


Regresión composicional: una aplicación a la distribución de las horas en proyectos de consultoría

Autora:
Andrea Lagoa Cereijo

Director:
Jacobo De Uña Álvarez

Máster en Técnicas Estadísticas

8 de julio de 2013

Índice general

Capítulo 1

Introducción

El presente trabajo se ha desarrollado en el marco de la realización de unas prácticas en la empresa Azteca de Ingeniería Consulting S.L., ubicada en Vigo, durante el verano de 2012. El objetivo de las prácticas se centró en la creación de una base de datos para la clasificación e imputación de las horas dedicadas por cada trabajador de la empresa a diferentes proyectos, y su posterior análisis estadístico.

La naturaleza de los datos analizados durante la realización de las prácticas ha hecho necesario abordar el tratamiento de los datos composicionales. Esta problemática surge cuando se trata el análisis de datos que representan partes de un total, como porcentajes o proporciones. Su principal característica es que su suma está restringida a ser una constante, igual a 1 en el caso de proporciones, a 100 para porcentajes, o una constante c en otras situaciones. Las restricciones a las que están sujetas estos datos implican que los métodos estadísticos clásicos no son adecuados para su estudio y modelización, principalmente debido a que el espacio muestral de los vectores composicionales es muy diferente al espacio euclídeo real asociado a datos sin restricciones. El primero en advertir la inadecuación de las técnicas estadísticas multivariantes para los datos composicionales fue Pearson en 1897, pero no fue hasta los años 80, con la publicación de la monografía de Aitchison (1986) cuando se dispone de una metodología específica para el análisis estadístico de este tipo de datos, basada en transformaciones para trasladar los datos al espacio real donde poder aplicar las técnicas estadísticas habituales.

Los datos composicionales son habituales en ciencias como geología o geoquímica, pero también son frecuentes en biología, sociología, ingeniería, ciencias medioambientales... Juegan también un papel importante en el campo económico, por ejemplo la distribución del presupuesto familiar en distintas partidas de gasto, la composición relativa de una cartera de inversión, la distribución de las ventas de un producto en distintas regiones, o, como analizaremos en este trabajo, la distribución de las horas trabajadas en un proyecto entre los diferentes departamentos de la empresa.

Las prácticas en Azteca Ingeniería se realizaron en conjunto con otra alumna del Máster en Técnicas Estadísticas, Rosa Caeiro Sánchez, que también basó la realización de su trabajo de fin de máster en el análisis de datos composicionales. Sin embargo, el enfoque de este trabajo es diferente por dos motivos principalmente:

- Ofrece una aproximación basada en un modelo de regresión multivariante, en contraposición al modelo de regresión de Dirichlet, cuyos coeficientes son de difícil interpretación.
- Se aborda más en profundidad el tratamiento de las componentes nulas en los datos composicionales, presentando un estudio de simulación para comprobar qué método de reemplazo de ceros es más adecuado en el contexto de los modelos de regresión.

En el primer apartado de este trabajo se explicará más en detalle la labor realizada durante las prácticas, y se identificarán los objetivos y fases del proyecto. En la segunda parte se revisarán los principios teóricos de los datos composicionales, comentando los principales problemas en su tratamiento. En la tercera parte se introducirán los modelos de regresión adecuados para los datos composicionales. Por último, se presentan los datos con los que se hará el análisis, en donde se estudiará cómo se distribuyen las horas dedicadas a diferentes proyectos en función de una serie de variables explicativas.

Capítulo 2

Desarrollo de las prácticas

Este trabajo es el resultado de la realización de unas prácticas mediante un Convenio de Cooperación Educativa entre la Universidad de Vigo y la empresa Azteca Ingeniería Consulting S.L. Las prácticas tuvieron una duración de tres meses, de junio a agosto de 2012, con una dedicación de 5 horas diarias de lunes a viernes. El principal objetivo de las prácticas fue el diseño y creación de una base de datos para la clasificación e imputación de las horas dedicadas por cada trabajador de la empresa a diferentes proyectos.

