UNIVERSIDAD AUTÓNOMA METROPOLITANA

Un modelo de regresión para datos en el Símplex D-dimensional

TESIS

Para obtener el título de:

MAESTRA EN CIENCIAS
MATEMÁTICAS APLICADAS E INDUSTRIALES

Presenta:
Angelica Amador Rescalvo

Director
Dr. Gabriel Núñez Antonio

México CDMX.,10/03/2017

Con todo mi amor y cariño a mis padres
Damiana y Leonardo.
Por su ejemplo de perseverancia,
constancia, apoyo y amor incondicional
que los caracteriza.
Todo ha sido posible gracias a ustedes.

Índice general

Introducción

En este trabajo de tesis se hace una revisión de los conceptos asociados al análisis de *datos composicionales*, haciendo énfasis en los problemas que se pueden presentar si estos son analizados con técnicas convencionales empleadas en el análisis de datos en $\mathbb{R}^k$. Se muestra la forma de analizarlos e implementar relaciones de regresión donde algunas variables involucradas sean de tipo composicional, poniendo énfasis en el caso de variables de respuesta composicional.

Aunque Francis Galton inventó la correlación, Karl Pearson fue el reponsable de su análisis, desarrollo y promoción como concepto científico en el ámbito estadístico (ver, Aldrich, 1995). Desde finales del siglo XIX Pearson señalaba el peligro en el que se podía incurrir cuando se interpretan correlaciones entre cocientes cuyo denominador y numerador contienen partes comunes. Así, de algún modo Pearson sugería que el análisis de variables composicionales, como proporciones de un todo, sería complicado.

Una característica de los datos composicionales es que las proporciones de una composición son de manera natural sujetas a una suma constante. Alrededor de 1960, el geólogo Felix Chayes retomó la aplicación y el análisis multivariado de datos composicionales. Él trató de separar lo que llamó correlación real de la correlación espuria e intento evitar el problema de la *clausura*. Por su parte, Aitchison, en los 80's, reconoció que las composiciones ofrecían información relativa, no absoluta, sobre las partes o componentes.

El hecho de que los log-cocientes sean más fáciles de trabajar matemáticamente comparados con los cocientes y que la transformación log-cociente ofrezca un mapeo uno-a-uno sobre el espacio real euclideano, ofreció un soporte para la construcción de una metodología basada sobre una variedad de transformaciones log-cocientes. Estas transformaciones permitieron el uso de procedimientos estadísticos multivariados no-restringidos a los datos transformados. Sin embargo, este enfoque propuesto por Aitchison presenta ciertas dificultades, derivadas del hecho de que implícitamente se asume la usual geometría Euclideana para el espacio muestral asociado a los datos composicionales, el Símplex unitario D-dimensional.

A principios del siglo XXI, varios investigadores reconocieron que las operaciones internas

en el Símplex (perturbación, potenciación, y la métrica correspondiente) definen un espacio vectorial, ver por ejemplo, Billheimer et al.(1997 y 2001) y Pawlowsky-Glahn y Egozcue (2001). Bajo este enfoque, el espacio muestral de vectores aleatorios composicionales es representado por el Símplex con una métrica diferente a la métrica Euclideana en el espacio real.

La presentación de este trabajo está organizado de la siguiente manera. En el Capítulo 1 se muestra la naturaleza de los datos composicionales, así como el Símplex como el espacio muestral natural para este tipo de variables. Se revisan las operaciones básicas para trabajar con este tipo de datos y se presenta la manera de hacer análisis descriptivo para datos composicionales. Se revisan además algunas transformaciones relevantes que mapean el Símplex D-dimensional al espacio Euclideano real.

En el Capítulo 2 se presenta y analiza el problema de regresión. En particular se revisan los conceptos de regresión lineal múltiple y regresión lineal multivariable, lo anterior como un paso de transición a los modelos de regresión donde se tengan variables composicionales, ya sea como variables independientes o como variables dependientes, este último caso siendo el objetivo central de este trabajo de tesis. El Capítulo finaliza con la presentación de un par de ejemplos.

En el Capítulo 3 se muestra la manera de aplicar los conceptos revisados y el análisis en una base de datos real. La base está asociada a los votos de las elecciones de 2015 en el Estado de Colima, para las elecciones de gobernador y diputados federales.

Finalmente, se presentan las conclusiones derivadas de este trabajo en el análisis y modelación de datos composicionales.

Capítulo 1

Preliminares

1.1. Naturaleza de los datos composicionales

Los datos composicionales son vectores con elementos no negativos que se expresan como proporciones y están sujetos a que la suma de sus elementos es una constante K. Estos datos aparecen en diferentes disciplinas como la biología, ecología, petrología y economía entre otras, en forma de proporciones de un todo. A continuación se presentan algunos ejemplos en diferentes disciplinas que ilustran la aplicabilidad de éste tipo de datos.

Composiciones Geoquímicas de rocas

En petrología es de gran importancia el análisis de composiciones geoquímicas de rocas. Comúnmente tales composiciones son expresadas en porcentajes por peso de 10 o más óxidos o son porcentajes por peso de algunos minerales básicos. Así, dado diferentes composiciones se puede tener interés en describir la variación de estas composiciones.

Sedimentos del lago Ártico en diferentes profundidades

En sedimentología, especímenes de sedimentos son separados en partes mutuamente exclusivos y exhaustivos (por ejemplo, arena, limo y arcilla) y las proporciones de estas partes por peso son llamadas composiciones. Algunas interrogantes que surgen al analizar estos datos son: ¿La composición de sedimentos depende de la profundidad del agua? Si es así, ¿cómo se puede modelar esa dependencia?

Estudio de la composición de la leche

Para mejorar la calidad de la leche de vaca, se estudia la composición de la leche que produce una de cada treinta vacas antes y después de una dieta estrictamente controlada y un régimen hormonal por un periodo de ocho semanas. Se decide tener el control de otras treinta vacas

criadas en las mismas condiciones, pero sobre un régimen regular al establecido. El propósito del experimento es determinar si el nuevo régimen ha producido algún cambio significativo en la composición de las diferentes proteínas de la leche.

En 1897 Karl Pearson identificó que existían problemas en el análisis e interpretación de los datos composicionales y a mediados del siglo XX, Chayes (1960) identificó dificultades derivadas de la restricción en la suma constante de sus componentes. A continuación se revisan algunos de los problemas que se pueden presentar en el análisis de datos composicionales.

1.1.1. Algunos problemas en el análisis de datos composicionales

Chayes (1960) menciona que ignorar la restricción de suma unitaria o incorporarla indebidamente en un modelado estadístico, puede provocar tener resultados erróneos, es decir, análisis inadecuados que llevan a tener inferencias dudosas o distorsionadas.
Algunos de los inconvenientes por ignorar esta restricción son: Sesgo en las correlaciones, incoherencias subcomposicionales, problemas al establecer relaciones lineales y problemas en el uso de operaciones clásicas equivalentes a las operaciones en $\mathbb{R}$. A continuación se ejemplifican algunos de estos problemas.

Sesgo en las correlaciones

Pearson (1897) fue el primero en identificar el problema de las malas correlaciones entre proporciones. El problema radica en que las correlaciones entre proporciones no son libres de tomar cualquier valor en el intervalo [-1,1].
Como ejemplo de lo anterior se analizó la correlación de 2 muestras de espesor de estratos X y Y, los primeros datos son el espesor expresado en metros y posteriormente el espesor en porcentaje.
A continuación se muestra en la *Tabla* 1.1, el espesor de estratos presentes en los datos *X* y *Y* (estos datos se obtuvieron de Rollinson, 1993).

Tabla 1.1. Espesor de estratos.

Datos	X (metros)	Y (metros)	X (porcentaje)	Y (porcentaje)
1	50.0	50.0	50.0	50.0
2	60.0	85.0	41.4	58.6
3	70.0	110.0	38.9	61.1
4	75.0	140.0	34.9	65.1
5	80.0	170.0	32.0	68.0
6	90.0	200.0	31.0	69.0
corr	0.9914		-1	

Los resultados del análisis pueden dejar ver que trabajar los datos en diferentes cantidades provoca que la correlación cambie. Para visualizar el comportamiento de la correlación de los datos expresados en metros y el de los datos representados como porcentaje se muestran las siguientes gráficas.

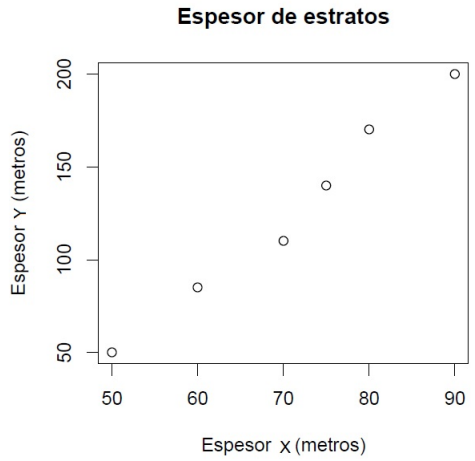


Figura 1.1. La correlación entre X y Y en metros es .99

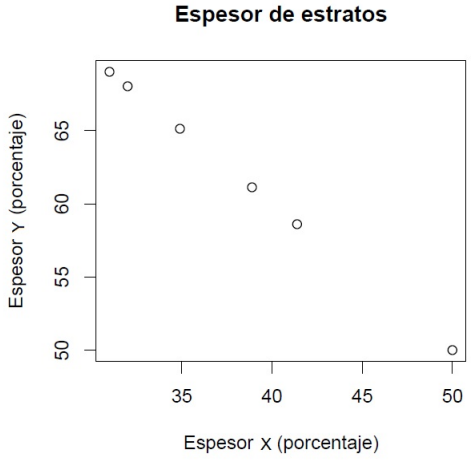


Figura 1.2. La correlación entre X y Y expresados como porcentajes es -1

La correlación de los datos expresados en metros como se muestra en la *Figura* 1.1 es .99, se observa que conforme aumenta el espesor X el espesor Y también aumenta, es decir, existe una relación positiva y proporcional entre las 2 variables. Sin embargo, cuando se trabaja con datos en forma de porcentaje (ver *Figura* 1.2) la correlación es -1, ahora la relación de estas

dos variables es negativa. La correlación no se conserva al trabajar con los datos en forma de porcentaje, causando problemas en el análisis de los datos.

Incoherencias en las subcomposiciones

Una subcomposición se define como cualquier subconjunto de variables (componentes) de una composición en la que se mantiene la condición de suma constante. Se espera que la correlación entre 2 partes sea la misma en la composición total así como en la subcomposición. Sin embargo, esto no ocurre y las correlaciones son diferentes. En el siguiente ejemplo se muestra la correlación de datos sobre estudios de muestras de suelo en diferentes situaciones: 1) al ser tomados de la composición completa W y 2) cuando se toman de una subcomposición Z. Los datos se obtuvieron de Aitchison (1997).

Sea $W = (w_1, w_2, w_3, w_4)$ la composición total, sea $Z = \left(\frac{w_1}{\sum_{i=1}^{3} w_i}, \frac{w_2}{\sum_{i=1}^{3} w_i}, \frac{w_3}{\sum_{i=1}^{3} w_i}\right) = (z_1, z_2, z_3)$ la subcomposición de W, de lo anterior se obtiene la siguiente tabla.

Tabla 1.2. Incoherencias en las correlaciones de subcomposiciones.

COMPOSICIÓN	SUBCOMPOSICIÓN
W	Z
(w_1, w_2, w_3, w_4)	$z_1, z_2, z_3)$
(0.1,0.2,0.1,0.6)	(0.25,0.50,0.25)
(0.2,0.1,0.1,0.6)	(0.50,0.25,0.25)
(0.3,0.3,0.2,0.2)	(0.375,0.375,0.25)

De los datos que se muestran en la *Tabla* 1.2, se toman los elementos (w_1, w_2) y (z_1, z_2) de la composición y subcomposición, respectivamente. Como puede verse la correlación que existe entre (w_1, w_2) de la composición de W es 0.5 y la correlación entre las partes de la composición Z es -1.
Con lo anterior se puede concluir que no se conserva la relación existente entre elementos en composiciones y subcomposiciones. La correlación de los elementos w_1 y w_2 en W es 0.5, pero cuando se analiza la correlación de estos mismos elementos en la subcomposición Z, la correlación es distinta, ahora es de -1. Así, existe un cambio al medir la relación que existe entre los elementos de una composición y en una subcomposición.

Dificultades para establecer relaciones lineales

El principal problema al establecer un modelo lineal de la forma $y = a + bx$ es que con las operaciones clásicas de suma y producto, cae fuera del espacio muestral en el que se encuentran los datos composicionales. La restricción de suma unitaria no se mantiene. Por

esta razón es necesario definir operaciones equivalentes en el Símplex a la suma y producto en $\mathbb{R}$ y así poder construir modelos adecuados para los distintos tipos de relación de regresión.

Los problemas mencionados anteriormente sobre el análisis de datos composicionales se conocen desde hace más de 100 años. En 1982 Aitchison propone que estas dificultades se pueden atacar si las partes que los conforman son tratadas como magnitudes relativas. Es decir, en cocientes $\left(\frac{x_i}{x_j}\right)$, $1 \leq i, j \leq n$, $i \neq j$ donde n es el número de elementos del vector. Esto se debe a que las composiciones proporcionan información sobre la magnitud relativa de sus partes. Por otro lado, cuando se trabaja con cocientes, las correlaciones espurias desaparecen y se eliminan los problemas que se tienen con las subcomposiciones.

En la siguiente sección se define el espacio muestral asociado a los datos composicionales.

1.2. El Símplex como espacio muestral

En términos más generales los números enteros de una composición

$$1, 2, \ldots, D$$

indican la cantidad de partes que conforman el dato composicional y las letras con subíndice $x_1, x_2, \ldots, x_D$ denotan los componentes. Así formalmente se tiene la siguiente definición.

Definición 1.1. Una *composición* x de D-partes es un vector $Dx1$ con componentes positivos $x_1, \ldots, x_D$ cuya suma es 1.

◇

Es importante tener en cuenta que la composición es completamente especificada por las d-partes de un subvector $x_1, \ldots, x_d$ donde $d = D - 1$ y $x_D = 1 - x_1 - x_2 - \cdots - x_d$.

1.2.1. El Símplex D-dimensional

Se debe tener cuidado en la dimensionalidad de una composición, ya que el espacio muestral se define en términos de un determinado subvector $(x_1, ..., x_d)$, pero la dimensión del vector es $D = d + 1$.

Definición 1.2. El *Símplex D−dimensional* es el conjunto definido por:

$$\mathcal{L}^D = \left\{ (x_1, \ldots, x_d, x_D) : x_1 > 0, \ldots, x_D > 0; \sum_{i=1}^{D} x_i = 1 \right\}$$

◇

Los datos composicionales cuentan con una estructura geométrica algebraica diferente a la de los reales. En la siguiente sección se muestran las operaciones básicas para trabajar en este espacio muestral, denominado el Símplex.

1.3. Operaciones con datos composicionales

Para que los elementos de estudio se encuentren en el espacio muestral Símplex se realiza la operación clausura, la cual se define a continuación.

Definición 1.3. El operador *Clausura C*, es una transformación que hace corresponder a cada vector $x = (x_1, x_2, ..., x_D) \epsilon \mathbb{R}_+^D$ su dato composicional asociado. Es decir,

$$C(x) = \left(\frac{x_1}{\sum_{i=1}^{D} x_i}, ..., \frac{x_D}{\sum_{i=1}^{D} x_i} \right),$$

donde $C(x) \epsilon \mathcal{L}^D$.

◇

A continuación se definen en el Símplex las operaciones de perturbación y potenciación, y son equivalentes a la suma y multiplicación en $\mathbb{R}$, respectivamente.

Definición 1.4. La operación *perturbación* entre $x = (x_1, x_2, ..., x_D) \epsilon \mathcal{L}^D$ y $y = (y_1, y_2, ..., y_D) \epsilon \mathcal{L}^D$ está definida por

$$x \oplus y = C\left(x_1 y_1, ..., x_D y_D\right),$$

donde C es el operador clausura.

◇

El conjunto de perturbaciones forma un grupo el cual tiene como elemento neutro la composición $\mathbb{E} = \left(\frac{1}{D}, ..., \frac{1}{D} \right)$ donde D es la dimensión del Símplex correspondiente. Este elemento neutral juega el papel del vector cero en los datos composicionales. La composición *inversa* de x, denotada por $\ominus x$, está dada por $\ominus x = C\left(\frac{1}{x_1}, ..., \frac{1}{x_D} \right)$, donde C es la clausura mencionada en la Definición 1.3. Esta operación juega el rol de la resta en datos composicionales.

Definición 1.5. La operación *potenciación* de $x = (x_1, x_2, ..., x_D) \epsilon \mathcal{L}^D$ por un número $\alpha \epsilon \mathbb{R}^+$ se define de la siguiente manera:

$$\alpha \odot x = C\left(x_1^{\alpha}, ..., x_D^{\alpha}\right),$$

donde C es el operador clausura.

◇

Se puede comprobar que con la operación de perturbación y potenciación por un escalar $\left(\mathcal{L}^D, \oplus, \odot\right)$ es un *Espacio vectorial*, (Billheimer et al. 2001).

A continuación se definen medidas estadísticas que consideran que la geometría del Símplex tiene una estructura diferente al espacio de los reales.

1.4. Estadística Descriptiva

En esta sección se presentan algunas medidas estadísticas de gran utilidad para la descripción de datos composicionales que serán utilizadas posteriormente. En $\mathbb{R}$ una medida estadística empleada con mayor frecuencia es una medida de ubicación. Se define a continuación la media geométrica en datos composicionales.

Definición 1.6. Sea $x_i = (x_{i1}, x_{i2}, ..., x_{iD}) \in \mathcal{L}^D$, la *media geométrica composicional* de x_i, $i = 1, 2, ..., n$ se define como

$$g(x_1, ..., x_i, ..., x_n) = C(g_1, g_2, ..., g_D),$$

donde

$$g_j = \left(\prod_{i=1}^{n} x_{ij}\right)^{\frac{1}{n}}$$

$j = 1, 2, ..., D$ y C el operador clausura.

◇

El patrón de variación de una composición $x = (x_1, x_2, ..., x_D) \in \mathcal{L}^D$, está determinada por la *matriz de variación*, $\tau_{ij} = var\left(ln\frac{x_i}{x_j}\right)$, $i, j = 1, 2, ..., D$. La matriz de variación es simétrica dado que $ln\left(\frac{a}{b}\right) = -ln\left(\frac{a}{b}\right)$ y $var(c) = var(-c)$.
Pequeños valores de τ_{ij} implican pequeñas varianzas en el $ln\left(\frac{a}{b}\right)$, entre más pequeño sea el valor de τ existe una mejor proporcionalidad entre los 2 elementos.

Es importante definir la norma, producto interior y distancia en el Símplex, estos elementos serán de gran ayuda para definir líneas, ángulos, ortogonalidad, etc., en el Símplex.

Definición 1.7. Sean $x = (x_1, x_2, ..., x_D)$ y $y = (y_1, y_2, ..., y_D)$ dos elementos de $\mathcal{L}^D$. El *producto interior de Aitchison* está definido como

$$< x, y >_A = \sum_{i=1}^{D} ln\frac{x_i}{g(x)} \cdot ln\frac{y_i}{g(y)},$$

donde $g(.)$ es la media geométrica .

◇

Definición 1.8. Sea $x = (x_1, x_2, ..., x_D) \epsilon \mathcal{L}^D$. La *norma de Aitchison* de x se define como

$$\|x\|_a = \sqrt{\frac{1}{2D}\sum_{i=1}^{D}\sum_{j=1}^{D}\left(ln\frac{x_i}{x_j}\right)^2}$$

◇

Definición 1.9. La *distancia de Aitchison* entre $x = (x_1, x_2, ..., x_D)$ y $y = (y_1, y_2, ..., y_D)$ dos elementos de $\mathcal{L}^D$ se define como

$$d_a(x, y) = \sqrt{\frac{1}{2D}\sum_{i=1}^{D}\sum_{j=1}^{D}\left(ln\frac{x_i}{x_j} - ln\frac{y_i}{y_j}\right)^2}$$

◇

Anteriormente se mencionó que $\mathcal{L}^D$ es un espacio vectorial y que está dotado de un producto interior. El producto interior de Aitchison, por esta razón $\left(\mathcal{L}^D, \oplus, \odot, <, >\right)$ es un *Espacio Euclídeo*.
A continuación se muestra un método para visualizar gráficamente datos composicionales.

1.4.1. Representaciones gráficas

Existen métodos gráficos para describir datos composicionales. En el caso D=3, uno de los más importantes son los diagramas ternarios. Para ilustrar este tipo de gráfico, sean n datos composicionales de los componentes de una roca $x_i = (Pb, Cu, Zn)$. Estos datos se representan en el diagrama ternario como se muestra en la Figura 1.3. Posteriormente se da una breve explicación de como ubicar estos datos composicionales.

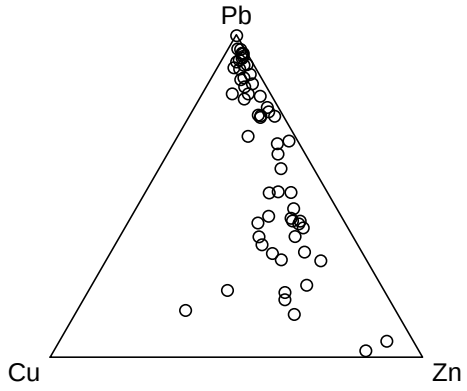


Figura 1.3. Conjunto de datos composicionales con 3 componentes.

Los diagramas ternarios son gráficos que se emplean para representar tres variables, cuya suma es un valor constante. Estos diagramas se representan con un triángulo equilátero, donde cada punto dentro de él representa un dato composicional compuesto por 3 componentes, tal como se muestra en la Figura 1.3.

Como ejemplo de lo anterior, sea X un dato composicional conformado por 3 componentes (A, B, C). En el diagrama ternario cada uno de los vértices del triángulo representa una componente. Cada vértice equivale al 100 % del componente que le pertenece del dato composicional y este porcentaje decrece conforme el punto avanza hacia el lado opuesto, como se muestra en los siguientes diagramas (Figura 1.4).

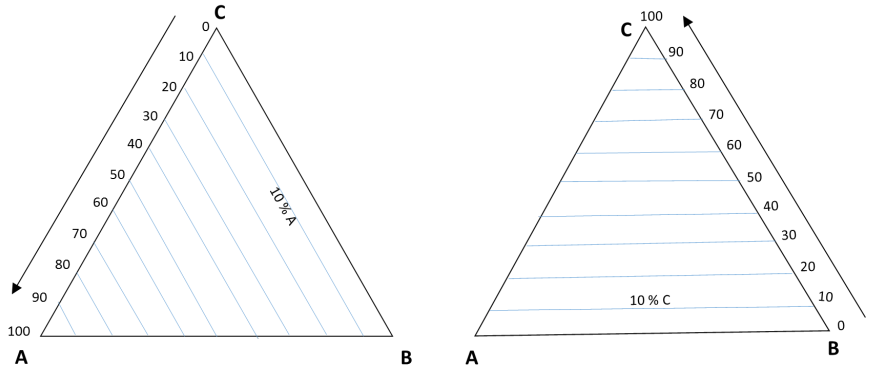


Figura 1.4. Incrementos de 2 elementos del dato composicional en el diagrama ternario.

El vértice A equivale al 100 % de la concentración de este componente y va disminuyendo conforme la línea se acerca al segmento $\overline{BC}$, por ejemplo, existe 0 % del componente C en el segmento $\overline{AB}$. Para poder graficar un dato composicional X, primero se localiza el nivel de porcentaje de cada uno de los elementos del dato composicional en el diagrama ternario, posteriormente se ubica la intersección de los tres segmentos y este punto es el que le pertenece al dato composicional X en el diagrama ternario.
Por ejemplo: Sea $X = (3, 6, 8)\ \epsilon\ \mathbb{R}^3$ se aplica el operador clausura y se tiene el dato composicional X_c= (0.18, 0.35, 0.47) $\epsilon\ \mathcal{L}^3$, (Ver Figura 1.5).

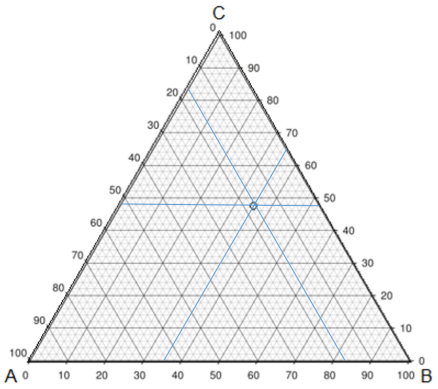


Figura 1.5. Diagrama ternario que muestra el dato composicional X_c

Los diagramas ternarios son de gran ayuda para observar de manera gráfica patrones de variabilidad que puedan existir en los datos, realizar pruebas de significancia, análisis en modelos de regresión, etc. Sin embargo, al realizar análisis estadísticos se puede estar interesado en trabajar solamente con una cierta cantidad de componentes del dato composicional, denominado subcomposiciones y se estudia en la siguiente sección.

1.4.2. Subcomposiciones

Hablar de subcomposiciones es tomar un subconjunto de las componentes del dato composicional en el Símplex D-dimensional y por medio de un operador construir un nuevo espacio r-dimensional en el Símplex con $r < D$. Así se tiene la siguiente definición

Definición 1.10. Si S es un subconjunto de las partes $(1, \ldots, D)$ de una composición x de D-partes, y x_S es el subvector de componentes de x, entonces $C(x_S)$ es llamada una *subcomposición* de S-partes, con C el operador *clausura*.

◇

Por ejemplo, sea $X = (x_1, x_2, x_3, x_4, x_5)$ la composición de una roca Hongite, se puede estar interesado en la subcomposición geoquímica de los primeros 3 elementos (x_1, x_3, x_5).

Con la definición anterior y la propuesta de Aitchison para datos composicionales, se obtiene el siguiente resultado.

Propiedad 1.1. La tasa de 2 componentes de una subcomposición es igual a la tasa de los correspondientes 2 componentes en la composición completa. Es decir, sí $S = C(X_s) = C\left(x_1, x_2, ..., x_p\right)$ entonces

$$\frac{s_i}{s_j} = \frac{x_i}{x_j}$$

$1 \leq i, j \leq p$ y $p < D$

$\diamond$

Como se puede observar de la propiedad 1.1, una subcomposición conserva la relación de proporción existente en la relación de los datos en la composición total. Sin embargo bajo la propuesta de Aitchison, se dificulta entender la geometría del Símplex $\mathcal{L}^D$, razón por la cual propone tomar logaritmos de los cocientes entre las partes, así los datos son mapeados a $\mathbb{R}_+^{D-1}$ y bajo esta transformación es posible utilizar cualquier método de estadística clásica.

Una medida que explica que tanta relación existe entre 2 variables es la covarianza, que se estudia en la siguiente sección.

1.4.3. Estructura de Covarianza

En esta sección se presenta el concepto de covarianza, la cual es relevante en el análisis de la variabilidad de datos composicionales.

Como ya se ha mencionado anteriormente una composición x puede ser completamente determinada por d cocientes tales como $\frac{x_i}{x_D}$, $(i = 1, ..., d)$.

De acuerdo a la propuesta de Aitchison los datos composicionales se trabajan en términos de magnitudes relativas. Por esta razón la covarianza queda definida como la covarianza de las magnitudes relativas del dato composicional, tal como se muestra a continuación

$$Cov\left(\tfrac{x_i}{x_k}, \tfrac{x_j}{x_l}\right),\ i, j, k, l \text{ valores de } 1, ..., D$$

Aun así se tiene la dificultad de solo trabajar en el ortante positivo $\mathbb{R}_+^{D-1}$. Para trabajar en todo el espacio $\mathbb{R}^{D-1}$ Aitchison propone tomar logaritmos de los cocientes entre las partes.Así,

$$\sigma_{ij,kl} = Cov\left\{log\left(\frac{x_i}{x_k}\right), log\left(\frac{x_j}{x_l}\right)\right\}.$$

Formalizando las ideas anteriores, se tienen las siguientes definiciones.

Definición 1.11. La estructura de *covarianza* de una composición x de D-partes es el conjunto

$$\sigma_{ij,kl} = Cov\left\{log\left(\frac{x_i}{x_k}\right), log\left(\frac{x_j}{x_l}\right)\right\} \quad i, j, k, l = 1, ..., D.$$

◇

Cuando i, j, k, l toman solo 2 valores, se obtiene la definición de varianza, es decir,

Definición 1.12. Para 2 elementos i y j de una composición x de $D-$partes la *varianza logcociente* se define como

$$var\left\{log\left(\frac{x_i}{x_j}\right)\right\}.$$

◇

Existen diferentes formas útiles y equivalentes para describir los patrones de variabilidad, cada una de estas covarianzas posee diferentes propiedades como se menciona a continuación.

1.4.4. Especificación de la estructura de covarianza

En esta sección se estudiarán algunas de las estructuras de covarianza como la matriz de covarianza logcociente y la matriz de covarianza logcociente centrada, así como algunas de sus características, que serán de gran ayuda para el análisis posterior de datos composicionales.

Matriz de covarianza logcociente

A continuación se define la covarianza en el Símplex como la relación existente entre 2 variables al ser tratadas como logcociente de los elementos.

Definición 1.13. Para una composición x de D-partes la matriz de $d \times d$

$$\Sigma = Cov\left\{log\left(\frac{x_i}{x_D}\right), log\left(\frac{x_j}{x_D}\right)\right\} \quad \forall\ i, j = 1, ..., d$$

es denominada la matriz de *covarianza logcociente*.

◇

Esta matriz de covarianza, generalmente es una matriz no singular y es asimétrica en el tratamiento de las partes.

Matriz de covarianza logcociente centrada

Una manera de conservar la forma especificada anteriormente de la matriz de covarianza y al mismo tiempo obtener un modelo simétrico de todas las D-partes es reemplazando el divisor x_D por la media geométrica $g(x)$.

Definición 1.14. Para una composición x de D-partes la matriz de $D \times D$

$$\Gamma = Cov\left\{log\left(\frac{x_i}{g(x)}\right), log\left(\frac{x_j}{g(x)}\right)\right\} \quad i, j = 1, ..., D$$

es llamada la matriz de *covarianza logcociente centrada*.

◇

La matriz de covarianza logcociente centrada tiene una estructura de matriz de covarianza, generalmente es una matriz singular y es simétrica en el tratamiento de las partes.

1.4.5. Logcocientes y logcontrastes

Continuando con el estudio de la variabilidad en el Símplex, se mostrará que los componentes logcocientes son muy importantes en el análisis de los datos composicionales.
Considérese la combinación lineal $a'Y$ con $Y = \left(log\left(\frac{x_1}{x_D}\right), \cdots, log\left(\frac{x_d}{x_D}\right)\right) \epsilon \mathbb{R}$ y $a \epsilon \mathbb{R}$. Se obtiene que

$$\begin{aligned} a_1 log\left(\tfrac{x_1}{x_D}\right) + \cdots + a_d log\left(\tfrac{x_d}{x_D}\right) &= a_1 logx_1 + \cdots + a_d logx_d - (a_1 + ... + a_d)\, logx_D \\ &= a_1 logx_1 + \cdots + a_d logx_d + a_D logx_D \end{aligned}$$

donde $a_1 + ... + a_d + a_D = 0$. Es decir, existe una relación 1-1 entre una combinación loglíneal de los elementos cocientes de Y y la combinación loglineal de los elementos de la composición x, se presentan características similares como las que se dan en un modelo lineal.

Definición 1.15. Un *logcontraste* de una composición x de D-partes es una combinación de la siguiente forma

$$a_1 logx_1 + \cdots + a_d logx_d + a_D logx_D = a' logx \quad \text{con} \quad a_1 + \cdots + a_D = 0$$

◇

La covarianza es de gran ayuda para indicar el grado de asociación entre 2 variables como se mencionó con anterioridad.

Definición 1.16. Una composición x de D-partes es un *logcociente no correlacionado* si

$$\sigma_{ij.kl} = Cov\left\{log\left(\frac{x_i}{x_k}\right), log\left(\frac{x_j}{x_l}\right)\right\} = 0, \quad \forall\ i, j, k, l = 1, ..., D,$$

diferentes entre sí.

◇

Este concepto puede reescribirse en términos de logcontrastes de la siguiente manera.

Definición 1.17. Una composición x de D-partes es un *logcontraste no correlacionado* si existen dos logcontrastes ortogonales no correlacionados, tal que:

$$Cov\left(a'log(X), b'log(X)\right) = 0 \quad \text{donde} \quad a'b = 0 \quad a, b \in \mathbb{R}$$

◇

1.5. Distribuciones en el Símplex

Una vez que se tiene una estructura de covarianza, se necesita encontrar una clase paramétrica en $\mathcal{L}^D$ para describir patrones de variabilidad en el Símplex.

1.5.1. La Distribución de Dirichlet

La distribución de Dirichlet $Dir(\alpha)$ es una familia de distribuciones de probabilidad multivariable, continua y parametrizada por un vector $\alpha = (\alpha_1, ..., \alpha_K)$ real de términos positivos. Esta distribución es la generalización multivariable de la distribución beta. La distribución de Dirichlet de orden $K \geq 2$ con parámetros $\alpha_1, ..., \alpha_K > 0$ tiene una función de densidad de probabilidad en el espacio Euclidiano $\mathbb{R}^{K-1}$ dada por:

$$f(x_1, ..., x_K|\alpha_1, ..., \alpha_K) = \frac{1}{B(\alpha)} \prod_{i=1}^{K} x_i^{\alpha_i - 1},$$

La distribución Beta es una distribución de probabilidad continua, cuya función de densidad se genera con valores $x \in [0, 1]$, de esta forma la distribución Dirichlet es una distribución definida en el Símplex abierto de $(K-1)$-dimensional definido por:

$$x_1, ..., x_{K-1} > 0.$$
$$x_1 + ... + x_{K-1} < 1$$
$$x_K = 1 - x_1 - ... - x_{K-1},$$

y cero en otro caso.

La constante de normalización es la función Beta multinomial $B(\alpha)$, la cual se puede expresar en términos de la función Gamma. Es decir,

$$B(\alpha) = \frac{\prod_{i=1}^{K} \Gamma(\alpha_i)}{\Gamma\left(\sum_{i=1}^{K} \alpha_i\right)}, \quad \text{con} \quad \alpha = (\alpha_1, ..., \alpha_K).$$

Desafortunadamente, esta familia paramétrica no es adecuada para la descripción de la variabilidad de datos composicionales, principalmente cuando en los datos composicionales se tienen patrones cóncavos, debido a que los contornos de isoprobabilidad de $D(\alpha)$ son convexos. Además la clase de Dirichlet no soporta un grado suficiente de dependencia composicional.

1.5.2. La Distribución Logística Normal

En busca de alternativas a la clase Dirichlet, McAlister (1879) se percató que los datos composicionales adoptaban patrones similares a los de una normal $\mathcal{N}(\mu, \sigma^2)$ en la línea real por medio de la siguiente transformación $w = exp(y)$ donde $w \epsilon \mathbb{R}^1_+$ y $y\epsilon \mathbb{R}^1$ cuya transformación inversa es $y = log(w)$ con $w \epsilon \mathbb{R}^1_+, y \epsilon \mathbb{R}^1$.
De esta manera, se obtiene la clase de distribución lognormal, aplicando estas transformaciones se introduce una distribución en el Símplex.

Propiedad 1.2. Los datos en el Símplex siguen una *Distribución Lognormal*, si estos datos siguen una distribución normal multivariada en $\mathbb{R}^d$, los datos son llevados a los reales por medio de una transformación inyectiva de $\mathbb{R}^d$ a $\mathcal{L}^D$.

$\diamond$

Un punto importante a señalar es que los parámetros de covarianza Σ de la distribución en $\mathcal{L}^D$ son precisamente la matriz de covarianza de los logcocientes, Σ, de esta manera se puede adoptar un modelo logístico normal para la descripción de los patrones de variabilidad en el Símplex. Por lo tanto, se tiene la ventaja de poder hacer uso de procedimientos basados en la normal multivariada. Es decir; se transforman cada una de las componentes composicionales x en logcociente, así se puede trabajar en $\mathbb{R}^d$ y hacer uso de todos los supuestos y procedimientos asociados a la distribución normal multivariada.

Función de densidad.

De acuerdo con la función de densidad de la distribución $\mathcal{N}^d(\mu, \Sigma)$ se tiene

$$f(y \mid \mu, \Sigma) = (2\pi)^{\frac{-d}{2}} |\Sigma|^{\frac{-1}{2}} exp\left\{\frac{-1}{2}(y-\mu)'\Sigma^{-1}(y-\mu)\right\}$$

con $y \epsilon \mathbb{R}^d$. Aplicando la transformación logística y el teorema de cambio de variable $y \to x$

se obtiene la correspondiente función de densidad de la distribución logística Normal $\mathcal{L}^D(\mu, \Sigma)$:

$$f(x \mid \mu, \Sigma) = (2\pi)^{\frac{-d}{2}} |\Sigma|^{\frac{-1}{2}} (x_1 \cdots x_D)^{-1} exp\left\{\frac{-1}{2}(y-\mu)' \Sigma^{-1}(y-\mu)\right\}$$

donde $y = log(x_i/x_D), \quad i = 1, \ldots, d.$

Para usar distribuciones en $\mathbb{R}^{D-1}$ inducidas por composiciones en el Símplex, Aitchison (1986) propone hacer uso de las siguientes transformaciones inyectivas, basadas en logaritmos de cocientes entre las partes de un dato composicional, la transformación logcociente aditiva y la transformación logcociente centrada.

1.5.3. Transformación logística normal aditiva.

Las siguientes transformaciones se basan en tomar un dato composicional x en $\mathcal{L}^D$ expresado en logcocientes y por medio de la transformación llevarlo al espacio $\mathbb{R}^{D-1}$. De esta manera será posible utilizar técnicas multivariantes en el espacio de los reales. Con este tipo de transformaciones se resuelven los problemas que se mencionaron en secciones anteriores, sobre los problemas existentes al analizar los datos composicionales (la restricción de suma constante, correlaciones espurias y problemas en las subcomposiciones).

Definición 1.18. La *transformación logcociente aditiva* (*alr*) es una transformación 1-1 de $x = (x_1, ..., x_D) \in \mathcal{L}^D$ a $Y \in \mathbb{R}^{D-1}$ definida por

$$Y = alr(x) = \left(log\frac{x_1}{x_D}, log\frac{x_2}{x_D}, ..., log\frac{x_{D-1}}{x_D}\right).$$

◇

Desafortunadamente la transformación *alr* es asimétrica respecto a las partes de la composición, debido a que la componente utilizada como denominador x_D cobra especial protagonismo. Por otro lado, con estas coordenadas no es posible usar el producto interno habitual y la distancia (ver Egozcue y Pawlowsky-Glahn ,2005).

Definición 1.19. *La transformación logcociente centrada* (*clr*) de una composición $x = (x_1, ..., x_D) \in \mathcal{L}^D$ de D-partes a $Z \in \mathbb{R}^D$ se define como

$$Z = clr(x) = \left(log\frac{x_1}{g(x)}, log\frac{x_2}{g(x)}, ..., log\frac{x_D}{g(x)}\right)$$

donde $g(x)$ es la media geométrica de x.

◇

La transformación *clr* es simétrica e isométrica en las partes, pero la imagen de $\mathcal{L}^D$ está restringida a un subespacio de $\mathbb{R}^D$ y las matrices de covarianza y correlación de los datos *clr* transformados son singulares ($det = 0$).

En el 2003 Egozcue y sus colaboradores proponen la transformación logcociente isométrica. Esta evita los problemas existentes en las transformaciones anteriores.

Definición 1.20. *La transformación logcociente isométrica* (*ilr*) de una composición x de D-partes $\epsilon\ \mathcal{L}^D$ se define como

$$ilr(x) = (\langle x, e_1\rangle_A\ , ..., \langle x, e_{D-1}\rangle_A)$$

donde $(e_1, ..., e_{D-1})$ es una base ortonormal del Símplex.

◇

Esta transformación (*ilr*) es isométrica, la ventaja es que transforma los datos composicionales en coordenadas en un sistema ortogonal, es decir, se puede usar cualquier técnica estadística multivariante para su estudio.

La razón por la que existen diferentes tipos de transformaciones se debe a que ninguna tiene las propiedades perfectas, al tratar con las operaciones básicas en el Símplex.

A continuación en la Tabla 1.3 se muestran algunas propiedades de las 3 transformaciones mencionadas con anterioridad.

Tabla 1.3. Propiedades de las transformaciones.