Desde hace unos años, cada trabajador de Azteca debía presentar semanalmente un documento indicando a qué proyectos y tareas se había dedicado durante la semana. Sin embargo no existía homogeneidad a la hora de preparar y presentar estos partes de trabajo, por lo que el grado de detalle y la claridad en la descripción de las tareas realizadas variaba considerablemente de unos trabajadores a otros. Con este proyecto, se pretendía crear una estructura predefinida de tareas, consensuada con todos los trabajadores y que abarcara la variedad de actividades realizadas en los diferentes departamentos.

En Azteca se plantean, en un futuro, desarrollar e implantar un software CRM (de las siglas en inglés *Customer Relationship Manager*, o administrador de las relaciones con los clientes), una aplicación informática que permita analizar y procesar toda la información necesaria para dar soporte a la estrategia de negocio de la empresa. Por ello, y para facilitar más adelante una posible sincronización de la información con el CRM, se decide utilizar el sistema de gestión de bases de datos MySQL.

2.1. Objetivos

Desde el inicio de las prácticas se establecieron una serie de objetivos que se debían cumplir durante la realización de las mismas. Estos objetivos se pueden resumir en los siguientes puntos:

1. Obtener los *perfiles de los trabajadores*, conociendo las tareas y proyectos a los que cada trabajador dedica sus horas.
2. Obtener los *perfiles de los proyectos*, en función del número de trabajadores y la dedicación por tareas.
3. Estimar las cargas horarias de *proyectos futuros*.

Una vez diseñada la base de datos e introducidos los datos de los partes de trabajo en ella, los dos primeros objetivos se alcanzaron obteniendo de la base de datos una serie de informes. Estos informes, basados en técnicas de estadística descriptiva, permiten analizar cómo se distribuyen las horas trabajadas entre los proyectos, los trabajadores y las tareas. Además, estos informes pueden obtenerse para cualquier rango de fechas, permitiendo conocer la evolución de la distribución de las horas trabajadas a lo largo de diferentes meses o años. Los informes diseñados fueron los siguientes:

- Departamento-tarea: para cada departamento de la empresa, total de horas dedicadas y cómo se distribuyen entre las diferentes tareas.
- Proyecto-tarea: para cada proyecto, cómo se distribuyen las horas entre las diferentes tareas.
- Proyecto-trabajador: para cada proyecto, cuántas horas le dedica cada trabajador.
- Proyecto-tarea-trabajador: para cada proyecto, qué tareas se le dedican (y número de horas) y qué trabajadores las realizan.
- Proyecto-trabajador-tarea: para cada proyecto, cuántas horas le dedica cada trabajador y cómo las distribuye por tareas.
- Trabajador-proyecto: para cada trabajador, distribución de sus horas trabajadas por proyectos.
- Trabajador-tarea: para cada trabajador, distribución de sus horas trabajadas por tareas.
- Trabajador-proyecto-tarea: para cada trabajador, distribución de sus horas trabajadas por proyectos y, dentro de cada proyecto, horas dedicadas a cada tarea.
- Trabajador-tarea-proyecto: para cada trabajador, distribución de sus horas trabajadas por tareas y, para cada tarea, horas dedicadas por proyectos.

Una vez obtenido un informe la empresa puede realizar gráficas que faciliten la interpretación y análisis de los datos, como se muestra en las Figuras 2.1 y 2.2. No se muestran los nombres de los trabajadores ni de las tareas para proteger información sensible de la empresa.

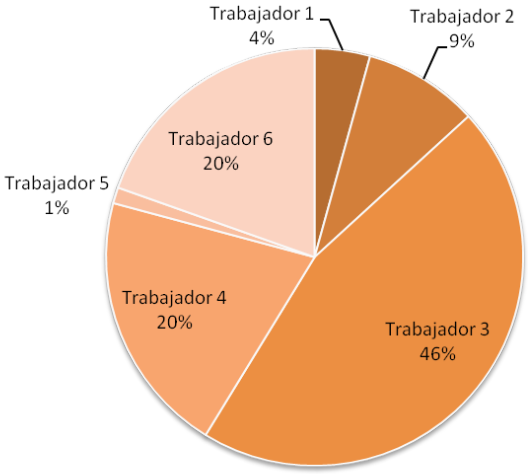

Figura 2.1: *Ejemplo de informe proyecto-trabajador, con la proporción de horas dedicadas por diferentes trabajadores al proyecto 17 durante 2012*

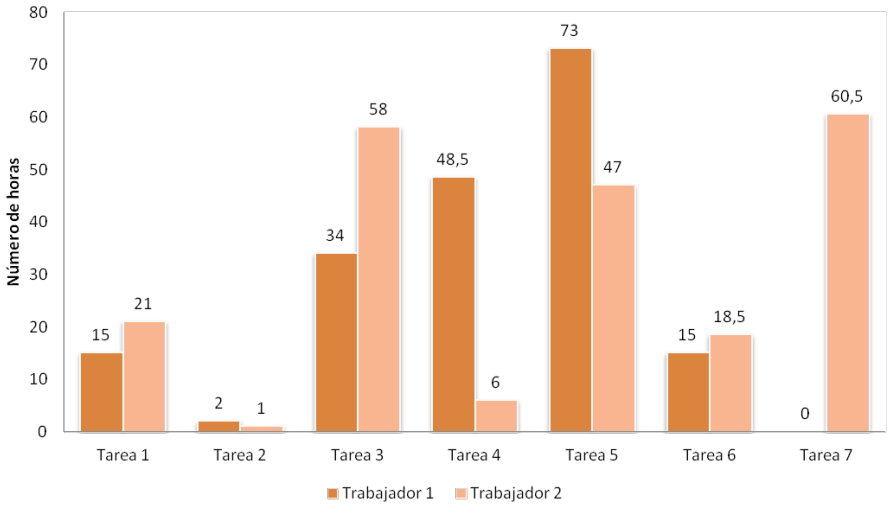

Figura 2.2: *Ejemplo de informe trabajador-tarea con el total de horas dedicadas por dos trabajadores a diferentes tareas en noviembre de 2011*

En cuanto al tercer objetivo, es necesario recurrir a técnicas estadísticas de predicción. Se pretende que, dado un futuro proyecto, se pueda conocer cómo se van a repartir las horas dedicadas al mismo entre los diferentes departamentos de la empresa (técnico, administrativo, comercial e I+D+i). Por lo tanto, hay que tener en cuenta las especiales características de esta información, ya que al trabajar con porcentajes de horas totales por proyecto, nos encontramos ante datos composicionales.

Toda esta información facilitará la toma de decisiones en la empresa, permitiendo la realización de cronogramas de proyectos, presupuestos y cálculo de ratios de productividad.

Por otro lado, y puesto que el proyecto se desarrolla en el marco de un Convenio de Cooperación Educativa entre la Universidad de Vigo y Azteca, un objetivo añadido es la *formación de los estudiantes en prácticas* en la operativa diaria de una consultoría de ingeniería y, por extensión, de una empresa de servicios. Al mismo tiempo, se transmite la importancia del tratamiento y el procesado de la información para la toma de decisiones mediante el uso de herramientas estadísticas y de gestión.

2.2. Fases del proyecto

A grandes rasgos, las **fases** de las prácticas desarrolladas en la empresa fueron las siguientes:

1. Definición de objetivos y recopilación de la información disponible.
2. Análisis de la información, identificación de la muestra disponible y planificación del proyecto.
3. Diseño de la base de datos.
4. Introducción de los datos en la base de datos.
5. Identificación de las variables respuesta y realización de informes descriptivos.
6. Análisis estadístico, mediante la selección de un modelo de regresión adecuado y posterior validación.
7. Análisis de los resultados obtenidos y redacción de informe de resultados.

El primer paso de este proyecto consistió en la recopilación de toda la información disponible, es decir, de los partes de trabajo semanales entregados por los trabajadores en los últimos años. Existían datos disponibles desde el año 2008, pero se decide que se empezará a introducir la información en la base de datos desde el año 2010, por ser el año más completo en cuanto a número de partes y detalle de los mismos. Una vez conocida la cantidad total de partes disponibles, se pudo proceder a planificar la ejecución del proyecto.

La parte con mayor carga de trabajo durante los meses de realización de las prácticas fue la de introducción de los partes de trabajo en la base de datos. Esto implicaba la necesidad de ir consensuando con los diferentes departamentos de la empresa una lista de tareas aceptada por todos y que se adaptara a cualquier actividad desarrollada en Azteca. Esta fase consumió cerca de dos terceras partes del total de la duración de las prácticas.