Transformación Logcociente Aditiva
$alr(x \oplus y) = alr(x) + alr(y)$
$alr(\lambda \odot x) = \lambda.alr(x)$
$\langle x, y\rangle_A \neq alr(x).alr\,'(\mathrm{y})$

Transformación Logcociente Centrada
$clr(x \oplus y) = clr(x) + clr(y)$
$clr(\lambda \odot x) = \lambda.clr(x)$
$\langle x, y \rangle_A = clr(x).clr\,'(\mathrm{y})$
$\langle x, y \rangle_A = \langle clr(x), clr(y) \rangle$
Transformación Logcociente Isométrica
$ilr(x \oplus y) = ilr(x) + ilr(y)$
$ilr(\lambda \odot x) = \lambda.ilr(x)$
$\langle x, y \rangle_A = ilr(x).ilr\,'(\mathrm{y})$
$\langle x, y \rangle_A = \langle ilr(x), ilr(y) \rangle$

Donde "." representa, el producto en los reales y " $'$ " representa la transpuesta de ese vector.

Capítulo 2

Regresión de datos composicionales

El análisis de regresión es una técnica estadística para modelar la relación entre variables. En algunos campos del conocimiento, la regresión es la técnica estadística más utilizada. En este capítulo se explican brevemente los modelos de regresión existentes en $\mathbb{R}$ y los modelos de regresión en el Símplex, así como la relación que existe entre estos dos tipos de modelos.

2.1. Regresión lineal simple

El modelo $y = \beta_0 + \beta_1 x + \epsilon$ con $\epsilon \sim \mathcal{N}(0, \sigma^2)$ es denominado *modelo de regresión lineal simple*, donde x es la variable independiente, y es la variable dependiente (o variable respuesta) del modelo, β_0 y β_1 los coeficientes de regresión desconocidos. La respuesta media en cualquier valor de x es $E(y|x) = E(\beta_0 + \beta_1 x + \epsilon) = \beta_0 + \beta_1 x$, β_0 representa la ordenada de y, su valor es el punto en el que la línea recta de los valores promedios cruza el eje y, y β_1 es la pendiente. En el caso de una variable independiente continua β_1 representa el cambio que existe en y cuando la variable independiente x aumenta en una unidad.

Como se mencionó anteriormente, los parámetros β_0 y β_1 son desconocidos y estos se estiman con los datos de la muestra. Sean n pares de datos $(x_1, y_1), ..., (x_n, y_n)$, un método para estimar los coeficientes de regresión es el método de mínimos cuadrados, el cual consiste en minimizar la suma de los cuadrados del error entre los puntos estimados en la recta y los puntos observados, es decir, se tiene que minimizar

$$S(\beta_0, \beta_1) = \sum_{i=1}^{n} (y_i - \beta_0 - \beta_1 x_i)^2,$$

Por lo tanto, los estimadores de mínimos cuadrados, $\hat{\beta}_0$ y $\hat{\beta}_1$, satisfacen

$$\frac{\partial S}{\partial \beta_0} = -2 \sum_{i=1}^{n} (y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i) = 0$$

$$\frac{\partial S}{\partial \beta_1} = -2\sum_{i=1}^{n}(y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i)x_i = 0$$

Despejando de las ecuaciones anteriores, los estimadores correspondientes a cada uno de los parámetros, se concluye que los correspondientes estimadores se calculan de la siguiente manera:

$$\widehat{\beta_0} = \overline{y} - \widehat{\beta_1}\overline{x}$$

$$\widehat{\beta_1} = \frac{\sum_{i=1}^{n} xy - \overline{xy}}{\sum_{i=1}^{n} x^2 - \overline{x}^2}$$

Donde,

$\overline{x}$ es la media de los valores de la variable independiente y $\overline{y}$ es la media de los valores observados.

Los estimadores por mínimos cuadrados $\widehat{\beta_0}$ y $\widehat{\beta_1}$ son estimadores insesgados de los parámetros β_0 y β_1; es decir

$$E(\widehat{\beta_1}) = \beta_1,$$

$$E(\widehat{\beta_0}) = \beta_0.$$

La suma de los cuadrados de los residuales, está dada por

$$SS_{RES} = \Sigma_i^n (y_i - \widehat{y_i})^2$$

Al sustituir $\widehat{y_i} = \widehat{\beta_0} + \widehat{\beta_1} x_i$ en la ecuación anterior, se obtiene,

$$SS_{RES} = \Sigma_i^n {y_i}^2 = n(\bar{y})^2 - \widehat{\beta_1} S\,xy$$

donde

$$S\,xy = \sum_{i=1}^{n} y_i(x_i - \overline{x})$$

y

$$SS_T = \sum_{i=1}^{n} {y_i}^2 - n(\bar{y})^2 = \sum_{i=1}^{n} (y_i - \bar{y})^2$$

De esta forma se tiene que

$$SS_{RES} = SS_T - \widehat{\beta_1} S\,xy$$

Más adelante se muestra que el valor esperado de SS_{RES} es $E(SS_{RES}) = (n-2)\sigma^2$. De esta manera un estimador insesgado de σ^2 está dado por

$$\widehat{\sigma}^2 = \frac{SS_{RES}}{n-2}.$$

Un modelo de regresión donde interviene más de una variable independiente se denomina *modelo de regresión múltiple* y se expresa de la siguiente manera:

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + ... + \beta_k X_k + \epsilon.$$

Este modelo es una generalización del modelo de regresión lineal simple. Una forma cómoda de representar los datos de este modelo es expresándolo en su forma matricial,

$$y = X\beta + \epsilon$$

$$\begin{pmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{pmatrix} = \begin{pmatrix} 1 & x_{11} & x_{12} & \cdots & x_{1k} \\ 1 & x_{21} & x_{22} & & x_{2k} \\ \vdots & & & & \vdots \\ 1 & x_{n1} & x_{n2} & \cdots & x_{nk} \end{pmatrix} \cdot \begin{pmatrix} \beta_0 \\ \beta_1 \\ \vdots \\ \beta_k \end{pmatrix} + \begin{pmatrix} \epsilon_1 \\ \epsilon_2 \\ \vdots \\ \epsilon_n \end{pmatrix}$$

donde

$$Y = \begin{pmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{pmatrix}, \quad X = \begin{pmatrix} 1 & x_{11} & x_{12} & \cdots & x_{1k} \\ 1 & x_{21} & x_{22} & & x_{2k} \\ \vdots & & & & \vdots \\ 1 & x_{n1} & x_{n2} & \cdots & x_{nk} \end{pmatrix}$$

$$\beta = \begin{pmatrix} \beta_0 \\ \beta_1 \\ \vdots \\ \beta_k \end{pmatrix} \; y \; \epsilon = \begin{pmatrix} \epsilon_1 \\ \epsilon_2 \\ \vdots \\ \epsilon_n \end{pmatrix}.$$

Por el método de mínimos cuadrados se obtiene que el estimador del vector de parámetros β es

$$\hat{\beta} = (X'X)^{-1}X'y.$$

Y el estimador de σ^2, es

$$\widehat{\sigma}^2 = \frac{SS_{RES}}{n-k-1}.$$

donde

$$\begin{aligned} SS_{RES} &= \sum_{i=1}^{n}(y_i - \hat{y}_i)^2 \\ &= (\text{y - } \hat{y})'(\text{y - } \hat{y}) \\ &= \left[y - X(X'X)^{-1}X'y\right]'\left[y - X(X'X)^{-1}X'y\right] \\ &= \text{y}'\left[I - X(X'X)^{-1}X'\right]y. \end{aligned}$$

El valor esperado de SS_{Res}, se calcula como sigue

$$E(SS_{RES}) = E\left(y'\left[I - X(X'X)^{-1}X'\right]y\right)$$

$$= traza\left(\left[I - X(X'X)^{-1}X'\right]\sigma^2 I\right) + E(y)'\left[I - X(X'X)^{-1}X'\right]E(y)$$

$$= (n - k - 1)\sigma^2.$$

Por lo tanto,

$$E(MS_{RES}) = E\left(\frac{SS_{RES}}{n - k - 1}\right) = \sigma^2$$

Se sigue que $\hat{\sigma}^2$ es un estimador insesgado de σ^2.

Análisis de varianza

En estadística el análisis de varianza (ANOVA) permite determinar si diferentes tratamientos muestran diferencias significativas o por el contrario puede suponerse que sus medias poblacionales no difieren. El análisis de varianza se basa en la descomposición de la variación total expresada de la siguiente manera:

$$SS_T = SS_{RES} + SS_{REG}.$$

$$\sum_{i=1}^{n}(y_i - \bar{y}_i)^2 = \sum_{i=1}^{n}(y_i - \hat{y}_i)^2 + \sum_{i=1}^{n}(\hat{y}_i - \bar{y}_i)^2.$$

Donde

La suma de los cuadrados de los errores (SS_{RES}) es igual $\sum_{i=1}^{n}(y_i - \hat{y}_i)^2$, la suma total de los cuadrados (SS_T) se define como $\sum_{i=1}^{n}(y_i - \bar{y}_i)^2$ y su diferencia $\sum_{i=1}^{n}(y_i - \bar{y}_i)^2 - \sum_{i=1}^{n}(y_i - \hat{y}_i)^2$ se denomina la suma de los cuadrados de la regresión (SS_{REG}).

Por medio de un paquete estadístico como se verá en capítulos posteriores, se calcula la tabla de análisis de varianza, la cual contiene la siguiente información.

Tabla 2.1. Análisis de Varianza.

Fuente de variación	Df	Sum Sq	Mean Sq	F value	Pr(>F)
Regresión	k	SS_{REG}	$\frac{SS_{REG}}{k} = MS_{REG}$	$F = \frac{MS_{REG}}{MS_{RES}}$	Valor de p
Error	$n-k-1$	SS_{RES}	$\frac{SS_{RES}}{n-k-1} = MS_{RES}$		

Df representa los grados de libertad de la regresión, es decir; el número de variables independientes.

Sum Sq representa la suma de cuadrados de regresión.

Mean Sq representa la media de los cuadrados, que son las sumas de cuadrados divididas entre sus respectivos grados de libertad.

F value representa los cuadrados medios para la regresión dividida entre los cuadrados medios para el error.

Pr(>F) es el estadístico F (prueba de Fisher), con el que se prueba la hipótesis nula de que ninguna de las variables independientes están relacionadas linealmente con las variables dependientes.

En un modelo de regresión lineal simple, la prueba de hipótesis es definida como H_0=$\beta_1 = 0$ la hipótesis nula. Si se acepta la hipótesis nula, entonces el modelo lineal no podría ser útil. Por otra parte la hipótesis nula análoga en la regresión múltiple se define como $H_0 : \beta_1 = , ..., = \beta_k$, es decir; las medias poblacionales son iguales. Ésta hipótesis establece que ninguna de las variables independientes tiene alguna relación lineal con la variable dependiente. La hipótesis alternativa $H_1 : \beta_1 \neq \beta_i$, $i = 1, ..n$ variables, es decir; al menos dos medias poblacionales son distintas.
El estadístico de prueba para esta hipótesis es

$$F = \frac{SS_{REG}/k}{SS_{RES}/(n-k-1)}$$

éste es un estadístico F con el que se prueba la hipótesis nula de que ninguna de las variables independientes están relacionadas linealmente con las variables dependientes.

2.2. Análisis de regresión multivariable

Ante la necesidad de utilizar más variables tanto en las variables explicativas como en las variables respuesta se introduce el modelo de regresión multivariante, que se define de la siguiente manera.

Considérese un modelo definido por:

$$\left[Y_1, Y_2, ..., Y_p\right] = \left[a_1, a_2, ..., a_p\right] + \left[X_1, X_2, ..., X_q\right] . \begin{pmatrix} b_{11} & b_{12} & \cdots & b_{1p} \\ b_{21} & b_{22} & & b_{2p} \\ \vdots & & & \vdots \\ b_{q1} & b_{q2} & \cdots & b_{qp} \end{pmatrix} + \left[u_1, u_2, ..., u_p\right]$$

Con u_K los errores aleatorios. Así, en forma matricial se tiene

$$Y = XB + U,$$

donde:
$Y(n \times p)$ es una matriz observada de p variables respuesta de n observaciones.

$X(n \times (q + 1))$ es una matriz conocida de q variables explicativas, con unos en la primera columna.

$B((q+1)\times p)$ es una matriz de parámetros de regresión desconocidos, donde el primer renglón es la ordenada.

$U(n \times p)$ es una matriz de errores aleatorios.

Se asume que para cada X los renglones de U son no correlacionados, cada uno con media 0 y matriz de varianza común Σ.

Las columnas de Y representan observaciones de variables dependientes las cuales son explicadas en términos de las variables independientes X. Se tiene así,

$$Y = \begin{pmatrix} y_1' \\ y_2' \\ \vdots \\ y_n' \end{pmatrix}, \quad X = \begin{pmatrix} x_1' \\ x_2' \\ \vdots \\ x_n' \end{pmatrix}, \quad U = \begin{pmatrix} u_1' \\ u_2' \\ \vdots \\ u_n' \end{pmatrix},$$

donde

$$y_i = (y_{i1}, \quad y_{i2}, ..., y_{ip}), \quad x_i = (x_{i1}, x_{i2}, ..., x_{iq}), \quad u_i = (u_{i1}, u_{i2}, ..., u_{ip}) \quad \text{con} \quad i = 1, 2, ..., n.$$

$B = (\beta_{(1)}, \beta_{(2)}, ..., \beta_{(p)})$ con $\beta_{(j)} = (\beta_{1j}, \beta_{2j}, ..., \beta_{qj})'$ para $j = 1, ..., p$.

Se asume que u_i es distribuido normalmente, esto es que U es una matriz $N_p(0, \Sigma)$. Para determinados análisis la representación por renglones es más conveniente. En este caso el modelo se representa de la siguiente manera

$$y_i = B' x_i + u_i \quad \text{con} \quad i = 1, ..., n.$$

A continuación se estiman los parámetros desconocidos del modelo de regresión multivariable, por medio del método de máxima verosimilitud. Sea Y una muestra aleatoria simple de tamaño n, donde $y_i \sim N_p(B, \Sigma)$ entonces la función de densidad conjunta de Y es:

$$f_Y(y|B, \Sigma) = \frac{1}{(2\pi)^{np/2}|\Sigma|^{n/2}} exp\left[-\frac{1}{2}\sum_{i=1}^{n}(y_i - B' x_i)' (\Sigma)^{-1} (y_i - B' x_i)\right].$$

Para facilitar algunos desarrollos primero se trabajará con el exponente de la segunda parte de la función de densidad.

$$-\frac{1}{2}\sum_{i=1}^{n}(y_i - B' x_i)' (\Sigma)^{-1} (y_i - B' x_i)$$

Se observa que Σ^{-1} es una matriz simétrica de dimensión $(p \times p)$ y $(y_i - B' x_i)$ tiene dimensión $(p \times 1)$, por lo tanto se puede aplicar la siguiente propiedad.

Propiedad 2.1. Sea A una matriz simétrica de $k \times k$ y x es un vector de $k \times 1$. Entonces $x'Ax = tr(Axx')$.

◇

Aplicando la propiedad 2.1 se tiene que,

$$\begin{aligned}
&-\frac{1}{2}\sum_{i=1}^{n} tr\left[(\Sigma)^{-1} (y_i - B' x_i)(y_i - B' x_i)'\right] \\
&= -\frac{1}{2} tr\left[(\Sigma)^{-1} \sum_{i=1}^{n}(y_i - B' x_i)(y_i - B' x_i)'\right] \\
&= -\frac{1}{2} tr\left[(\Sigma)^{-1} (Y' - B'X')(Y' - B'X')'\right] \\
&= -\frac{1}{2} tr\left[(\Sigma)^{-1} (Y - XB)'(Y - XB)\right] \qquad (2.1)
\end{aligned}$$

Las dimensiones del producto de las matrices son: $M = (\Sigma)^{-1} (Y - XB)'$ una matriz de $(p \times n)$ y $H = (Y - XB)$ una matriz de $(n \times p)$.

Propiedad 2.2. Si A es una matriz de tamaño $p \times n$ y sea B una matriz de $n \times p$ entonces $tr(AB) = tr(BA)$.

$\diamond$

Utilizando la propiedad 2.2 en la ecuación (2.1) se obtiene

$$-\frac{1}{2} tr\left[(Y - XB)(\Sigma)^{-1}(Y - XB)'\right].$$

Se sustituye en $f_Y(y : B, \Sigma)$, y se aplica logaritmo a la verosimilitud. Se tiene así

$$\ell(B, \Sigma) = -\frac{n}{2} ln|2\pi\Sigma| - \frac{1}{2} tr\left[(Y - XB)(\Sigma)^{-1}(Y - XB)'\right]$$

El método de máxima verosimilitud ofrece como estimador de los parámetros aquel valor que maximiza $\ell(B, \Sigma)$.

Se asume que $n \geq p + q$ y que X es de rango completo q por lo tanto $(X'X)^{-1}$ existe. Sea $P = I_n - X(X'X)^{-1}X'$, una matriz $P(n \times n)$, simétrica idempotente de rango $(n - q)$ que se proyecta sobre un subespacio de $\mathbb{R}^n$ ortogonal al espacio columna de X (es decir, $PX = 0$), se tiene el siguiente resultado.

Teorema 1: Los estimadores de máxima verosimilitud de los parámetros del modelo de regresión están dados por

$$\hat{B} = (X'X)^{-1}X'Y$$
$$\hat{\Sigma} = n^{-1}Y'PY.$$

Demostración:
Sea $\hat{Y} = X\hat{B}$, consideremos la siguiente igualdad $\hat{B} = (X'X)^{-1}X'Y$, sustituimos en $\hat{Y}$ y se obtiene que:

$$\hat{Y} = X\hat{B} = X(X'X)^{-1}X'Y$$

y la matriz residual es

$$\hat{U} = Y - \hat{Y} = PY \tag{2.2}$$

Entonces

$$Y - XB = \hat{U} + X\hat{B} - XB = \hat{U} + X(\hat{B} - B) \tag{2.3}$$

Sustituyendo las igualdades (2.2) y (2.3) en la función de verosimilitud $f_Y(y : B, \Sigma)$ queda de la siguiente manera:

$$\ell(B,\Sigma) = -\frac{n}{2}ln|2\pi\Sigma| - \frac{1}{2}tr\left[(\hat{U} + X(\hat{B} - B))\,(\Sigma)^{-1}\,(\hat{U} + X(\hat{B} - B))'\right]$$

por la propiedad 2.2 de la traza de un producto, se tiene

$$\ell(B,\Sigma) = -\frac{n}{2}ln|2\pi\Sigma| - \frac{1}{2}tr\left[(\Sigma)^{-1}\,(\hat{U} + X(\hat{B} - B))'(\hat{U} + X(\hat{B} - B))\right]$$

se resuelve el producto de vectores

$$(\hat{B} - B)'(\hat{U}X)' + (\hat{U}'X)(\hat{B} - B) = 0 \;\text{ ya que }\; \hat{U}'X = Y'PX = 0$$

Así,

$$\ell(B,\Sigma) = -\frac{n}{2}ln|2\pi\Sigma| - \frac{1}{2}tr\left[(\Sigma)^{-1}\,(\hat{U}'\hat{U} + (\hat{B} - B))'X'X(\hat{B} - B))\right],$$

por hipótesis

$$\hat{\Sigma} = n^{-1}Y'PY = n_{-1}\hat{U}'\hat{U}$$

sustituyendo $\hat{U}'\hat{U} = n\hat{\Sigma}$ en $\ell(B,\Sigma)$, se obtiene que

$$\ell(B,\Sigma) = -\frac{n}{2}ln|2\pi\Sigma| - \frac{1}{2}tr\left[(\Sigma)^{-1}\,(n\hat{\Sigma} + (\hat{B} - B))'X'X(\hat{B} - B))\right]$$

Finalmente se tiene

$$\ell(B,\Sigma) = -\frac{n}{2}ln|2\pi\Sigma| - \frac{n}{2}tr\left[(\Sigma)^{-1}\,(\hat{\Sigma})\right] - \frac{1}{2}tr\left[(\Sigma)^{-1}\,(\hat{B} - B)'X'X(\hat{B} - B)\right]$$

Se puede observar que solo el último término involucra B y ℓ es maximizada cuando $B = (\hat{B})$. Por lo tanto la función de verosimilitud se reduce a

$$\begin{aligned}\ell(B,\Sigma) &= -\frac{n}{2}ln|2\pi\Sigma| - \frac{n}{2}tr\left[(\Sigma)^{-1}\,(\hat{\Sigma})\right] \\ &= -\frac{np}{2}ln2\pi - \frac{n}{2}ln|\Sigma| - \frac{n}{2}tr\left[(\Sigma)^{-1}\,(\hat{\Sigma})\right].\end{aligned}$$

□

Antes de continuar revisaremos el siguiente teorema:

Teorema 2: Para cada matriz fija $A > 0$

$$f(\Sigma) = \frac{1}{|\Sigma|^{n/2}} exp\left[\frac{-1}{2} tr\left(\Sigma^{-1}A\right)\right]$$

es maximizada a través de $\Sigma > 0$ por $\Sigma = n^{-1}A$ y

$$f(n^{-1}A) = |n^{-1}A|^{-n/2} exp\left[-\frac{n}{2}\right].$$

La demostración del teorema anterior pueder verse (Mardia, 1995).

De acuerdo con el teorema 2 la función $\ell(B, \Sigma)$ se maximiza en $\Sigma = \hat{\Sigma}$.

El máximo valor de la función de log-verosimilitud está dada por:

$$\ell(\hat{B}, \hat{\Sigma}) = -\frac{n}{2} ln|2\pi| + ln|\hat{\Sigma}|^{np/2} = -\frac{n}{2} ln|2\pi\hat{\Sigma}| - \frac{np}{2}$$

Se puede mostrar que el estimador de máxima verosimilitud es insesgado. Es decir,

$$E(\hat{B}) = B$$

□

Análisis de varianza multivariante

El análisis de varianza multivariante (MANOVA) es una extensión del modelo ANOVA, ahora se trabaja con un modelo que contiene más de una variable dependiente. El análisis de varianza multivariante es una técnica que tiene en cuenta todas las variables dependientes de forma simultánea, lo que implica, que la correlación entre ellas, se tiene en cuenta en el análisis. El análisis mediante MANOVA contrasta una sola hipótesis, que las medias de los n grupos son iguales en las k variables dependientes, es decir;

$$H_0 : \begin{pmatrix} \mu_{11} \\ \vdots \\ \mu_{1n} \end{pmatrix} = \begin{pmatrix} \mu_{i1} \\ \vdots \\ \mu_{in} \end{pmatrix} = \begin{pmatrix} \mu_{k1} \\ \vdots \\ \mu_{kn} \end{pmatrix}$$

El análisis multivariado de varianza es una técnica que tiene la ventaja de permitirnos contrastar hipótesis sobre los efectos de los tratamientos, permite también determinar la importancia de cada variable dependiente en el efecto observado. Por analogía al modelo ANOVA, se considera que la matriz de sumas totales (T) se obtiene sumando la matriz de sumas de cuadrados de la regresión (B) y la matriz de suma de cuadrados de los residuales (W):

$$T = B + W$$

Por lo anterior la matriz de suma de los cuadrados de la regresión B se puede expresar como $B = T - W$, que contiene en la diagonal principal las sumas de cuadrados y fuera de la diagonal las sumas de productos entre tratamientos.

En ANOVA, para calcular el estadístico de contraste se divide la variabilidad entre grupos, debida a los tratamientos, por variabilidad error. Estas variabilidades en MANOVA vienen dadas por B y W, respectivamente. Sin embargo, no pueden dividirse entre sí porque la división entre matrices no existe, pero se puede realizar obteniendo la inversa de W y multiplicando por B, es decir; $W^{-1}B$. Esta matriz será el equivalente de la división de MS_{REG}/MS_{RES} necesaria para determinar la F de Snedecor en la aproximación univariada.

Los valores propios de la matriz $W^{-1}B$ proporcionan los cocientes entre MS_{REG} y MS_{RES} en los variados. Por tanto, se determinan los valores propios (o autovalores) de la matriz resolviendo la ecuación siguiente:

$$\left|W^{-1}B - \lambda I\right| = 0$$

Para tomar decisiones sobre el efecto de la variable independiente se utiliza el siguiente estadístico *La traza de Pillai*.
La traza de Pillai puede interpretarse como la suma de las varianzas explicadas por cada variable independiente en los variados. Su ecuación es:

$$V = \sum_{i=1}^{k} \frac{\lambda_i}{1 - \lambda_i} = traza[B(B + W)^{-1}]$$

Donde k es el número de variables dependientes del modelo.

Una vez que se han introducido algunos de los modelos lineales en los reales, a continuación se muestran los modelos que se utilizan cuando alguna de las variables en el modelo son variables composicionales.

2.3. Regresión composicional

En los modelos de regresión lineal simple, regresión lineal múltiple y regresión multivariable las variables dependientes e independientes pueden ser variables composicionales, en los 3 casos los parámetros de estos modelos son datos composiciones en el Símplex. Sin embargo, la mayoría de procedimientos en los modelos lineales composicionales tienen un análogo a los métodos lineales clásicos.

Se pueden distinguir 3 tipos de regresión composicional, cuando solo Y es una variable composicional, cuando X es una variable composicional y cuando ambas variables X y Y son composicionales.

A continuación se analizaran los diferentes modelos mencionados anteriormente.

2.3.1. Variable independiente composicional

Hay diferentes formas de generalizar una regresión lineal con una variable independiente composicional, estas generalizaciones son por lo general multivariables lineales o no lineales.

El mapeo que generaliza este modelo de regresión debe adecuarse convenientemente para composiciones y esto se logra mediante el producto escalar de Aitchison, jugando un papel importante en modelos de regresión con la forma:

$$Y_i = a + \langle b, X_i \rangle_A + \varepsilon_i \tag{2.4}$$

donde:
Y_i la variable dependiente, $a \epsilon \mathbb{R}$, b, X_i son vectores composicionales en $\mathcal{L}^D$ con b desconocido y ε_i el error aleatorio $\epsilon \mathbb{R}$ con distribución normal. Además, $\langle r, s \rangle_A$ el producto interior de Aitchison como de muestra en la Definición 1.7 y $g(.)$ la media geométrica de la Definición 1.6. Recordemos que $\langle b, X_i \rangle_A$ cumple con la siguiente propiedad (ver Tabla 1.3)

$$\langle b, X_i \rangle_A = \langle ilr(b), ilr(X_i) \rangle,$$

esto gracias a la propiedad de isometría de la transformación (ilr). Por lo tanto, se tiene

$$\langle ilr(b), ilr(X_i) \rangle \epsilon \mathbb{R},$$

donde $\langle . \rangle$ es el producto interior en los reales.

Por lo anterior el modelo (2.4) se puede expresar como

$$\begin{aligned} Y_i &= a + \langle b, X_i \rangle_A + \epsilon_i \\ &= a + \langle ilr(b), ilr(X_i) \rangle + \epsilon_i \\ &= a + \Sigma_{j=1}^{D-1} ilr_j(b) ilr_j(X_i) + \epsilon_i. \\ &= a + \Sigma_{j=1}^{D-1} \beta_j ilr_j(X_i) + \epsilon_i. \end{aligned}$$

Haciendo $\beta_j = ilr_j(b)$ en el último paso.

Se puede ver que el modelo resultante es un modelo de regresión lineal simple con variables reales. En consecuencia, se puede analizar de manera clásica y aplicar la transformación ilr^{-1} para estimar a $b \in \mathcal{L}^D$.

El vector composicional b proporciona la dirección en la cual X puede ser perturbada para tener mayores efectos en Y. Es decir, b representa el cambio de Y cuando X aumenta. Así, si se tienen dos observaciones composicionales X_i y X_j que difieren por un vector unitario $\frac{b}{\| b \|}$ en la dirección de b, $X_j = X_i + \frac{b}{\| b \|}$, entonces

$$\begin{aligned} E\left[Y_j|X_j\right] &= a + \langle b, X_j \rangle_A + E(\epsilon) \\ &= a + \langle b, X_i \oplus \frac{b}{\| b \|} \rangle_A \\ &= a + \langle b, X_i \rangle_A + \frac{1}{\| b \|} \langle b, b \rangle_A \\ &= E\left[Y_i|X_i\right] + \| b \| \end{aligned}$$

Es decir, el valor esperado de Y_j difiere del valor esperado de Y_i por $\| b \|$.
Para ejemplificar la teoría mencionada anteriormente se muestra el siguiente ejemplo.

Regresión con variable independiente composicional

Se está diseñando un nuevo sustrato especializado para fresas y se está interesado en entender cómo influye este sustrato en el total de biomasa de fresa producida por temporada.

Se entiende por biomasa a la materia total de los seres que viven en un lugar determinado, expresada en peso por unidad de área o de volumen, es decir, se está midiendo la cantidad de fresa en Kg que se produce por temporada.

Los datos que se analizan son los que se muestran en la Tabla 2.2:

Tabla 2.2. Datos de Fresas

	C1	C2	C3	biomasa
1	0.3333	0.3333	0.3333	12.4063
2	0.1666	0.6666	0.1666	6.0037
3	0.6666	0.1666	0.1666	4.3887
4	0.1666	0.1666	0.6666	13.7103
5	0.4444	0.4444	0.1111	4.9393
6	0.1111	0.4444	0.4444	11.6531
7	0.4444	0.1111	0.4444	11.2670

Se definen las variables X y Y como:

La variable independiente X es la variable composicional del modelo y la variable dependiente Y es una variable que se encuentra en $\mathbb{R}$.

- X: Sustrato para las fresas, donde el sustrato está compuesto por 3 componentes que se definen como C1, C2, C3.
- Y: Cantidad de biomasa en Kg producida por temporada.

Se toma el log de la biomasa de las fresas ya que estos datos solo pueden ser positivos y la producción puede crecer de manera exponencial con luz y agua adecuada.

En la Figura 2.1 se observa el diagrama ternario que muestra la relación que existe entre los componentes del sustrato y la cantidad de fresa producida por temporada.
El grado de dependencia entre las fresas y los componentes del sustrato se mide con respecto al tamaño del círculo, si el radio del circulo crece, nos indica que la dependencia es mayor y viceversa, por este motivo se puede observar que la cantidad de fresas tiende a ser mayor cuando el sustrato tiene una cantidad mayor del tercer componente $C3$.
El propósito de este ejemplo es modelar la relación existente entre los componentes del sustrato y la biomasa de las fresas, por esta razón se hace uso del análisis de regresión para datos composicionales.

En este problema el modelo a utilizar es aquel en el que la variable explicativa X es una variable composicional. El modelo correspondiente es:

$$Y_i = a + \langle b, X_i \rangle_A + \varepsilon_i$$

$$Fresas_i = a + \langle b, Sustrato_i \rangle_A + \varepsilon_i$$

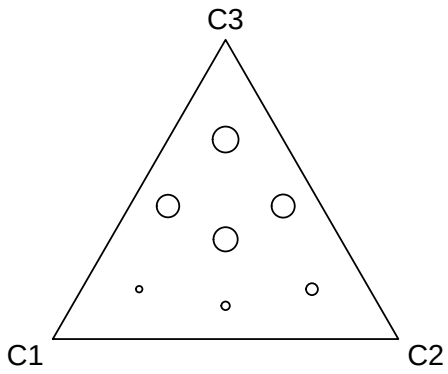


Figura 2.1. Relación entre sustrato y cantidad de fresa.

Se transforma el modelo para poder hacer análisis de regresión en $\mathbb{R}$, con ayuda de la isometría de la transformación (ilr) (Def. 1.20), se obtiene un modelo de regresión múltiple

$$Fresas_i = a + \Sigma_{j=1}^{D-1} ilr(b)_j ilr_j(Sustrato_i) + \epsilon_i$$

renombrando $ilr(b) = \beta$

$$= a + \Sigma_{j=1}^{D-1} \beta_j ilr_j(Sustrato_i) + \epsilon_i$$

Al ajustar el modelo de regresión se obtiene que los valores de los parámetros estimados correspondientes son:

$$\widehat{a} = 2.1261$$

Y el vector del parámetro estimado b, está dado por

$$\widehat{b} = (\ 0.2387,\ 0.2706,\ 0.4907\ )$$

El cual proporciona la dirección en el que la variable composicional $X = Sustrato$ debe ser perturbada para lograr mayores efectos en la variable $Y = Fresa$. Así, el modelo estimado está dado por

$$\widehat{Fresas}_i = 2.1261 + \langle (0.2387, 0.2706, 0.4907), Sustrato_i \rangle_A$$

Para comprender estos resultados se muestra en la Figura 2.2 el correspondiente diagrama ternario.

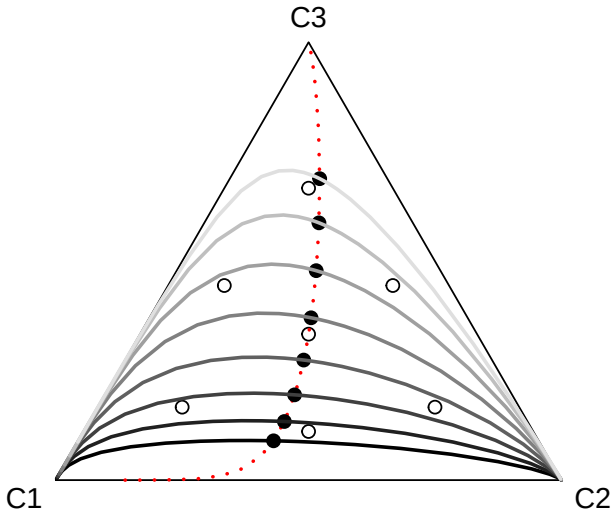


Figura 2.2. Líneas que corresponden a los datos composicionales

En la figura 2.2, se pueden identificar los datos originales en círculos, así como su representación gráfica de la forma $Y = f(X)$, las líneas corresponden a todos los datos composicionales $X's$ que satisfacen la variable dependiente Y. La sucesión de puntos rojos, nos indica la media de los puntos en una dirección b. Hay que recordar que si tenemos dos observaciones X_i y X_j y estas difieren por un vector unitario $\frac{b}{\| b \|}$, en la dirección de b, entonces el valor esperado de Y_j, difiere del valor esperado Y_i por $\| b \|$. Una vez que se estima el vector de b, se analizará su significancia.

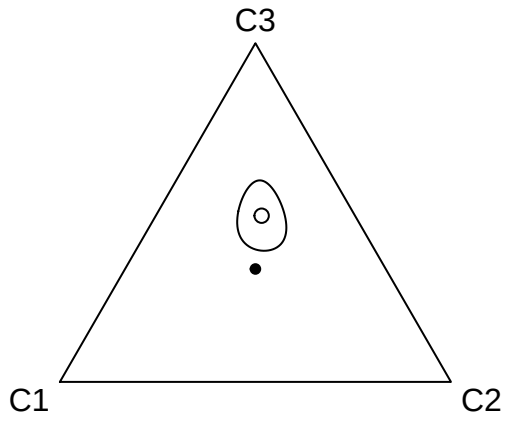


Figura 2.3. Elipse de confianza para la media del sustrato

En la Figura 2.3 el punto en negro corresponde al elemento neutro del Símplex, este punto representa a un vector que no puede inducir algún cambio en la respuesta biomasa de la fresa. La elipse es llamada elipse de confianza y en ella se encuentra un círculo, que corresponde al parámetro b. Como el elemento neutro no se encuentra dentro de la elipse de confianza del vector b, la cual se definirá más adelante, se asume con 95 de confianza, que existe una dependencia significativa entre la log-biomasa y la composición del sustrato.
Para corroborar este resultado se realiza un análisis de varianza, donde se obtiene,

Tabla 2.3. ANOVA para el sustrato.

```
            Df  Sum Sq Mean Sq F value  Pr(>F)
ilr(X)       2 1.13902 0.56951  8.6737 0.03511 *
Residuals    4 0.26264 0.06566

Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Por medio de la Tabla 2.3 se observa que la variable independiente X es significativa para el modelo, es decir, se asume con un nivel de confianza $(1-\alpha)$ con α=0.05 que el nuevo sustrato especializado para fresas es significativo en el total de biomasa de fresa producida por temporada.
Para analizar la normalidad de los datos y los residuales del modelo se presentan las siguientes gráficas:

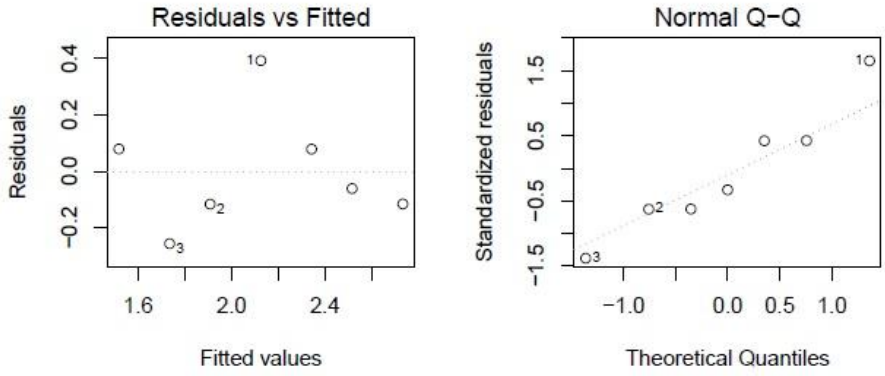


Figura 2.4. Normalidad y Residuales

En el panel (a) de la Figura 2.4 se observa que no existe un patrón repetitivo en los datos y que los residuales se encuentran alrededor del cero. De la grafica Normal $Q - Q$, que se muestra en el panel (b) de esta figura, se observa que no hay evidencia que indique que los datos no se puedan modelar con una distribución normal. Adicionalmente, para estos datos se realizó una prueba de Shapiro- Wilk con la que se obtuvieron los siguientes resultados.

```
data: residest
W = 0.90273, p-value = 0.3478
```

Recordemos que la hipótesis nula indica que la población está distribuida normalmente y la hipótesis alternativa que los datos no son distribuidos normalmente, como el p-valor es mayor que el nivel de significancia α= 0.05 no se rechaza la hipótesis nula.
Para averiguar que los errores cumplen el supuesto de la Homocedasticidad se hace una prueba de hipótesis basada en el coeficiente de Spearman.