Una vez completada la base de datos, y ya que su sistema de gestión se realiza en MySQL, fue necesario realizar algunos ajustes para poder importar los datos al software estadístico R. Si se trabaja con el sistema operativo Windows, el primer paso consiste en descargar e instalar el *driver* de MySQL y añadirlo a la lista de DSN de usuario en "Origen de datos ODBC". A continuación se instala en R el paquete RODBC, disponible en: `cran.r-project.org/web/packages/RODBC/`. Este paquete posibilita conectar R con bases de datos, ya sean en Excel, Acces, MySQL,... permitiéndole realizar consultas sobre ellas e incluso modificarlas. Una vez instalada y cargada la librería en R es necesario establecer la conexión con la base de datos proporcionándole el DSN del driver instalado en el paso anterior. De esta forma ya se pueden realizar consultas sobre una base de datos externa, y es posible tratar ficheros muy grandes de los que sólo se importan a R las observaciones o variables de interés.

Por ejemplo, si hemos instalado el driver y lo hemos denominado *driverMySQL*, vinculándolo a la base de datos en la que la información de interés está almacenada en una tabla denominada *datos*, el código en R necesario para acceder a todo el contenido de la tabla es:

```
> library(RODBC)
> canal<-odbcConnect('driverMySQL')
> datos<-sqlQuery(canal,'SELECT * FROM datos')
```

A partir de este punto, ya se puede abordar el objetivo de estimar las cargar horarias de proyectos futuros, pero para ello es necesario reestructurar la información disponible. En la base de datos en la que se introducen los partes de trabajo cada *observación* indica: fecha, trabajador, proyecto, tarea y número de horas dedicadas. Sin embargo, para el análisis estadístico interesa disponer de la información por proyectos: deseamos conocer, para cada proyecto, la distribución porcentual de las horas entre los diferentes departamentos de la empresa. Por tanto, fue necesaria también una labor de reorganización de la base de datos que hiciera posible el acceso a la información en distintos formatos.

Capítulo 3

Datos Composicionales

3.1. Fundamentos matemáticos

Los datos composicionales constan de vectores cuyas D componentes son proporciones o porcentajes de un total. Su peculiaridad es que su suma se restringe a ser una constante c, que sería igual a 1 en el caso de proporciones.

Definición 3.1.1 *Un dato composicional* $\boldsymbol{x} = (x_1, x_2, \cdots, x_D)'$ *con* D *partes, es un vector con componentes estrictamente positivas, tal que la suma de todas ellas es igual a una constante* c*. Su espacio muestral es el símplex* $\boldsymbol{S}^D$*, definido por*

$$\boldsymbol{S}^D = \{(x_1, x_2, \cdots, x_D)' : x_i > 0; \sum_{i=1}^{D} x_i = c\} \tag{3.1}$$

Para el caso $D = 3$, el símplex S^3 suele representarse mediante un diagrama ternario, triángulo equilátero de altura c. Existe una correspondencia biunívoca entre los datos composicionales con 3 partes y los puntos del diagrama ternario. Un dato composicional $\mathbf{x} = (x_1, x_2, x_3)'$ se corresponde con el punto que dista x_1,x_2 y x_3, respectivamente, de los lados opuestos a los vértices 1, 2 y 3 (Figura 3.1).

Las operaciones fundamentales en el símplex son la *perturbación* y la *potenciación*. Para tener una idea intuitiva del significado de estas operaciones básicas en el símplex diremos que la perturbación es el equivalente a la traslación o suma en espacios reales y la potenciación es el equivalente al producto por escalar. Formalmente, estas operaciones se definen a continuación.

Definición 3.1.2 *Dadas dos composiciones de* D *partes* $\boldsymbol{x}, \boldsymbol{y} \in S^D$*, la perturbación se define como:*

$$\boldsymbol{x} \oplus \boldsymbol{y} = \mathcal{C}[x_1 y_1, \cdots, x_D y_D] \tag{3.2}$$