```
Spearman's rank correlation rho
data: Predicted and abs(residest)
S = 76, p-value = 0.4444
rho
-0.3571429
```

El coeficiente indica que la correlación existente entre las variables es de −0.3571, existiendo una asociación negativa entre ellas.

El coeficiente de determinación del modelo ajustado es $R^2 = 0.72$, el cual es muy cercano a 1. Por lo tanto, se puede decir que existe un buen ajuste global del modelo.

2.3.2. Variable dependiente composicional

En el modelo de regresión con variable dependiente composicional una o más covariables reales explican a la variable dependiente composicional Y. En esta sección se construye un modelo de regresión con respuesta composicional y una variable independiente real. Es decir, se define el modelo

$$Y_i = a \oplus X_i \odot b + \varepsilon_i$$

donde
a, b son constantes composicionales desconocidas, Y_i es una composición aleatoria, X_i es una variable real aleatoria, ε_i una variable aleatoria composicional con esperanza $\mathbb{E} = \left(\frac{1, ..., 1}{D}\right)$ y varianza constante. Se considera que ε sigue una distribución lognormal en el Símplex $\mathcal{N}_s^D(\mathbb{E}, \Sigma)$ con matriz de covarianza Σ.

La intersección a se puede interpretar como la composición esperada cuando $X = 0$, la pendiente b puede ser interpretada como la perturbación aplicada a la composición si X es incrementada en una unidad.

Para trabajar con este modelo en $\mathbb{R}$, se necesita reescribirlo en un contexto de regresión multivariada basado en el principio de trabajar con coordenadas utilizando alguna de las transformaciones mencionadas anteriormente. Por lo tanto, al utilizar la transformación en una composición, ésta es representada como un vector real. Es decir,

$$Y_i \quad = \quad a \oplus X_i \odot b + \varepsilon,$$

se transforma en

$$ilr(Y_i) = ilr(a) + X_i ilr(b) + ilr(\varepsilon_i).$$

Renombrando a las variables

$$ilrY_i = Y', \; ilr(a) = a', \; ilr(b) = \beta \text{ y } ilr(\varepsilon_i) = \epsilon,$$

se obtiene una ecuación lineal multivariada en $\mathbb{R}$ como sigue,

$$Y_i' = a' + X_i.\beta + \epsilon \; con \; \epsilon_i \sim N(0_{D-1}, \Sigma_{ilr}).$$

Las elipses de confianza se pueden usar para realizar pruebas de hipótesis sobre los parámetros de un modelo de regresión. A continuación se revisa brevemente este enfoque.
Para generar las elipse de confianza hay que tener en cuenta que cualquier ecuación de segundo grado,

$$Ax^2 + 2Bxy + Cy^2 + Dx + Ey + F = 0,$$

puede escribirse en su forma matricial de la siguiente manera:

$$\begin{pmatrix} x & y \end{pmatrix}\begin{pmatrix} A & B \\ B & C \end{pmatrix}\begin{pmatrix} x \\ y \end{pmatrix}+\begin{pmatrix} D & E \end{pmatrix}\begin{pmatrix} x \\ y \end{pmatrix} + \mathrm{F} = 0.$$

Por lo anterior, se definen las elipses en el símplex.

Elipses en el Símplex

Para obtener elipses de confianza con respecto a la media se necesita un vector $x*$ de dimensión $D-1$ real, al cual se definirá como el vector de medias μ, $\mu = (\mu_1, ..., \mu_{D-1})$ y una matriz real definida positiva $\Sigma = (\sigma_{ij})$ los cuales determinan la ecuación de una elipse $\epsilon_{D-1}(x^*)$ con centro en μ,

$$\epsilon_{D-1}(x^*) = (x^* - \mu)\Sigma(x^* - \mu)^t = r^2, \quad r > 0. \tag{2.5}$$

Al realizar el producto de matrices, 2.5 se puede escribir como:

$$\epsilon_{D-1}(x^*) = \sum_{j=1}^{D-1}\sum_{i=1}^{D-1}\sigma_{ij}x_i^*x_j^* - 2\sum_{j=1}^{D-1}\sum_{i=1}^{D-1}\sigma_{ij}\mu_i x_j^* + c = 0, \tag{2.6}$$

con $c = \mu\Sigma\mu^t - r^2$.

Usando la descomposición espectral de la matriz Σ, se obtiene que la matriz Σ puede ser escrita de la siguiente manera:

$$\Sigma = \sum_{i=1}^{D-1}\lambda_i\nu_i^t\nu_i, \tag{2.7}$$

donde λ_i denotan los eigenvalores ortonormales de Σ y ν_i denotan los eigenvectores ortonormales de Σ.

Sustituyendo (2.7) en (2.5) se obtiene:

$$(x^* - \mu)\left(\sum_{i=1}^{D-1}\lambda_i\nu_i^t\nu_i\right)(x^* - \mu)^t = r^2.$$

Así en términos del producto interior Euclidiano, se tiene que

$$\sum_{i=1}^{D-1} \lambda_i \left(\langle \nu_i, x^* \rangle_E\right)^2 - 2 \sum_{i=1}^{D-1} \lambda_i \langle \nu_i, \mu \rangle_E \langle \nu_i, x^* \rangle_E + c = 0$$

Los vectores ν_i determinan la dirección de los ejes de la elipse y sus longitudes son determinadas por los eigenvalores λ_i. Haciendo uso de las propiedades de norma y producto interior de Aitchison mencionadas en la Definición 1.8 y 1.7 respectivamente, se obtiene la ecuación de una elipse en el Símplex D−dimensional $\epsilon_D^S(x)$

$$\epsilon_D^S(x) = \sum_{i=1}^{D-1} \lambda_i \left(\langle e_i, x \rangle_A\right)^2 - 2 \sum_{i=1}^{D-1} \lambda_i \langle e_i, \mu \rangle_A \langle e_i, x \rangle_A + c = 0$$

TEOREMA. La forma analítica de la elipse en el Símplex $\epsilon_D^S(x)$ está determinada por:

$$\sum_{i=1}^{D-1} \sum_{j=i+1}^{D} \sum_{k=1}^{D-1} \sum_{l=k+1}^{D} a_{ijkl} ln\frac{x_i}{x_j}\frac{x_k}{x_l} + \sum_{i=1}^{D-1} \sum_{j=i+1}^{D} b_{ij} ln\frac{x_i}{x_j} + c = 0,$$

donde

$$a_{ijkl} = \frac{1}{D^2} \sum_{m=1}^{D-1} \lambda_m ln\frac{e_{mi}}{e_{mj}} ln\frac{e_{mk}}{e_{ml}},$$

y

$$b_{ij} = \frac{-2}{D} \sum_{m=1}^{D-1} \lambda_m \langle e_i, \mu \rangle_A ln\frac{e_{mi}}{e_{mj}},$$

y

$$K = \sum_{i=1}^{D-1} \lambda_i (\langle e_m, \mu \rangle_A)^2 - c^2,$$

y

$$c = \sum_{i=1}^{D-1} \lambda_i (\langle e_m, \mu \rangle_A)^2 - c^2.$$

En la ecuación anterior μ representa el centro de la elipse en el Símplex y el vector $e_i = (e_{i1}, ..., e_{iD})$ la dirección de los ejes de la elipse.

Regiones de confianza

Cuando se trabaja con distribuciones normales multivariantes se tiene densidad constante sobre elipses o elipsoides tal como se mencionó con anterioridad

$$\epsilon_{D-1}(x^*) = (x^* - \mu)\Sigma(x^* - \mu)^t = r^2,$$

a los cuales se les conoce como contornos de la distribución, estos elipsoides están centrados en μ, es decir la media de los datos y la semilongitud de los ejes brindan información sobre los intervalos de confianza.

Las dimensiones de la elipse dependen directamente de las varianzas estimadas de los estimadores de los coeficientes de regresión:

- A mayor varianza, mayor será la elipse y viceversa.
- La inclinación de la elipse depende de la covarianza entre los dos coeficientes estimados.
- La elipse asciende de izquierda a derecha si tal covarianza es positiva, descendiendo en caso contrario.

Debido a que se trabaja con una distribución normal p- multivariante, la región que concentra el $100(1-\alpha)$ % de la distribución es una hiperelipsoide de dimensión p. La finalidad es exhibir un hiperrectángulo, el cual contenga la región de confianza de cada variable, es decir,

$$n(\bar{x} - \mu)\Sigma^{-1}(\bar{x} - \mu) \leq \frac{P(n-1)}{(n-p)} F_{p,n-p}(\alpha)$$

donde $\bar{x}$= ilr(x), p es el número de variables, n es el número de datos y $F_{p,n-p}(\alpha)$ es el quantil de una distribución de Fisher con p y $n-p$ grados de libertad, respectivamente. Así

$$n(\bar{x} - \mu)\Sigma^{-1}(\bar{x} - \mu) \leq c^2 = \frac{P(n-1)}{(n-p)} F_{p,n-p}(\alpha),$$

y se obtiene la siguiente desigualdad

$$\sqrt{(\bar{x} - \mu)\Sigma^{-1}(\bar{x} - \mu)} \leq \frac{c\sqrt{\lambda_i}}{\sqrt{n}} = \sqrt{\lambda_i}\sqrt{\frac{P(n-1)}{n(n-p)} F_{p,n-p}(\alpha)}.$$

De esta manera, los intervalos para el vector medio μ se obtienen mediante:

$$\textit{límite}_i = \sqrt{\lambda_i}\sqrt{\frac{P(n-1)}{n(n-p)} F_{p,n-p}(\alpha)},$$

y los intervalos correspondientes quedan definidos como:

$$[media_i \pm límite_i]$$

Donde, se cuenta con los intervalos para la media de μ en $\mathbb{R}^{D-1}$. Para poder interpretar estos intervalos en el Símplex, se debe aplica la transformación inversa, ilr^{-1}, tanto al límite inferior como al límite superior del intervalo obtenido.

Elipses de confianza para parámetros

Así como las elipses ayudan a encontrar una región de confianza para el vector de medias, son de gran utilidad para obtener regiones de confianza para los parámetros del modelo de regresión. Se obtiene la elipse de confianza para el parámetro correspondiente y por medio del gráfico asociado se observa si el parámetro es significativo o no para el modelo.

Para poder corroborar los resultados que se obtienen de las elipses de confianza, se utiliza prueba de hipótesis para cada una de las variables utilizadas en el modelo de regresión.
El método de prueba de hipótesis que se emplea en este análisis es:

Hipótesis nula Ho: La $i-esima$ variable no tiene influencia en el modelo, dadas las $(i-1)$ variables restantes del modelo. Si la variable es una variable continua, se dice que la pendiente del coeficiente $b_g = \mathbb{E}$. Si es un factor, se dice la composición tiene la misma esperanza para todos los niveles.

Hipótesis alternativas H_1: la variable independiente tiene influencia en el modelo, es decir, el coeficiente $b_g \neq \mathbb{E}$.
Es decir;

$$\begin{aligned} H_0 &: \{b_1 = ... = b_m = 0\} \\ H_1 &: \left\{\text{Al menos uno } \ b_g \neq 0\right\} \end{aligned}$$

El p-valor de la prueba ayuda a concluir que una variable puede ser removida si su influencia es no significativa, de esta manera se puede obtener un modelo parsimonioso, esto se debe a que no se puede asegurar que el parámetro sea diferente de cero.
Éste tipo de elipse es de gran utilidad para efectuar contrastes de hipótesis, al examinar si el punto establecido en dicha hipótesis se encuentra dentro o fuera de la región de confianza.
Si dicho punto está dentro, no se rechaza la hipótesis nula y si está fuera se rechaza.

Valores predichos y observados

Una vez que los parámetros son estimados, éstos se evalúan en el modelo sin el término error, de esta forma se obtiene

$$\hat{y} = x\hat{\beta}$$

Residuales

Los residuales son aquella diferencia existente entre los valores predichos $\hat{y}_i$ y los valores observados y_i

$$r_i = Y_i \ominus \hat{Y}_i = ilr^{-1}(ilr(Y_i) - ilr(\hat{Y}_i))$$

Los residuales son de gran ayuda para revisar los supuestos de un modelo de regresión: errores distribuidos normalmente y que sean homocedasticos.
Con la finalidad de tener una mejor comprensión sobre el modelo revisado se presenta el siguiente ejemplo, donde se analiza un modelo de regresión donde la variable dependiente es composicional.

Ejemplo con una variable dependiente composicional

La petrología se encarga del estudio de las rocas, su origen y su evolución, en el siguiente ejemplo se usa un conjunto de datos de proporciones de cuatro tipos petrográficos de granos de sedimento, que conforman la variable composicional Y, el modelo que se analizará es el siguiente:

$$Y_i = a \oplus Xgrano_i \odot b_1 \oplus Xposicion_i \odot b_2 \oplus Xdescarga_i \odot b_3 \oplus Xrelieve_i \odot b_4 + \varepsilon_i$$

Donde la variable composicional esta conformada por:

- Y_{1i} fragmentos de roca poliminerálica Rf,
- Y_{2i} granos de un solo cuarzo de cristal Qm
- Y_{3i} granos que contienen muchos cristales de cuarzo Qp.
- Y_{4i} granos de Mica M.

Los datos fueron tomados de Aitchison, 1986. El siguiente problema tiene como finalidad determinar la proporción que contiene un grano de roca $Xgrano$ de cada uno de los componentes de la variable composicional Y, de acuerdo a su posición $Xposicion$.
Así, las variables explicativas X_s del modelo son:

- $Xgrano$ representa el tamaño del sedimento $(fino, medio, aspero)$

- *Xposicion*: es una variable dicotómica que describe si el río pertenece a la parte norte o sur de la cuenca de drenaje.

Se definen las siguientes covariables continuas:

- *Xdescarga* es una covariable continua que describe el fluido del volumen anual de agua de rio por área,

- *Xrelieve* es la pendiente del río en tanto por ciento.

Sin olvidar que se trabaja con una mezcla de covariables continuas y categóricas, se analizará la dependencia de estas variables, explicativas y de respuesta composicional.

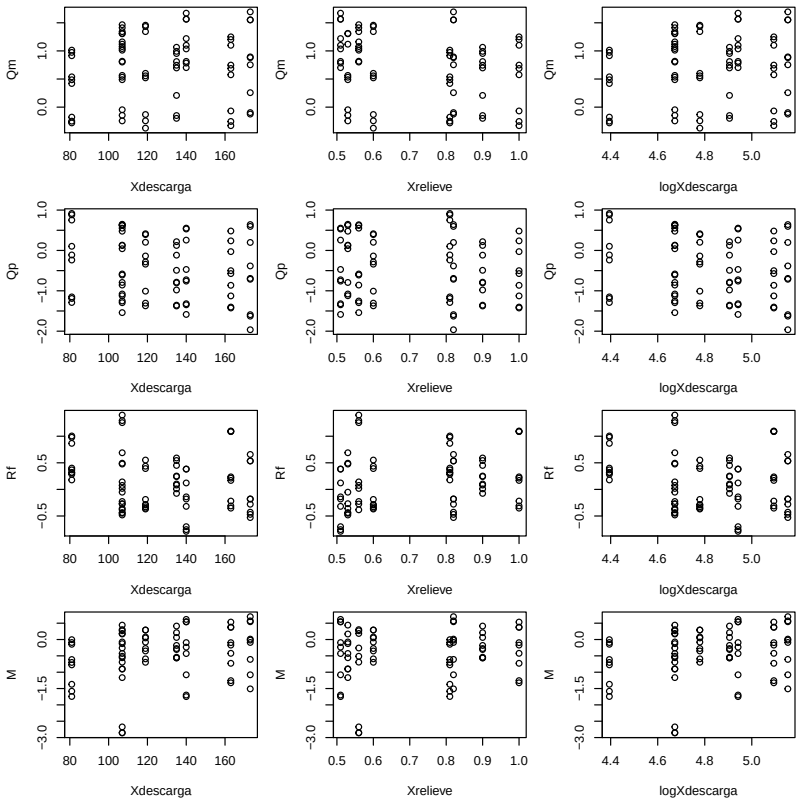


Figura 2.5. Relación entre variables

La Figura 2.5 muestran la relación de cada componente de Y, con respecto a las covariables *Xdescarga, Xrelieve y Log(Xdescarga)* debido a que esta cantidad es siempre positiva y puede crecer de manera exponencial,

Para poder analizar el modelo de regresión en $\mathbb{R}$, se aplica la transformación *ilr* al modelo de regresión composicional, el cual queda de la siguiente manera:

$$Y_i = ilr(a)+ilr(Xgrano_i.b_1)+ilr(Xposicion_i.b_2)+ilr(Xdescarga_i.b_3)+ilr(Xrelieve_i.b_4)+ilr(\varepsilon_i)$$

Se resuelve el modelo de regresión en el espacio de los reales, utilizando R-Project, y se aplica la transformación ilr^{-1} a los resultados, para obtener los parámetros en el Símplex, los cuales se muestran en la Tabla 2.4.

Cada uno de los parámetros de las covariables son la perturbación aplicada a la composición de las rocas si Y aumenta una unidad.

Tabla 2.4. Parámetros estimados del modelo Composicional.

	Variable	Qm	Qp	Rf	M
a	(Intercept)	0.0695	0.0565	0.8738	1.594e-05
b1	log(Xdescarga)	0.1541	0.0887	0.0487	7.083e-01
b2	Xrelieve	0.0753	0.1515	0.6283	1.446e-01
b3	Xgranomedium	0.1198	0.4961	0.2692	1.148e-01
b3	Xgranocoarse	0.0329	0.6429	0.3009	2.319e-02
b4	Xposicionsouth	0.3481	0.2141	0.2978	1.398e-01

Posteriormente se aplica al modelo el análisis de varianza multivariable, ver Tabla 2.5, con la cual permitirá probar estadísticamente que el modelo está bien estructurado, al analizar la significancia de cada una de las covariables.

Tabla 2.5. Análisis de Varianza del modelo Multivariable.

	Df	Pillai	approx F	num Df	den Df	Pr(>F)	
(Intercept)	1	0.93194	292.115	3	64	< 2.2e-16	***
log(Xdescarga)	1	0.51149	22.337	3	64	5.182e-10	***
Xrelieve	1	0.72312	55.717	3	64	< 2.2e-16	***
Xgrano	2	1.15063	29.351	6	130	< 2.2e-16	***
Xposicion	1	0.38186	13.179	3	64	8.435e-07	***
Residuals	66						

Signif. codes: 0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

De acuerdo con la información de la Tabla 2.5, muestra que las covariables del modelo de estudio son significativas para el modelo, con un nivel de significancia mayor de 0.001 .

Posteriormente se muestra en la Figura 2.6 las elipses de confianza alrededor de la media de los datos.

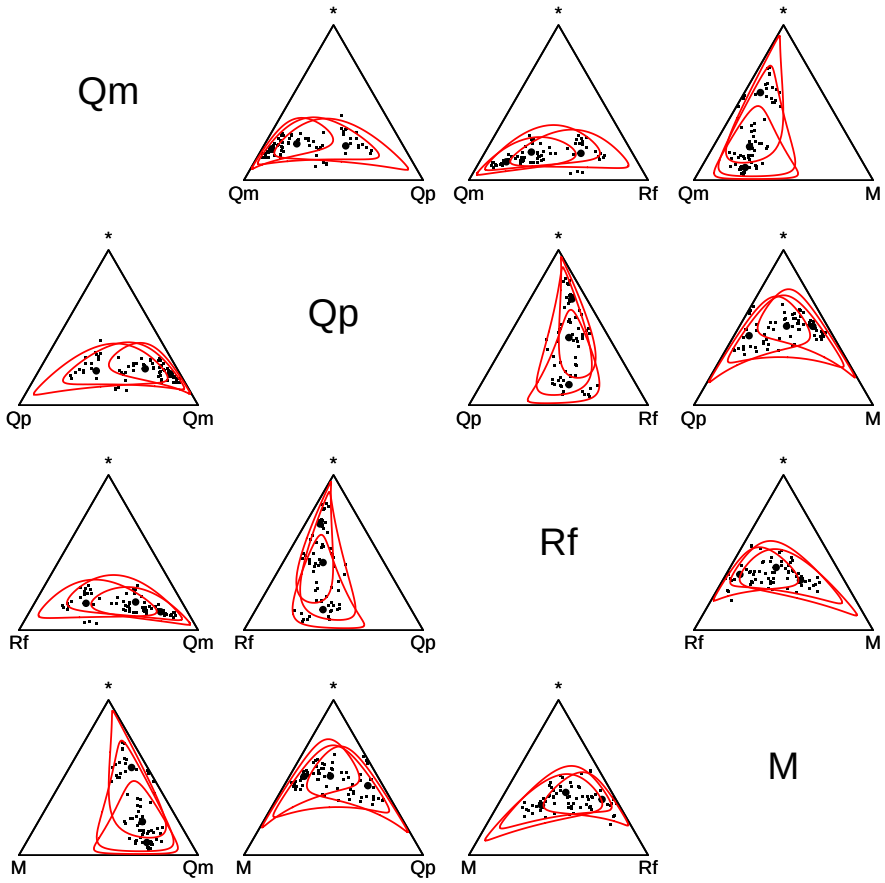


Figura 2.6. Elipses de confianza de los datos composicionales

Por la Figura 2.7 se puede comprobar que los datos siguen una distribución normal, debido a que los puntos están cerca de la recta.

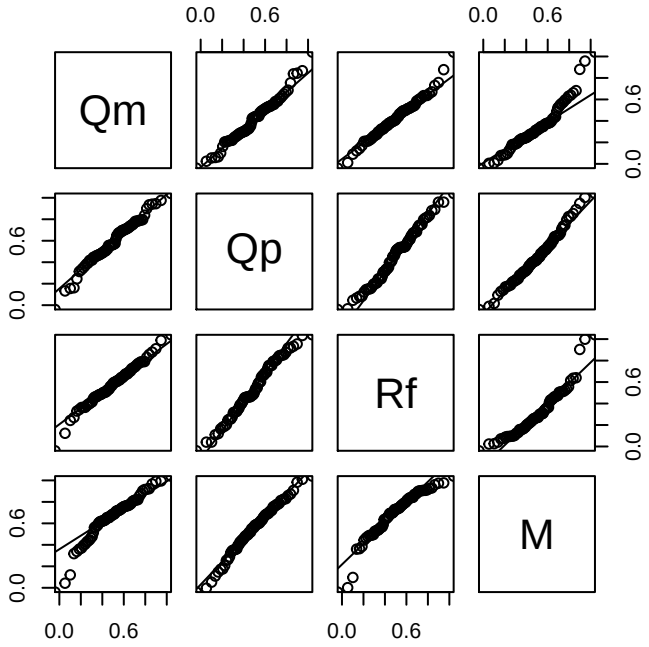


Figura 2.7. Normalidad de los datos transformados

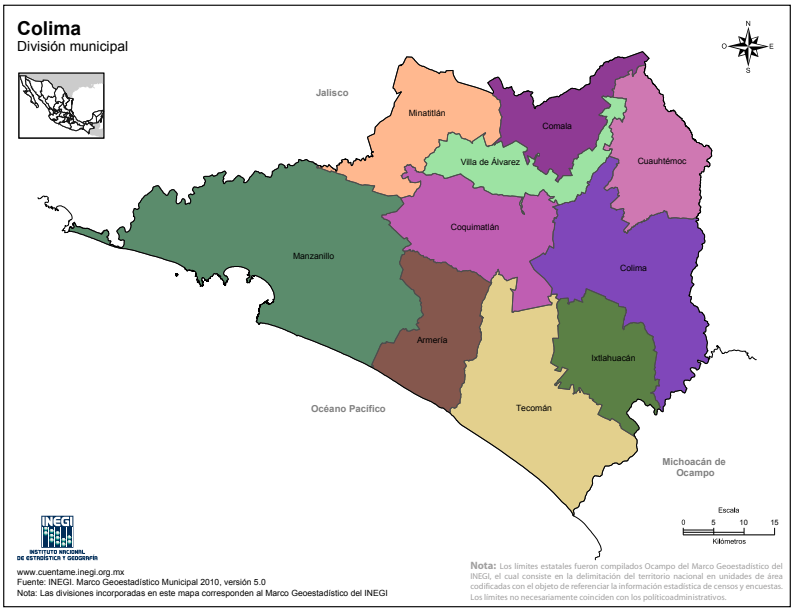
Capítulo 3

Un modelo composicional para describir datos de elecciones

En la actualidad existe gran interés en hacer pronósticos en el comportamiento de preferencias electorales, a través de votaciones a favor de partidos políticos de un determinado lugar. En general, los votos a favor de las fuerzas electorales en un ejercicio electoral, poseen restricciones que los vuelven factibles de analizar como datos composicionales. Lo anterior ya que la suma total de la proporción de votos de cada uno de los partidos políticos es uno y el conjunto de valores es positivo.

En este capítulo se muestra la manera de aplicar los conceptos discutidos en los capítulos anteriores. Particularmente se estudia un problema real referente a las elecciones ordinarias que se llevaron a cabo en el Estado de Colima el 7 de Junio del 2015, para Gobernador y Diputados Federales.

El estado de Colima se divide en 10 municipios: Armería, Colima, Cómala, Coquimatlán, Cuauhtémoc, Ixtlahuacán, Manzanillo, Minatitlán, Tecomán y Villa de Álvarez, que a su vez es dividido en 16 distritos. Un distrito es la división geográfica en que se organiza el territorio de un país con fines electorales, así todos los electores se ubican conforme a su domicilio en un distrito electoral. Cada distrito se divide en secciones, que corresponden a las casillas en donde los electores depositan su voto durante el día de la elección.



Como se mencionó anteriormente, en junio del 2015 se realizaron elecciones para Gobernador en el estado de Colima. Sin embargo, el Tribunal Electoral del Poder Judicial de la Federación (TEPJF) anuló la votación.

3.1. Problema

El propósito de este ejercicio es llevar a cabo estimaciones sobre el porcentaje de votos que obtendría cada partido político en las elecciones extraordinarias en el Estado de Colima en enero del 2016, para gobernador. Los partidos políticos con registro ante el Instituto Electoral del Estado fueron:

- PAN: Partido Acción Nacional
- PRI: Partido Revolucionario Institucional
- PRD: Partido de la Revolución Democrática
- PT: Partido del Trabajo
- PV: Partido Verde Ecologista de México

- PM: Partido Movimiento Ciudadano
- PA: Partido Nueva Alianza
- MORENA: Partido Movimiento de Regeneración Nacional
- ES: Partido Encuentro Social
- PH: Partido Humanista

Los partidos Verde Ecologista de México, Nueva Alianza y del Trabajo, estuvieron en coalición con el PRI, en apoyo al candidato que registró este último para participar en la elección extraordinaria de gobernador.

3.1.1. Información sobre la muestra

Los datos fueron recabados de la página del Instituto Electoral del Estado de Colima, esta base de datos consta de 903 registros, donde cada uno de ellos representa la información de las casillas con las que cuenta el estado.
Esta base de datos cuenta con el número de votos que obtuvo cada uno de los partidos políticos en las elecciones para Diputados Federales en junio de 2015. El número de votos obtenido por cada partido es considerado como una variable independiente, es decir, las variables (X_S) del modelo de regresión. En esta base también se encuentra el número de votos que obtuvo cada partido político en la elección ordinaria para Gobernador, estos datos representaran la variable dependiente (Y) del modelo.
Las variables que representan cada una de las variables independientes del modelo, es decir, los votos para diputados, son:

- $PAN.x$
- $PRI.x$
- $PRD.x$
- $PVEM.x$
- $PT.x$
- $MC.x$
- $PANAL.x$

- $MORENA.x$
- $PH.x$
- $ES.x$
- $C\_PRI\_PVEM.x$
- $NO\_REGISTRADOS.x$
- $NULOS.x$

Hay que mencionar que

El Distrito I Electoral Federal de Colima está formado por los municipios de Colima, Cómala, Coquimatlán, Cuauhtémoc, Ixtlahuacán y Villa de Álvarez. Por otra parte, el Distrito II Electoral Federal de Colima está formado por los municipios de Armería, Manzanillo, Minatitlán y Tecomán.
Por su parte, las variables que representan los votos en la elección ordinaria para gobernador, son los siguientes.

- $PAN.y$
- $PRI.y$
- $PRD.y$
- $PVEM.y$
- $PT.y$
- $MC.y$
- $PANAL.y$
- $MORENA.y$
- $PH.y$
- $ES.y$
- $C\_PRI\_PVEM.y$
- $NO\_REGISTRADOS.y$
- $NULOS.y$

Se debe notar que tanto el vector de votos para Diputados como el vector de votos para Gobernador, se pueden considerar como datos composicionales. De la base de datos conjunta se tomó una muestra de 350 registros. Con la muestra anterior se ejemplificará el ajuste de un modelo de regresión, donde la variable de respuesta sea la variable composicional votos para gobernador en el Estado de Colima.
Lo anterior, con el objetivo de estimar la verdadera cantidad de votos obtenida por cada partido político en las elecciones para gobernador del Estado de Colima en junio de 2015.

3.1.2. Especificación del modelo

Como se mencionó con anterioridad el modelo composicional con el que se trabaja en este ejemplo, es aquel en el que la variable dependiente es una variable composicional. Sea

$$PV_{ji} : \text{La proporción de votos del partido } i \text{ de la } j-\acute{e}sima \text{ casilla,}$$
$$i = 1, ..., 13 \; y \; j = 1, ..., 350.$$

Se debe notar que la proporción de votos de cada uno de estos partidos es positiva y, que el cambio en uno de ellos afecta inversamente al número de votos del resto de los partidos. Con esto se puede asegurar que:

$$PV_{ji} \; \epsilon \; [0, 1]$$

Donde i representa los 10 partidos políticos mencionados anteriormente, la coalición, No registrados y los votos Nulos. Se sabe que la suma de todos los votos es el total de las personas votantes. Así,

$$\sum_{i=1}^{13} PV_{ji} = 1$$

Por lo revisado en el Capítulo 2, el modelo a utilizar cuando la variable dependiente es composicional se expresa de la siguiente forma.

$$Y_j = a \oplus X_j \odot b + \varepsilon_j$$

Es decir, ahora se tiene el siguiente problema

$$\left(PV_{j1}, PV_{j2}, ..., PV_{j13}\right) = \alpha_0 + X_{j1}b_1 + ... + X_{j14}b_{14} + \epsilon$$

Con $j = 1, ..., n$ donde n es el número de casillas votantes composicionales de los partidos políticos y 14 son los votos para diputado de cada partido político, coalición, No registrados, los votos Nulos y el distrito electoral al que pertenecen. En términos de los partidos políticos que participaron en las elecciones, se tiene

$$\begin{aligned}(PAN.y_j, PRI.y_j, ..., NULOS.y_j) \;\; = \;\; & a \oplus PAN.x_j \odot b_1 \oplus ... \oplus NULOS.x_j \odot b_{13} \\ & \oplus DIST\_FED_j \odot b_{14} + \varepsilon_j\end{aligned}$$

3.1.3. Ajuste del modelo

Como primer paso para analizar el modelo de regresión anterior, es necesario transformar el modelo a R^{D-1}, donde D es el número de componentes de la variable composicional. Para lograr lo anterior, como se discutió en el Capítulo 2, Definición 1.20, se aplica la transformación logcociente isométrica (ilr), es decir,

$$\begin{aligned} ilr(PAN.y_j, PRI.y_j, ..., NULOS.y_j) &= a \oplus PAN.x_j \odot ilrb_1 \oplus ... \oplus NULOS.x_j ilrb_{13} \\ &\quad \oplus DIST\_FED_j ilrb_{14} + \varepsilon_j \end{aligned}$$

Para poder visualizar la metodología discutida en capítulos anteriores, se ejemplifica un modelo reducido en 3 dimensiones. Lo anterior, para trabajar en el Símplex $\mathcal{L}^3$ y poder realizar ciertas visualizaciones.

Análisis de un modelo particular en $\mathcal{L}^3$

En esta sección se analiza una parte del modelo de votaciones, este modelo composicional sigue conservando el mismo esquema que el anterior, el modelo que representa este tipo de regresión es:

$$Y_i = a \oplus X_i \odot b + \varepsilon_i$$

El modelo particular que se analizará en esta sección queda expresado de la siguiente manera:

$$\begin{aligned} (PAN.y_i, PRI.y_i, MC.y_i) &= a \oplus PAN.x_i \odot b_1 \oplus PRI.x_i \odot b_2 \oplus MC.x_i \odot b_3 \\ &\quad \oplus PVEM.x_i \odot b_4 \oplus DIST\_FED_i \odot b_5 \\ &\quad \oplus CASILLA_i \odot b_6 + \varepsilon_i \end{aligned}$$

La Figura 3.1 muestra la relación existente entre los 3 partidos políticos que conforman la variable dependiente composicional.

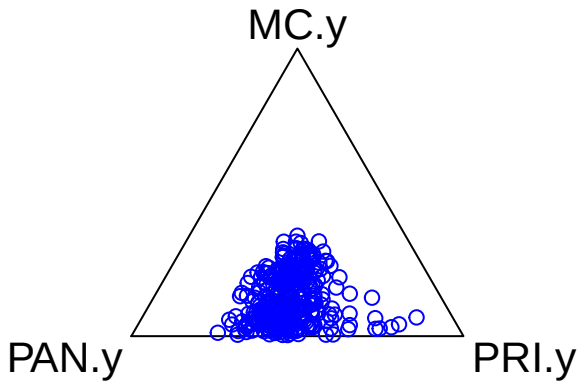


Figura 3.1. Relación de los 3 partidos políticos

El diagrama ternario de la Figura 3.1 muestra la dispersión de las variables composicionales compuestas por los partidos políticos *PAN*, *PRI y MC*. Se puede observar que la proporción de votos es muy similar entre los partidos *PAN* y *PRI*, existiendo una mínima ventaja para el partido revolucionario institucional, pues algunas de las casillas cuenta con una proporción mayor para éste partido. Lo anterior, se observa en el diagrama en el momento que los círculos comienzan a acercarse al vértice correspondiente al partido del *PRI*.
Empleando la transformación (ilr) el modelo anterior se puede llevar a un espacio real de dimensión 2. Así se pueden aplicar métodos como el de mínimos cuadrados para ajustar el modelo. Es decir,

$$\begin{aligned} ilr(PAN.y_i, PRI.y_i, MC.y_i) &= a \oplus PAN.x_i \odot ilrb_1 \oplus PRI.x_i \odot ilrb_2 \oplus MC.x_i \odot ilrb_3 \\ &\quad \oplus PVEM.x_i \odot ilrb_4 \oplus DIST\_FED_i \odot ilrb_5 \\ &\quad \oplus CASILLA_i \odot ilrb_6 + \varepsilon_i \end{aligned}$$

Se puede observar que ahora se tiene un modelo multivariante en el espacio $\mathbb{R}^3$. Una vez estimado el modelo, los parámetros estimados son llevados nuevamente al Símplex con la transformación $ilr^{-1}(b)$. Los resultados que se obtienen son los que se presentan a continuación.