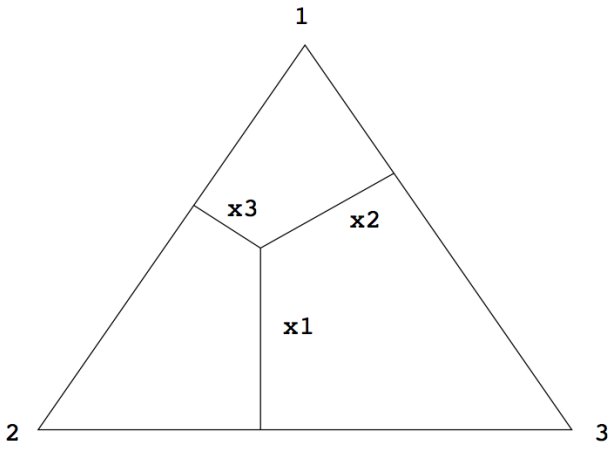

Figura 3.1: *Representación de un dato composicional* $(x_1, x_2, x_3)'$ *en el símplex* S^3

Definición 3.1.3 *Dada una composición* $\boldsymbol{x} \in S^D$ *y un número real* α*, la potenciación es:*

$$\alpha \otimes \boldsymbol{x} = \mathcal{C}[x_1^\alpha, \cdots, x_D^\alpha] \tag{3.3}$$

Definición 3.1.4 *El operador* $\mathcal{C}$ *se denomina clausura, es una transformación que hace corresponder a cada vector* $\boldsymbol{w} = (w_1, w_2, \cdots, w_D)'$ *de* $\mathbb{R}_+^D$ *su dato composicional asociado* $\mathcal{C}(\boldsymbol{w}) = c\boldsymbol{w}/(w_1 + w_2 + \cdots + w_D)$*, con* c *la constante de clausura. De esta forma las componentes del nuevo vector suman* c *y por tanto pertenecen a* S^D*.*

Además, es posible obtener una estructura del espacio vectorial Euclídeo en el símplex, añadiendo el producto interno, la norma y la distancia a las definiciones anteriores. La estructura de espacio Euclídeo implica que sobre el símplex pueden definirse todos los entes y conceptos geométricos usuales tales como líneas, ángulos, ortogonalidad, paralelismo, etc.

Definición 3.1.5 *El producto interno de* $\boldsymbol{x}, \boldsymbol{y} \in S^D$ *se define como:*

$$\langle \boldsymbol{x}, \boldsymbol{y} \rangle_a = \frac{1}{2D} \sum_{i=1}^{D} \sum_{j=1}^{D} ln\frac{x_i}{x_j} ln\frac{y_i}{y_j} \tag{3.4}$$

Definición 3.1.6 *La norma de* $\boldsymbol{x} \in S^D$ *es:*

$$\|\boldsymbol{x}\|_a = \sqrt{\frac{1}{2D} \sum_{i=1}^{D} \sum_{j=1}^{D} \left(ln\frac{x_i}{x_j} \right)^2} \tag{3.5}$$

Definición 3.1.7 *La distancia entre* $\boldsymbol{x}$ *e* $\boldsymbol{y} \in S^D$ *se define como:*

$$d_a(\boldsymbol{x}, \boldsymbol{y}) = \|\boldsymbol{x} \ominus \boldsymbol{y}\|_a = \sqrt{\frac{1}{2D} \sum_{i=1}^{D} \sum_{j=1}^{D} \left(ln\frac{x_i}{x_j} - ln\frac{y_i}{y_j} \right)^2} \tag{3.6}$$

En ocasiones puede interesar centrar la atención en un subconjunto de los componentes, es decir, en *subcomposiciones*, que son las proyecciones del símplex S^D sobre un sub-símplex de dimensión menor, por ejemplo S^d, obtenidas mediante la clausura de un subvector formado por d de las partes de una composición en S^D.

Definición 3.1.8 *Si S es un subconjunto cualquiera de las partes $1, 2, \ldots, D$ de un dato composicional $\boldsymbol{x} \in S^D$ y $\boldsymbol{x}_S$ simboliza el subvector formado por las correspondientes partes de $\boldsymbol{x}$, entonces $\boldsymbol{s} = \mathcal{C}(\boldsymbol{x}_S)$ recibe el nombre de subcomposición de las S partes de $\boldsymbol{x}$.*

La formación de una subcomposición tiene la buena propiedad de conservar la magnitud relativa de las partes.