```
                PAN.y  PRI.y  MC.y
(Intercept) 0.4296 0.4252 0.1450
PAN.x       0.3350 0.3323 0.3325
PRI.x       0.3327 0.3352 0.3320
```

```
MC.x        0.3247 0.3269 0.3482
PVEM.x      0.3337 0.3323 0.3338
DIST2       0.3990 0.4303 0.1705
CASILLA2    0.3356 0.3359 0.3284
CASILLA3    0.3397 0.3421 0.3181
CASILLA4    0.3009 0.3198 0.3792
CASILLA5    0.3174 0.3511 0.3313
CASILLA6    0.3023 0.3212 0.3763
CASILLA7    0.3226 0.3492 0.3281
CASILLA10   0.2949 0.2790 0.4260
```

El *intercept* = (0.4296, 0.4252, 0.1450) es el valor del parámetro a en el modelo;
Este es el valor a de la esperanza composicional cuando las variables X_s del modelo anterior son iguales a cero, el valor del gradiente b es la perturbación que se le aplica a cada cantidad de votos para gobernador cuando los votos para diputados incrementan en una unidad.

A continuación se muestran en la Figura 3.2 las elipses de confianza para los parámetros, para poder visualizar qué tanto los parámetros correspondientes influyen en el modelo.

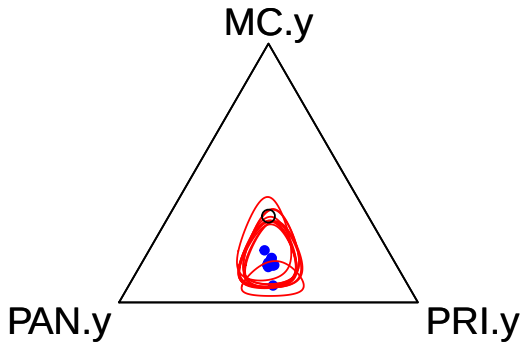


Figura 3.2. Elipses de confianza de los parámetros del modelo con 3 partidos como componentes.

Si la elipse de confianza correspondiente al $(1-\alpha)\times 100\,\%$ contiene al elemento neutro de la perturbación (punto negro) $\mathbb{E} = \left(\frac{1}{3}, \frac{1}{3}, \frac{1}{3}\right)$ el correspondiente parámetro está muy cercano o puede llegar a ser cero, así que en estos casos, el parámetro no influye en el modelo. Para verificar el resultado anterior, se realiza un análisis de varianza para revisar la significancia de las variables.

Tabla 3.1. Análisis de Varianza

	Df	Pillai	approx F	num Df	den Df	Pr(>F)	
(Intercept)	1	0.89954	1504.36	2	336	< 2.2e-16	***
PAN.x	1	0.52023	182.17	2	336	< 2.2e-16	***
PRI.x	1	0.53860	196.11	2	336	< 2.2e-16	***
MC.x	1	0.62629	281.54	2	336	< 2.2e-16	***
PVEM.x	1	0.12064	23.05	2	336	4.166e-10	***
DIST	1	0.37504	100.82	2	336	< 2.2e-16	***
CASILLA	7	0.01495	0.36	14	674	0.9843	
Residuals	337						

Signif. codes: 0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

El contraste de hipótesis que se analiza en este caso es:

$$\begin{aligned} H_0 &: \left\{ b_{g1} = ... = b_{gm} = 0 \right\} \\ H_1 &: \left\{ \text{Al menos uno } \; b_{gj} \neq 0 \right\} \end{aligned}$$

Como se observa en la Tabla 3.1, la variable $CASILLA$ tiene valor de p=0.9843 por lo que pude concluirse que no existe evidencia que demuestre que el tipo de casilla influya en el modelo de regresión por esta razón será eliminada del modelo.

Así, el modelo con el que se trabajará es el siguiente:

$$\begin{aligned} (PAN.y, PRI.y, MC.y) = \; & a \oplus PAN.x \odot b_1 \oplus PRI.x \odot b_2 \oplus MC.x \odot b_3 \\ & \oplus PVEM.x_i \odot b_4 \oplus DIST\_FED \odot b_5 + \varepsilon_i \end{aligned}$$

El siguiente diagrama muestra gráficamente como se distribuyen los votos de los 3 partidos dependiendo el distrito federal al que pertenecen. Como se muestra en la Figura 3.3, el Distrito I se representa con color azul (*Código* 1) y el Distrito II color rojo (*Código* 2). En este

conjunto de datos se puede ver que los votantes del distrito II optaron por dividir sus votos solo entre los partidos *PRI* y *PAN*, sin embargo el distrito I optó por tener votaciones divididas entre los 3 partidos políticos.

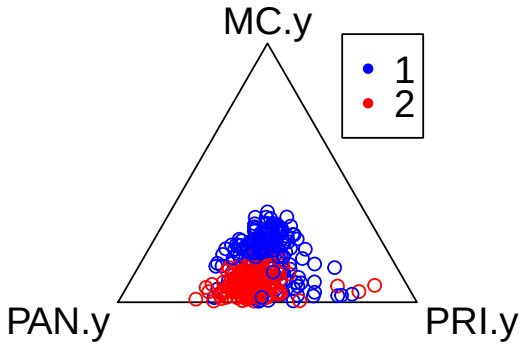


Figura 3.3. Representación de Distrito Federal Electoral

Los nuevos parámetros estimados se presentan en la Tabla 3.2 y se representan en un diagrama ternario como se muestra en la Figura 3.4:

Tabla 3.2. Parámetros estimados en el Símplex

	PAN.y	PRI.y	MC.y
(Intercept)	0.4293	0.4257	0.1449
PAN.x	0.3350	0.3323	0.3325
PRI.x	0.3327	0.3351	0.3320
MC.x	0.3247	0.3269	0.3483
PVEM.x	0.3337	0.3323	0.3338
DIST2	0.3995	0.4300	0.1704

Se visualiza nuevamente las elipses de confianza, para el nuevo modelo.

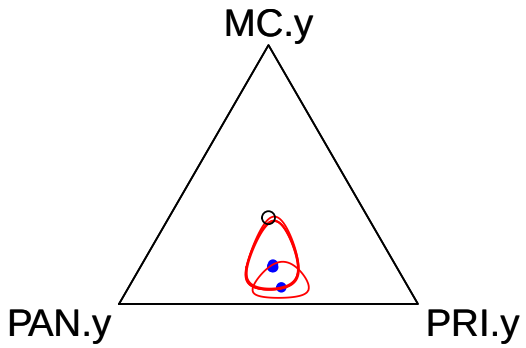


Figura 3.4. Elipses de confianza para nuevo modelo

Ahora es posible observar en la Figura 3.4 que ninguna de las elipses contienen al punto neutro, es decir todos los parámetros son significativos para el modelo y este resultado puede verificarse de manera más formal con los resultados presentados en la siguiente tabla.

Tabla 3.3. Análisis de Varianza

```
            Df  Pillai approx F num Df den Df    Pr(>F)
(Intercept)  1 0.89860  1519.79      2    343 < 2.2e-16 ***
PAN.x        1 0.51770   184.09      2    343 < 2.2e-16 ***
PRI.x        1 0.53767   199.45      2    343 < 2.2e-16 ***
MC.x         1 0.62378   284.35      2    343 < 2.2e-16 ***
PVEM.x       1 0.11983    23.35      2    343 3.116e-10 ***
DIST         1 0.37313   102.08      2    343 < 2.2e-16 ***
Residuals  344
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

Con lo anterior se concluye que todas las variables son relevantes para el modelo.
Una vez que encuentran los parámetros estimados, se obtiene el modelo estimado

$$\begin{aligned}\hat{Y}_j = (\hat{PAN}.y_j, \hat{PRI}.y_j, \hat{MC}.y_j) \quad = \quad & \hat{a} \oplus PAN.x_j \odot \hat{b_1} \oplus PRI.x_j \odot \hat{b_2} \oplus MC.x_j \odot \hat{b_3} \\ & \oplus PVEM.x_i \odot \hat{b_4} \oplus DIST\_FED_j \odot \hat{b_5}\end{aligned}$$

Se le llaman datos predichos a cada uno de los votos ($PA\hat{N}.y_j$, $PR\hat{I}.y_j$, $M\hat{C}.y_j$). En particular, el vector medio de los valores predichos resulta ser

PAN.y	PRI.y	MC.y
0.4739	0.4252	0.1008

Se estima que el porcentaje de votaciones para gobernador en el caso de que se tratara solo de los 3 partidos mencionados en el modelo sería la siguiente, la proporción de votos para el PAN sería mucho mayor con un porcentaje de votos de 47.39 % mientras que los del PRI 42.53 % y finalmente los del MC con una minoría de 10.08 %.(Ver Figura 3.5)

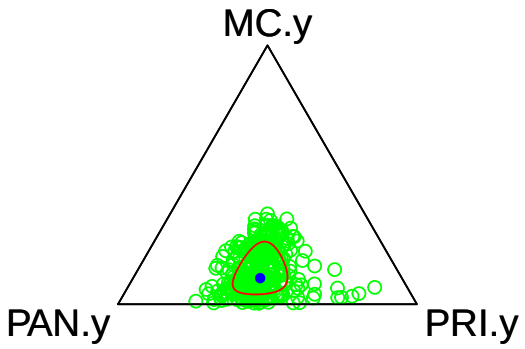


Figura 3.5. Concentración alrededor de la media de los valores predichos para cada partido.