Al ser la suma de las componentes, siempre positivas, de un vector de datos composicionales una constante, un cambio en una componente conlleva un cambio en, al menos, una de las demás componentes. Esto provoca que una fila cualquiera de la matriz de covarianzas de una muestra de vectores de proporciones siempre tenga al menos un elemento negativo y que su suma sea igual a 0. Esto implica que la matriz de covarianzas es singular y que las correlaciones no varían libremente en el habitual intervalo [-1,1]. Ya en 1897 Pearson advirtió de la existencia de una falsa correlación entre las partes de una composición. que falsea la imagen de las relaciones de dependencia y pueden conducir a interpretaciones erróneas.

Otro aspecto importante a tener en cuenta en el análisis de datos composicionales es que no existe ninguna relación entre la matriz de covarianzas de una subcomposición y la de la composición de procedencia: el signo de la covarianza entre dos partes puede ir fluctuando cuando nos movemos de la composición inicial a subcomposiciones de dimensión cada vez más pequeña.

Otra de las dificultades importantes es la falta de familias paramétricas suficientemente flexibles para modelar los conjuntos de datos composicionales. Una opción es la distribución de Dirichlet y sus generalizaciones, que se obtienen mediante la clausura de vectores aleatorios con componentes independientes. Esto impide su uso en la modelización de fenómenos con relaciones de dependencia no inducidas por la suma constante.

3.2. Transformaciones log-cociente

Una metodología adecuada para el análisis de datos composicionales debe tener en cuenta las características del símplex como espacio muestral sobre el que se definen. La idea principal es que las composiciones sólo proporcionan información sobre la magnitud relativa de sus partes, y se asume que el valor de la suma de sus partes es irrelevante. Por lo tanto, cualquier afirmación sobre una composición debe hacerse en términos de los cocientes entre las partes, los cuales

medirán dicha relación relativa. De esta forma, una función aplicable sobre datos composicionales deberá ser invariante por cambios de escala y expresable en términos de cocientes entre las partes, asegurando el cumplimiento del principio de *coherencia subcomposicional*. Este principio requiere que cuando se examina un subconjunto de las partes de una composición los resultados del análisis no sean contradictorios con los obtenidos de la composición original.

La metodología propuesta por Aitchison (1986) proporciona, por primera vez, una forma de evitar la restricción de la suma constante. Esta metodología se basa en la transformación de una composición definida sobre S^D en un vector que involucre los cocientes entre las partes y que esté definido sobre el espacio real. De esta manera, el problema queda expresado en términos de tales vectores transformados, con lo que se tiene la posibilidad de resolverlo utilizando las técnicas estadísticas multivariantes habituales en espacios reales. Trabajando con los cocientes desaparecen los problemas de las correlaciones espurias. Por otra parte, la magnitud relativa entre las partes de una subcomposición no cambia en relación a la magnitud relativa entre las partes de la composición original. Por lo tanto, cuando trabajamos con funciones invariantes por cambios de escala, se cumple el principio de coherencia subcomposicional.

Aitchison propone dos tipos de transformaciones de los datos, basadas en los logaritmos de cocientes entre las partes de un datos composicional.

Definición 3.2.1 *La transformación log-cociente aditiva (alr) de* $\boldsymbol{x} \in S^D$ *a* $\boldsymbol{y} \in \mathbb{R}^{D-1}$ *se define como:*

$$\boldsymbol{y} = alr(\boldsymbol{x}) = \left(ln\frac{x_1}{x_D}, ln\frac{x_2}{x_D}, \cdots, ln\frac{x_{D-1}}{x_D} \right) \tag{3.7}$$

Un inconveniente de la transformación alr es su asimetría respecto a las partes de la composición, ya que la utilizada como denominador cobra especial protagonismo. Por otro lado, no es una transformación isométrica, los ángulos y distancias en el símplex no pueden asociarse con ángulos y distancias en el espacio real (con la métrica Euclídea).

Definición 3.2.2 *La transformación log-cociente centrada (clr) de* $\boldsymbol{x} \in S^D$ *a* $\boldsymbol{z} \in \mathbb{R}^D$ *se define como:*

$$\boldsymbol{z} = clr(\boldsymbol{x}) = \left(ln\frac{x_1}{g(\boldsymbol{x})}, ln\frac{x_2}{g(\boldsymbol{x})}, \cdots, ln\frac{x_D}{g(\boldsymbol{x})} \right) \tag{3.8}$$

donde $g(\boldsymbol{x})$ *es la media geométrica de las D partes de* $\boldsymbol{x}$*.*

La transformación clr es simétrica e isométrica, pero la imagen de S^D queda realmente restringida a un subespacio de $\mathbb{R}^D$ y la matriz de covarianzas de los datos clr-transformados es singular.

Egozcue y otros (2003) proponen la transformación *log-cociente isométrica (ilr)*, que salva los principales inconvenientes de las dos transformaciones anteriores.

Definición 3.2.3 *La transformación ilr de una composición* $\boldsymbol{x}$ *se define como:*

$$ilr(\boldsymbol{x}) = (\langle \boldsymbol{x}, \boldsymbol{e}_1 \rangle, \ldots, \langle \boldsymbol{x}, \boldsymbol{e}_{D-1} \rangle) \tag{3.9}$$

donde $\boldsymbol{e}_1, \ldots, \boldsymbol{e}_{D-1}$ *es una base ortonormal del símplex.*

Es decir, las componentes del vector ilr-transformado son las coordenadas de la composición $\mathbf{x}$ respecto a la base $\mathbf{e}_1, \ldots, \mathbf{e}_{D-1}$. Esta transformación es isométrica. El principal problema es determinar cuál es la base ortonormal más apropiada para un problema concreto, la que proporciona las expresiones que hacen más fácil la interpretación de los resultados.

La existencia de más de una transformación nos lleva a la situación de deber elegir entre una de ellas como paso previo a la aplicación de cualquier método estadístico multivariante. Tradicionalmente, en las aplicaciones que exigen simetría en el tratamiento de sus componentes, como por ejemplo una clasificación no paramétrica, se utiliza la trasformación clr. Para la modelización de conjuntos de datos composicionales con distribuciones multivariantes, se ha venido utilizando mayoritariamente la trasformación alr. De esta forma se evita trabajar con distribuciones degeneradas.

3.3. Medidas de posición y variabilidad

Los estadísticos descriptivos estándar no son muy informativos en el caso de datos composicionales. En particular, la media aritmética y la varianza o desviación típica de los componentes individuales no se ajustan a la geometría de Aitchison como medidas de tendencia central y dispersión. A continuación se definen las conceptos de *centro*, *matriz de variación* y *varianza total*, propuestas por Aitchison (1986).

Definición 3.3.1 *Una medida de tendencia central para datos composicionales es la media geométrica composicional. Dada una muestra* $\boldsymbol{X} = (\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n)$ *de datos composicionales en* S^D*, se define como:*

$$\boldsymbol{g}(\boldsymbol{X}) = \mathcal{C}(g_1, g_2, \ldots, g_D) \tag{3.10}$$

con $g_i = \left(\prod_{j=1}^{n} x_{ij}\right)^{1/n}$, $i = 1, 2, \ldots, D$.

En la Figura 3.2 puede apreciarse cómo la media geométrica composicional, además de ser compatible con las operaciones básicas del símplex, representa el centro de gravedad de la nube de puntos composicionales de forma más adecuada que la habitual media aritmética. En general, esto ocurrirá siempre que la nube de puntos en S^3 tenga una forma aproximadamente cóncava.

Definición 3.3.2 *La dispersión en un conjunto de datos composicionales se puede describir con*

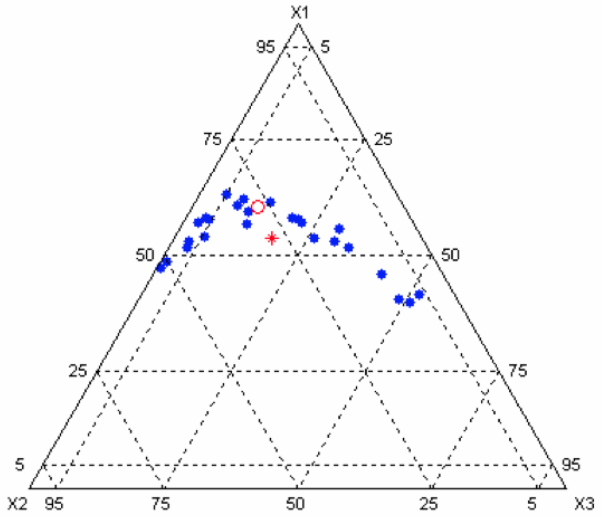

Figura 3.2: *Diagrama ternario con media geométrica composicional (o) y media aritmética (*) de un conjunto de datos composicionales*