En la Figura 3.6 se muestra la diferencia entre los valores predichos $\hat{Y}_j$ y los valores observados Y_j, es decir los residuales.

Nos interesa que estos valores sean muy pequeños, es decir que la diferencia entre los valores predichos y los estimados sea muy pequeña. En la Figura 3.6 se puede ver que la que la mayor parte de los datos en el diagrama anterior se concentran cerca del punto neutro de perturbación $\mathbb{E} = \left(\frac{1}{3}, \frac{1}{3}, \frac{1}{3}\right)$.

Al hacer análisis de la normalidad de los residuos, se utilizará un gráfico que muestra los cocientes de los residuos entre las variables, y posteriormente aplica una prueba de bondad de ajuste. A continuación se muestra la gráfica qqnorm correspondiente.

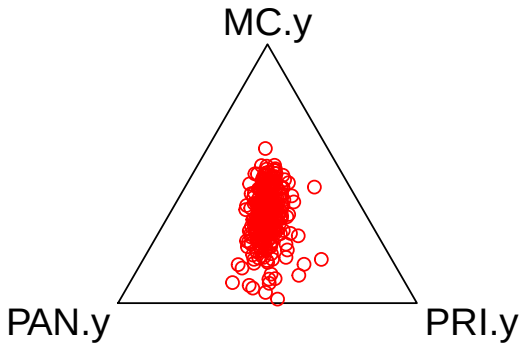


Figura 3.6. Residuos

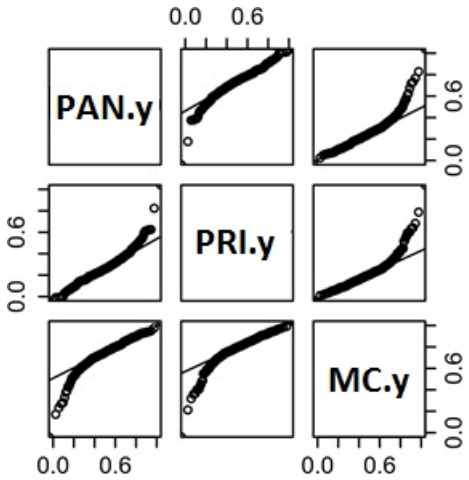


Figura 3.7. Normalidad de datos transformados

De la figura 3.7 se puede observar que las variables siguen una tendencia lineal, aunque algunos datos se alejan de la recta especialmente en los extremos. Otro análisis importante que se debe realizar es la homocedasticidad del modelo, este tipo de análisis ayuda a revisar si la varianza es constante o ésta varía, para este caso se hace uso de la transformación clr mencionada en la Definición 1.19, para no perder la dimensión de los datos.

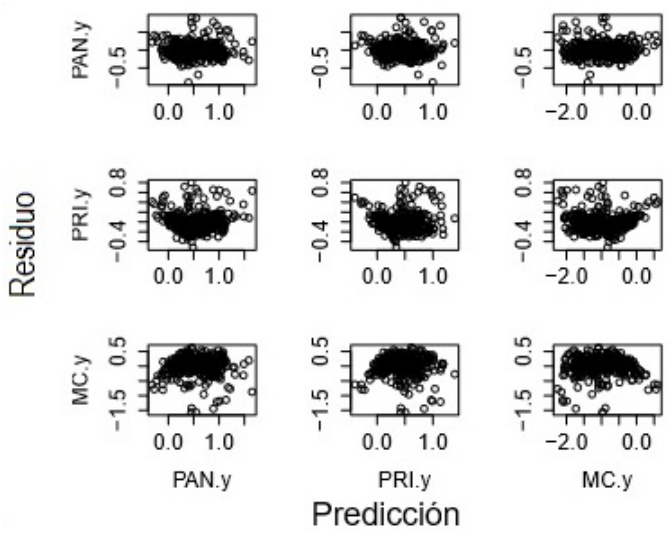


Figura 3.8. Homocedasticidad de los datos transformados

De la Figura 3.8 se puede apreciar que la varianza del error es constante a lo largo de los valores predichos, por lo que se puede concluir que sí existe homocedasticidad en el modelo. Para revisar la capacidad del modelo para explicar la variabilidad global de los datos se utiliza el coeficiente de determinación R^2. En este caso, se puede emplear ya que los datos transformados están en el espacio de los reales. Para el modelo ajustado se obtiene que R^2= 0.73.

Intervalos de confianza

Uno de los objetivos de este trabajo es dar a conocer un intervalo en el que se encuentre el verdadero porcentaje de votos alcanzado por cada uno de los partidos políticos en las elecciones a Gobernador en el Estado de Colima. Como se mencionó con anterioridad el procedimiento consiste en construir los intervalos correspondientes en el espacio real, posteriormente transformarlos a una región de confianza en el Símplex.

La representación de la región de confianza para dos variables correlacionadas y analizadas de manera conjunta, tienen forma de elipse, cuya inclinación depende de la correlación que exista entre las variables.
En este ejemplo, se toman los datos estimados de los 3 partidos políticos obtenidos anteriormente.

$$\hat{Y}_j = (P\hat{A}N.y_j, P\hat{R}I.y_j, M\hat{C}.y_j)$$

Se aplica la transformación *ilr* a $\hat{Y}_j$ para trabajar en el espacio de los reales, se obtiene una matriz de dimensión 350×2 y se calcula la matriz de covarianza de estos datos.

$$cov(ilr(\hat{Y}_j))=\begin{bmatrix}0.0429 & 0.0294\\ 0.0295 & 0.4820\end{bmatrix}$$

y el correspondiente vector de medias,

$$media(ilr(\hat{Y}_j))=\begin{bmatrix}-0.08 & -1.20\end{bmatrix}$$

Una vez que se cuenta con la matriz de covarianza, se utiliza la descomposición espectral, para obtener los valores propios de la matriz de covarianza Σ. En este caso, estos resultan ser

$$\lambda_1 = 0.0409 \quad \lambda_2 = 0.48402$$

Así, se obtiene la elipse de confianza, como se muestra en la Figura 3.9

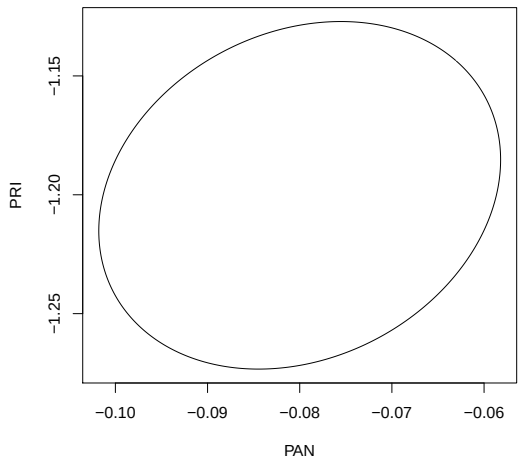


Figura 3.9. Elipse de confianza

En la Figura 3.9 se puede observar la elipse de confianza de los datos transformados. Una vez que se tienen los valores propios se aplica la teoría mencionada con anterioridad obtener los intervalos asociados,

$$límite_i = \sqrt{\lambda_i}\sqrt{\frac{p(n-1)}{n(p-n)} * F_{p,n-p}(\alpha)}$$

Por lo tanto los intervalos de confianza para $ilr(\bar{Y}_j)$, es decir, los intervalos de para cada uno de los partidos políticos en el espacio real son:

$$Intervalo = [media_j \pm límite_i] \quad con \quad j, i = 1, 2$$

$$\begin{aligned} \text{PAN} &\in \text{(-0.1012, -0.0587)} \\ \text{PRI} &\in \text{(-1.2734, -1.1269)} \end{aligned}$$

Para visualizar estos intervalos se muestra la Figura 3.10

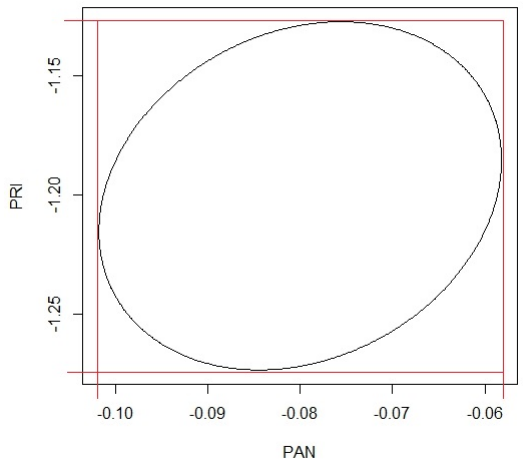


Figura 3.10. Intervalos de los partidos políticos

Como paso final se le aplica la transformación inversa (ilr^{-1}) a los intervalos para obtener los ejes de la elipse en el Símplex.

```
eje_1=(0.474, 0.423, 0.103)
eje_2=(0.485, 0.420, 0.095)
```

De acuerdo a la transformación se obtiene que los intervalos entre los que se encuentra la media en el Símplex son:

$$\begin{aligned} \text{PAN.y} &\in \text{(0.474,0.485)} \\ \text{PRI.y} &\in \text{(0.420, 0.423)} \\ \text{MC.y} &\in \text{(0.103,0.095)} \end{aligned}$$

Una vez que se ha presentado un análisis marginal con solo 3 partidos a continuación se trabajará con el modelo propuesto inicialmente, en el que se consideraran todos los partidos políticos involucrados en las elecciones para gobernador en Colima el 7 de Junio del 2015.

Regresión composicional con el modelo de 10 partidos

Al inicio de este capítulo se mencionó que el modelo que será utilizado para estos datos, así como las variables necesarias y su significado.

$$\begin{aligned}(PAN.y_j, PRI.y_j, ..., NULOS.y_j) &= a \oplus PAN.x_j \odot b_1 \oplus ... \oplus NULOS.x_j \odot b_{13} \\ &\oplus DIST\_FED_j \odot b_{14} + \varepsilon_j\end{aligned}$$

Donde la variable composicional Y_j tiene como componentes,

$$\begin{aligned}\mathrm{Y}_j &= (\mathrm{PAN.y}_j, PRI.y_j, PRD.y_j, PVEM.y_j, PT.y_j, MC.y_j, PANAL.y_j, MORENA.y_j, \\ &\mathrm{ES.y}_j, C\_PRI\_PVEM_j, NO\_REGISTRADOS.y_j, NULOS.y_j)\end{aligned}$$

Como primer paso se aplica la transformación *ilr* al modelo y utilizando el método de máxima verosimilitud se estiman los parámetros de cada variable independiente del modelo.

Los valores que se obtienen después de ajustar el modelo de regresión son los siguientes:

	PAN.y	PRI.y	PRD.y	PVEM.y	PT.y	MC.y
(Intercept)	0.3443	0.3146	0.0257	0.0312	0.0202	0.0886
PAN.x	0.0775	0.0769	0.0769	0.0766	0.0770	0.0768
PRI.x	0.0768	0.0774	0.0769	0.0765	0.0768	0.0769
PRD.x	0.0769	0.0770	0.0773	0.0768	0.0768	0.0772
PVEM.x	0.0768	0.0765	0.0770	0.0791	0.0768	0.0768
PT.x	0.0765	0.0761	0.0770	0.0777	0.0820	0.0773
MC.x	0.0767	0.0770	0.0755	0.0766	0.0765	0.0802
PANAL.x	0.0763	0.0765	0.0763	0.0754	0.0768	0.0776
MORENA.x	0.0766	0.0765	0.0763	0.0757	0.0766	0.0781
ES.x	0.0759	0.0765	0.0758	0.0774	0.0769	0.0779
C_PRI_PVEM	0.0760	0.0766	0.0728	0.0778	0.0746	0.0757
NO_REGISTRADOS.x	0.0826	0.0831	0.0766	0.0872	0.0950	0.0865
NULOS.x	0.0763	0.0761	0.0758	0.0772	0.0763	0.0767
DISTII	0.0718	0.0849	0.0975	0.1108	0.0577	0.0402
UBIII	0.0732	0.0751	0.0806	0.0813	0.0741	0.0511

	PANAL.y	MORENA.y	PH.y	ES.y	PRI_PVERDE_PANAL
(Intercept)	0.0183	0.0189	0.0239	0.0249	0.0201
PAN.x	0.0768	0.0768	0.0767	0.0768	0.0768
PRI.x	0.0768	0.0769	0.0768	0.0770	0.0769
PRD.x	0.0770	0.0770	0.0768	0.0766	0.0767
PVEM.x	0.0768	0.0765	0.0769	0.0763	0.0765
PT.x	0.0757	0.0765	0.0760	0.0765	0.0756
MC.x	0.0769	0.0764	0.0771	0.0762	0.0769
PANAL.x	0.0818	0.0754	0.0759	0.0773	0.0771
MORENA.x	0.0765	0.0826	0.0770	0.0746	0.0760
ES.x	0.0765	0.0760	0.0758	0.0820	0.0758
C_PRI_PVEM	0.0779	0.0773	0.0756	0.0740	0.0857
NO_REGISTRADOS.x	0.0787	0.0832	0.0590	0.0806	0.0719
NULOS.x	0.0761	0.0755	0.0759	0.0766	0.0773
DISTII	0.0747	0.0834	0.0859	0.0801	0.0746
UBIII	0.0729	0.0809	0.0909	0.0928	0.0768

	NO_REGISTRADOS.y	NULOS.y
(Intercept)	0.0483	0.0205
PAN.x	0.0768	0.0770
PRI.x	0.0768	0.0769
PRD.x	0.0767	0.0766
PVEM.x	0.0768	0.0767
PT.x	0.0757	0.0768
MC.x	0.0768	0.0766
PANAL.x	0.0769	0.0761
MORENA.x	0.0768	0.0762
ES.x	0.0773	0.0757
C_PRI_PVEM	0.0787	0.0766
NO_REGISTRADOS.x	0.0412	0.0738
NULOS.x	0.0775	0.0821
DISTII	0.0727	0.0651
UBIII	0.0719	0.0777

Como se mencionó con anterioridad una de las desventajas de hacer análisis con una cantidad mayor a 4 variables composicionales es el no poder visualizar de manera gráfica los resultados obtenidos. Afortunadamente los métodos de análisis pueden seguir siendo de gran ayuda en modelos grandes.

Para revisar si las variables correspondientes a cada uno de los parámetros estimados son relevantes se aplica un análisis de varianza multivariante y se obtienen los siguientes resultados.

Tabla 3.4. Análisis de Varianza

	Df	Pillai	approx F	num Df	den Df	Pr(>F)	
(Intercept)	1	0.99532	5741.6	12	324	< 2.2e-16	***
PAN.x	1	0.69289	60.9	12	324	< 2.2e-16	***
PRI.x	1	0.62015	44.1	12	324	< 2.2e-16	***
PRD.x	1	0.70583	64.8	12	324	< 2.2e-16	***
PVEM.x	1	0.55878	34.2	12	324	< 2.2e-16	***
PT.x	1	0.50973	28.1	12	324	< 2.2e-16	***
MC.x	1	0.59250	39.3	12	324	< 2.2e-16	***
PANAL.x	1	0.15998	5.1	12	324	6.861e-08	***
MORENA.x	1	0.23047	8.1	12	324	2.724e-13	***
ES.x	1	0.16285	5.3	12	324	4.296e-08	***
C_PRI_PVEM	1	0.15003	4.8	12	324	3.404e-07	***
NO_REGISTRADOS.x	1	0.13991	4.4	12	324	1.658e-06	***
NULOS.x	1	0.34067	14.0	12	324	< 2.2e-16	***
DIST	1	0.21601	7.4	12	324	4.061e-12	***
UBI	1	0.14816	4.7	12	324	4.576e-07	***
Residuals	335						

Signif. codes: 0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

De acuerdo con la información de la Tabla 3.4, muestra que las covariables del modelo de estudio son significativas para el modelo, con un nivel de significancia mayor de 0.001 .

Para finalizar con el modelo, se obtendrán las regiones de confianza para cada uno de los parámetros estimados en el modelo de regresión, por medio de la ecuación mencionada en el capítulo anterior.

$$\sqrt{(\bar{x}-\mu)\Sigma^{-1}(\bar{x}-\mu)} \leq \frac{c\sqrt{\lambda_i}}{\sqrt{n}} = \sqrt{\lambda_i}\sqrt{\frac{P(n-1)}{n(n-p)}F_{p,n-p}(\alpha)}$$

Así como los intervalos:

$$límite_i = \sqrt{\lambda_i}\sqrt{\frac{P(n-1)}{n(n-p)}F_{p,n-p}(\alpha)}$$

$$[media_i \pm límite_i]$$

Así las regiones de confianza en el espacio real resultantes son:

$$\begin{aligned}
\text{ilr}_{PAN} &\in \text{(-0.0919, -0.0680)} \\
\text{ilr}_{PRI} &\in \text{(-2.5712, -2.5188)} \\
\text{ilr}_{PRD} &\in \text{(-1.9350, -1.8765)} \\
\text{ilr}_{PVEM} &\in \text{(-1.4796,-1.3898)} \\
\text{ilr}_{PT} &\in \text{(0.3479, 0.4553)} \\
\text{ilr}_{MC} &\in \text{(-1.5808, -1.4538)} \\
\text{ilr}_{PANAL} &\in \text{(-1.2703,-1.0952)} \\
\text{ilr}_{MORENA} &\in \text{(-1.4851, -1.3005)} \\
\text{ilr}_{PH} &\in \text{(-1.1684, -0.9744)} \\
\text{ilr}_{ES} &\in \text{(-1.0477, -0.8051)} \\
\text{ilr}_{PRI_PVERDE_PANAL} &\in \text{(-0.3296, 0.0131)} \\
\text{ilr}_{NO_REGISTRADOS} &\in \text{(-0.4093, 0.0927)}
\end{aligned}$$

Aplicando la transformación inversa *ilr* se obtiene que las regiones de confianza que contiene al verdadero porcentaje de votos obtenidos en las elecciones pasadas para gobernador del Estado de Colima, resultan ser:

$$
\begin{aligned}
\text{PAN.y} &\in (0.3836, 0.4177) \\
\text{PRI.y} &\in (0.3484, 0.3668) \\
\text{PRD.y} &\in (0.0167, 0.0167) \\
\text{PVEM.y} &\in (0.0146, 0.0149) \\
\text{PT.y} &\in (0.0149, 0.0161) \\
\text{MC.y} &\in (0.0824, 0.0918) \\
\text{PANAL.y} &\in (0.0108, 0.0125) \\
\text{MORENA.y} &\in (0.0121, 0.0150) \\
\text{PH.y} &\in (0.0082, 0.0105) \\
\text{ES.y} &\in (0.0097, 0.0128) \\
\text{PRI\_ PVERDE\_PANAL} &\in (0.0098, 0.0139) \\
\text{NO\_ REGISTRADOS.y} &\in (0.0188, 0.0304) \\
\text{NULOS.y} &\in (0.0169, 0.0330)
\end{aligned}
$$

Recordemos que el PRI estaba en coalición con los partidos (PVEM y PANAL), por esta razón al final de las votaciones los votos son sumados y el intervalo que se obtiene de para el PRI.Y + COALICIÓN resulta ser (0.3838,0.4083).

Se observa que en las votaciones los partidos $PRI - COALICION$ y PAN tiene regiones de confianza similares a pesar de que el PRI sume votos de otro partidos.

Finalmente se comparan los resultados que se obtuvieron por medio de la regresión de datos composicionales contra los resultados oficiles proporcionados por el Instituto Electoral del Estado de Colima después de las votaciones del 7 de junio del 2015 en Colima, expresados en porcentaje se pueden encontrar en *(http://www.ieecolima.org.mx/resultados%2091-12/2015/gobernadorprincipal.html)*.

PARTIDO POLÍTICO	**RESULTADOS INE DE COLIMA**	**INTERVALOS DE PORCENTAJES COMPOSICIONALES**
PAN	39.65	(0.3836, 0.4177)
PRI	35.04	(0.3484, 0.3668)
PRD	1.96	(0.0167, 0.0167)
PVERDE	2.41	(0.0146, 0.0149)
PT	1.77	(0.0149, 0.0161)
MOV.CIU.	11.95	(0.0824, 0.0918)
ALIANZA	1.17	(0.0108, 0.0125)
MORENA	1.27	(0.0121, 0.0150)
HUMANISTA	0.61	(0.0082, 0.0105)
E.SOCIAL	0.83	(0.0097, 0.0128)
COALICIÓN	1.20	(0.0098, 0.0139)
NO_ REGISTRADOS.y	0.02	(0.0188, 0.0304)
NULOS	2.11	(0.0169, 0.0330)
PRI-PVERDE-ALIANZA-COALICIÓN	39.82	(0.3838,0.4083)

Conclusiones

El estudio de los datos composicionales es de suma importancia debido a la frecuencia con la que aparecen en problemas de la vida cotidiana, desafortunadamente no existe una gran difusión sobre los modelos para este tipo de datos o como se mencionó en un principio, las herramientas estadísticas empleadas para su estudio son utilizadas de manera errónea. Afortunadamente Aitchison y sus colaboradores proponen herramientas que ayudan a eliminar los problemas existentes en datos composicionales. Contar con una estructura geométrica propia de los datos composicionales, permite eliminar los problemas que existían en un principio. Además de proporcionar herramienta que facilita la comprensión de análisis estadístico en el Símplex, como el uso de diagramas ternarios, que permiten visualizar de manera gráfica el comportamiento de los vectores composicionales. El uso de transformaciones que ayuden a llevar los datos composicionales al espacio de los reales $\mathbb{R}$, permite utilizar cualquier método de estadística tradicional para su análisis.

El estudio de modelos de regresión donde alguna de las variables (dependiente e independientes) puede tomar el papel de una variable composicional, es importante reconocer el modelo de regresión composicional con el que se trabajará, posteriormente elegir de manera correcta la transformación que facilita el análisis del modelo. Uno de los objetivos de la presente tesis fue aplicar el modelo de regresión composicional donde la variable independiente fuera de tipo composicional. Por esta razón se analizaron datos sobre las elecciones de Colima llevadas a cabo en Junio del 2015 para Gobernador.

Gracias a las transformaciones existentes en el Símplex, se aplicó la teoría de elipses de confianza, logrando con esta herramienta un intervalo de confianza para cada proporción de votos de los partidos políticos. Al final del trabajo se pudo comparar los resultados proporcionados por el Instituto Electoral de Colima y los intervalos por medio de Regresión con datos composicionales. El proporcionar un capítulo con una aplicación, permite al lector entender la importancia de utilizar esta herramienta en problemas de la vida cotidiana, debido a que cumplen con los requisitos de ser analizados con las herramientas de datos composicionales, así como usar herramienta computacional como $R - Proyect$.

Apéndice A

Anexo

En el siguiente Anexo se encuentra el programa utilizado en R-Project, para el modelo de regresión cuando se cuenta con una variable independiente como dato composicional.

```
library(foreign)
library( compositions)
library(mvnormtest)
rm(list=ls())

source("coli.R")

dat<-read.csv("DATOS-ELECCIONES.csv",sep=",", header=T)

if(0){
  # partido.x: Elección de Diputados, 2015
  # partido y: Elección de Gobernador, 2015
  par(mfrow=c(2,2))
  plot(dat$PAN.x,dat$PAN.y)
  plot(dat$PRI.x,dat$PRI.y)
  plot(dat$PRD.x,dat$PRD.y)
  plot(dat$PVEM.x,dat$PVEM.y)
  plot(dat$PT.x,dat$PT.y)
  plot(dat$MC.x,dat$MC.y)
  plot(dat$MORENA.x,dat$MORENA.y)
  plot(dat$PANAL.x,dat$PANAL.y)
  plot(dat$PH.x,dat$PH.y)
  plot(dat$ES.x,dat$ES.y)

}
#-------------------------------------------------------------------------
dat$Municipio<-rep(0,length(dat[,1]))
```

```
n.estratos<-10
dat$Municipio[dat$MUNICIPIO=="ARMERIA"]<-1
dat$Municipio[dat$MUNICIPIO=="COLIMA"]<-2
dat$Municipio[dat$MUNICIPIO=="COMALA"]<-3
dat$Municipio[dat$MUNICIPIO=="COQUIMITLAN"]<-4
dat$Municipio[dat$MUNICIPIO=="CUAUHTEMOC"]<-5
dat$Municipio[dat$MUNICIPIO=="IXTLAHUACAN"]<-6
dat$Municipio[dat$MUNICIPIO=="MANZANILLO"]<-7
dat$Municipio[dat$MUNICIPIO=="MINATITLAN"]<-8
dat$Municipio[dat$MUNICIPIO=="TECOMAN"]<-9
dat$Municipio[dat$MUNICIPIO=="VILLA DE ALVAREZ"]<-10

#-------------------------------------------------------------------------
dat[is.na(dat)]<-0
base<-as.matrix(cbind(dat[,-c(1:3,8,13:15,27:29,33)]))
ncol.base<-length(base[1,])
#-------------------------------------------------------------------------


Nh<-c(1:n.estratos)*NA
for(k in 1:n.estratos){
  Nh[k]<- length(which(base[,37]==k))
}

Est<-list(
  d1=0, d2=0, d3=0, d4=0, d5=0, d6=0, d7=0, d8=0, d9=0, d10=0
)

# de cada uno de los n.estratos a las bases Est[[k]], k=1,...,n.estratos
n.casillas<-c(1:n.estratos)*0
for(k in 1:n.estratos){
  indx<-which(base[,37]==k)
  n.casillas[k]<- length(indx)
  if(n.casillas[k]==0){Est[[k]]<-matrix(0,1,ncol.base+1)}
  else{if(n.casillas[k]==1){
    base.aux<-matrix(c(base[indx,],0),1,ncol.base+1)
    Est[[k]]<-base.aux
  }else{
    Est[[k]]<-cbind(base[indx,], rep(0,n.casillas[k]) )
  }
  }
```

```
  rm(indx)
}

#-------------------------------------------------------------------------

nn<-350
nh.prop<-c(1:n.estratos)*NA
nh.op<-c(1:n.estratos)*NA
if(1){
  Sh.hat<-matrix(NA,n.estratos,2)
  for(k in 1:n.estratos){
    Sh.hat[k,1]<-apply(Est[[k]],2,sd)[9]
    Sh.hat[k,2]<-apply(Est[[k]],2,sd)[23]
  }

  Sh<-as.vector(Sh.hat[,2])
  ck<-sum(Nh*Sh)
  nh.op.aux<-nn*(Nh*Sh/ck)
  nh.prop.aux<-nn*(Nh/sum(Nh))
  nh.op<-round(nh.op.aux)
  nh.prop<-round(nh.prop.aux)
}
nh.prop.aux<-nn*(Nh/sum(Nh))
nh.prop<-round(nh.prop.aux)

# Se define el tamaño de muestra en cada estrato y se registra en la
# última columna de la base en cada estrato

nh<-nh.prop
sum(nh)

for(k in 1:n.estratos){
  ncol<-dim(Est[[k]])
  Est[[k]][,ncol[2]]<-nh[k]
}


Est.sample<-lapply(Est,GeneraMuestra.1)

#-------------------------------------------------------------------------
plan<rbind(Est.sample[[1]],Est.sample[[2]],Est.sample[[3]],Est.sample[[4]],
Est.sample[[5]],Est.sample[[6]],Est.sample[[7]],Est.sample[[8]],
Est.sample[[9]],Est.sample[[10]])
```

```
DIST_FED<-plan[,1]
DIST_LOC<-plan[,2]
SECCION <-plan[,3]
ID_CASILLA <-plan[,4]
EXT_CONTIGUA <-plan[,5]
UBICACION_CASILLA<-plan[,6]
TIPO_ACTA <-plan[,7]
LISTA_NOMINAL <-plan[,8]
PAN.x <-plan[,9]
PRI.x <-plan[,10]
PRD.x <-plan[,11]
PVEM.x<-plan[,12]
PT.x <-plan[,13]
MC.x <-plan[,14]
PANAL.x <-plan[,15]
MORENA.x <-plan[,16]
PH.x <-plan[,17]
ES.x<-plan[,18]
C_PRI_PVEM<-plan[,19]
NO_REGISTRADOS.x <-plan[,20]
NULOS.x <-plan[,21]
TOTAL.x <-plan[,22]
PAN.y <-plan[,23]
PRI.y <-plan[,24]
PRD.y <-plan[,25]
PVEM.y<-plan[,26]
PT.y <-plan[,27]
MC.y <-plan[,28]
PANAL.y<-plan[,29]
MORENA.y <-plan[,30]
PH.y <-plan[,31]
ES.y <-plan[,32]
PRI_PVERDE_PANAL<-plan[,33]
NO_REGISTRADOS.y <-plan[,34]
NULOS.y <-plan[,35]
TOTAL.y <-plan[,36]
Municipio<-plan[,37]
#------------------------------------------------------------------------------
VOTA<-cbind(DIST_FED,DIST_LOC, SECCION,ID_CASILLA, EXT_CONTIGUA,
      UBICACION_CASILLA, TIPO_ACTA, LISTA_NOMINAL, PAN.x, PRI.x,
      PRD.x,PVEM.x, PT.x,MC.x, PANAL.x, MORENA.x, PH.x, ES.x,
      C_PRI_PVEM,NO_REGISTRADOS.x, NULOS.x, TOTAL.x, PAN.y,
```

```
      PRI.y, PRD.y, PVEM.y,PT.y, MC.y ,PANAL.y,MORENA.y,PH.y,
      ES.y, PRI_PVERDE_PANAL,NO_REGISTRADOS.y ,NULOS.y ,TOTAL.y )
#------------------------------------------------------------------------------
#como 1er paso se representa el modelo con 3 variables para poder
#visualizar las gráficas expuestas con anterioridad

CASILLA<- factor(ID_CASILLA)
DIST<-factor(DIST_FED)
contrasts(CASILLA)
contrasts(DIST)

varY<-acomp(cbind(PAN.y, PRI.y,MC.y))

plot(varY, col=c("blue","red")[DIST])
legend(x=0.75,y=0.55,abbreviate(levels(DIST),minlength=1),
pch=20,col=c("blue","red"),yjust=0)

varX<-cbind(PAN.x, PRI.x, MC.x,PVEM.x)

modelp<- lm(ilr(varY)~ PAN.x+ PRI.x+ MC.x+PVEM.x+DIST)

modelo<-ilrInv(coef(modelp)[],orig=varY)
anova(modelp)

pred <- (predict(modelp))

prediccion <- ilrInv(predict(modelp,orig=varY))
names(prediccion)<-names(varY)
plot(prediccion)
#Revisión con Rcuadrada
R2(modelp)

##-----elipses de confianza
a <- ilrInv(coef(modelp)[1,],orig=varY)
b <- ilrInv(rbind(coef(modelp)[-1,]),orig=varY)
r<-mean(varY-mean(varY))
plot(modelo, col="green")
mu=a+b
Sigma1 <- ilrvar2clr(var(modelp))
ellipses(mu,Sigma1,2,col="red")
plot(pred,pch=20, col="blue")
Sigma1 <- ilrvar2clr(var(modelp))
me<- mean(pred)
```

```
var1<-cov(pred)
df1<-ncol(pred)-1
df2<-nrow(pred)-ncol(pred)+1
rconf<-sqrt(qf(p=0.95,df1,df2)*df1/df2)
plot(pred)
ellipses(me,var1,1,col="red")
par(new=TRUE)
plot(me,col="black")

#-------------------------------------------------------------------------

plot(prediccion ,col="blue")

Sigma1 <- ilrvar2clr(var(modelp))
ma<-mean(prediccion)
ellipses(ma,Sigma1,rconf,col="red")

#-------------------------------------------------------------------------

me<-mean(prediccion)

plot(varY,col="green")
par(new=TRUE,col="yellow")
plot(me,pch=20, col="blue")

Sigma2 <- ilrvar2clr(var(modelp))
ellipses(me,Sigma2,2,col="red")
# para los residuales
res<-resid(modelp)
residuo <- ilrInv(resid(modelp))
names(residuo)<- names(varY)
plot(residuo, col="red")
#-------------------------------------------------------------------------
qqnorm(residuo)
pr<-clr(prediccion)
re<-clr(residuo)
names(pr)<- names(clr(varY))
names(re)<- names(varY)
#---------- HOMOCEDASTICIDAD
opar <- par(oma=c(2,2,0,0),mar=c(4,4,1,1))
pairwisePlot(pr,re)
#mtext(text=c("Predicción","Residuo"),side=c(1,2),
at=0.5,line=2,outer=TRUE)
```

```
par(opar)
opar <- par(mfrow=c(3,3),mar=c(2,2,0,0), oma=c(3,3,0,0))
for(i in 1:3){for(j in 1:3){plot(log(Pr[,i]/pr[,j]),
log(re[,i]/residuo[,j]) ,pch=ifelse(i!=j,19,""))
    if(i==j){text(x=0,y=0,labels=colnames(re)[i],cex=1.5)}
    else{abline(h=0)}}}
mtext(text=c("Predicción","Residuo"),side=c(1,2),at=0.5,
line=2,outer=TRUE)
par(opar)
#-------------COLORES
opar <- par(oma=c(0,0,2,0),mar=c(4,4,1,0))
pairwisePlot(clr(prediccion),clr(residuo),col=as.numeric(DIST))
legend(locator(1),levels(DIST),col=as.numeric(1:3),pch=1,xpd=NA)
par(opar)
######
R2(modelp)
######
shapiro.test(residuo[])
#-------------------------------------------------------------------------
#####aqui termina lo del subconjunto##########
####aquí comienza lo del modelo original######
varY<-acomp(cbind(PAN.y, PRI.y,PRD.y, PVEM.y,PT.y, MC.y ,PANAL.y,
MORENA.y,PH.y, ES.y, PRI_PVERDE_PANAL ,NO_REGISTRADOS.y,NULOS.y  ))

varX<-cbind(PAN.x, PRI.x, PRD.x,PVEM.x, PT.x,MC.x, PANAL.x, MORENA.x,
      PH.x, ES.x,C_PRI_PVEM,NO_REGISTRADOS.x, NULOS.x)

DIST<-factor(DIST_FED)

DISL<-factor(DIST_LOC)

SEC<-factor(SECCION)

ID<-factor(ID_CASILLA)

EXT<-factor(EXT_CONTIGUA)

UBI<-factor(UBICACION_CASILLA)

ACTA<-factor(TIPO_ACTA)

LISTA<-factor(LISTA_NOMINAL)
```

```
model<- lm(ilr(varY)~ PRI.x+PRD.x+PVEM.x+PT.x+MC.x+PANAL.x+MORENA.x
          +ES.x+C_PRI_PVEM+NO_REGISTRADOS.x+NULOS.x+DIST+UBI)
model1<-ilrInv(coef(model)[],orig=varY)

a<-anova(model)
ilrInv(coef(anova(model))
plot(model)
#-------------------------------------------------------------------------
pred <- (predict(model))
pred

prediccion <- ilrInv(predict(model,orig=varY))
prediccion


 # media composicional
names(prediccion)<-names(varY)
mean(prediccion)

#------------------significancia del modelo

res<-resid(model)
r<-ilrInv(res)
mshapiro.test(t(res))
R2(model)

#------------------normalidad del modelo
qqnorm(ilrInv(resid(model),orig=varY))
qqnorm(Resid[,1])
ga<- predict(model)
mynew=data.frame(varX=munew)
(ga[,1],newdata=varX,interval="prediction")


#-------------------------------------------------------------------------
#HOMOSCEDATICIDAD

opar <- par(oma=c(3,3,0,0),mar=c(4,4,1,0))
pairwisePlot(ilr(Pred),ilr(Resid))
mtext(text=c("predicted values (clr)","residuals (clr)"),side=c(1,2),
            at=0.5,line=2,outer=TRUE)
par(opar)
```

```
# normalidad mediante pruebas de shapiro
StudentizedResiduals = rstudent(model)
qqnorm(StudentizedResiduals)
qqline(StudentizedResiduals)
shapiro.test(StudentizedResiduals)

#-------------------------------------------------------------------------
#R cuadrado ajustado

R2(model )
```

Bibliografía

[1] Aitchison, J. (1986). *The Statistical Analysis of Compositional Data.*, Chapman and Hall.

[2] Aitchison,J. *A concise Guide to Compositional Data Analysis.*

[3] Aldrich, J.(1995). *Correlations Genuine and Spurious in Pearson and Yule.*, Statistical Science.

[4] Alperin, M. (2013). *Introducción al análisis estadístico de datos geológicos.*, Editorial de la Universidad de La Plata.

[5] Bates,D. and Watts, D. (1988). *Nonlinear Regression Analysis and Its Applications.*, John Wiley Sons, Ltd.

[6] Boggs, S.(2009). *Petrology of sedimentary rocks.*,Cambridge University Presss, New York.

[7] Buccianti, A., Mateu- Figueras, G. and Pawlowsky, V. (2006). *Compositional data analysis in the Geosciences: From theory to practice.*, The Geological Society London.

[8] Buccianti, A., Mateu- Figueras, G. (2001). *Compositional Data Analysis (Theory and Applications.)*, John Wiley Sons, Ltd.

[9] Díaz, L. y Morales, M. (2012). *Estadística multivariada: inferencias y métodos.*, Editorial Universidad Nacional de Colombia.

[10] Hess, G. and Bay, J. (1997). *Generaty confidence intervals for compositions-based landscape indexes.*, Kluwer Academic Publishers.

[11] Hoskin, G., Galvis, M. y Masias, R. (2005). *Modelos de decisión electoral y perfiles de votante ene Colombia: elecciones presidenciales 2002.*, Departamento de ciencia política de la Universidad de los Andes.

[12] Hron, K. (2009). *Analytical representation of ellipses in the Aitchison Geometry and its application.*, Palacký University.

[13] Johnson, J. (1991). *Applied Multivariate Data Analysis.*, Springer.

[14] Levin, R. y Rubin, S. (2004). *Estadística para Administración y Economía.*, Pearson Educación de México, S.A. DE C.V.

[15] Maindonald, J. and Braun, J. (2010). *Data Analysis and Graphics Using R- an Example-Based Approach.*,Cambridge University Presss, New York.

[16] Mardia, K. and Kent, J. (1995). *Multivariate Analysis*, Academia Press.

[17] Márquez, J. y Aparicio, J. (2011). *Modelos estadísticos para sistemas electorales multipartidistas en Stata.*

[18] Miguez, F. (2007). *Introduction to R for multivariate data analysis.*,Turner Hall.

[19] Morales, I. (2014). *Efecto incumbente en elecciones municipales: un análisis de regresión discontinua para guatemala.*, Kluwer Academic Publishers. Revista de Análisis Económico Vol.29 N.2, pp. 113- 150.

[20] Myers, R y Geoffrey, G. (2010). *Generalized linear models.*, John Wiley & Sons.

[21] Pawlowsky- Glahn, V., Egozcue, J. and Delgado, R. (2015). *Modeling and Analysis of Compositional Data*, John Wiley & Sons, Ltd.

[22] Pearson, K. (1897). *Mathematical contributions to the theory of evolution. On a form of spurious correlation which may arise when indices are used in the measurement of organs.* Proceedings of Royal Society.

[23] Peña, D. (2002). *Análisis de datos multivariantes*.

[24] Rencher, A. (2002). *Methods of Multivariate Analysis.*, John Wiley & Sons, Ltd.

[25] Rollinson. (1993). *Principles of Igneous and Metamorphic Petrology.*, Universidad da Coruña, Cambridge.

[26] Sánchez, R. *Regresión de Dirichlet en datos composicionales: distribución de tareas de consultoría*, Universidad da Coruña.

[27] Van den Boogaart K. y Tolosana, R. (2013). *Analyzing Compositional Data with R*, Springer.