la matriz de variación:

$$\boldsymbol{T} = \begin{pmatrix} t_{11} & t_{12} & \cdots & t_{1D} \\ t_{21} & t_{22} & \cdots & t_{2D} \\ \vdots & \vdots & \ddots & \vdots \\ t_{D1} & t_{D2} & \cdots & t_{DD} \end{pmatrix}, \qquad t_{ij} = var\left(ln\frac{x_i}{x_j}\right) \tag{3.11}$$

Esta matriz es simétrica y tiene ceros en la diagonal principal. Aunque no puede expresarse como la matriz de covarianzas estándar de un vector, sí que está relacionada con las matrices de covarianzas de los vectores log-cociente transformados alr, clr, e ilr mediante simples operaciones matriciales (Egozcue y otros, 2003).

Definición 3.3.3 *Una medida de la dispersión global es la varianza total, dada por:*

$$totvar[\boldsymbol{X}] = \frac{1}{2D}\sum_{i=1}^{D}\sum_{j=1}^{D} var\left(ln\frac{x_i}{x_j}\right) = \frac{1}{2D}\sum_{i=1}^{D}\sum_{j=1}^{D} t_{ij} \tag{3.12}$$

La varianza total resume la matriz de variación en una única cantidad.

3.4. El problema de los ceros

Hasta ahora, se ha visto que el análisis estadístico de vectores de proporciones gira en torno a los cocientes entre sus partes. Sin embargo, si un vector presenta alguna componente nula no son aplicables las transformaciones log-cociente de la forma x_i/x_j. En general, en el estudio de

los datos composicionales, se consideran dos tipos de ceros:

1. *Ceros esenciales*: se corresponden con componentes que realmente toman un valor nulo. Pueden surgir por una excesiva desagregación o por la existencia de diferentes submuestras dentro del conjunto de datos. En Aitchison y Kay (2003) pueden encontrarse algunas propuestas para tratar con los ceros esenciales.

2. *Ceros por redondeo*: se corresponden con valores que no han podido observarse por limitaciones en los instrumentos de medida o en el procedimiento de recogida y tratamiento de los datos, o incluso por políticas que impiden que se registren cuantías pequeñas que no superan cierto umbral de detección.

En el caso de los ceros esenciales, en algunos casos se puede interpretar su presencia como un indicador de que la composición pertenece a un grupo diferente. Por ejemplo, como ilustra Aitchison (1986), en el análisis de la distribución de presupuestos domésticos puede haber familias con valores nulos en el grupo de gasto de "tabaco y alcohol", lo que podría llevar a analizar por separado las familias que consumen tabaco y bebidas alcohólicas de las que no. Otra solución podría ser la fusión o amalgamamiento de algunas de las partes. Sin embargo, el dato composicional con ceros por redondeo es un tipo especial de dato incompleto y, por lo tanto, es susceptible de ser tratado mediante técnicas de imputación o reemplazo. El procedimiento elegido debe preservar la estructura de covarianzas y ser coherente con las propiedades métricas específicas de los datos composicionales.

3.4.1. Estrategias de reemplazo

El primer método de reemplazo fue el **reemplazo aditivo** propuesto por Aitchison (1986). Dada una composición $\mathbf{x} \in S^D$ que contiene Z ceros por redondeo, el método de reemplazo aditivo reemplaza $\mathbf{x}$ por una nueva composición $\mathbf{r} \in S^D$ sin ceros de la siguiente forma:

$$r_j = \begin{cases} \frac{\delta(Z+1)(D-Z)}{D^2}, & \text{si} \quad x_j = 0 \\ x_j - \frac{\delta(Z+1)Z}{D^2}, & \text{si} \quad x_j > 0 \end{cases} \tag{3.13}$$

donde δ es un valor pequeño. Debido a la suma restringida (3.1) de los datos composicionales, es necesario modificar tanto el valor que es cero como los que no. Este método puede generalizarse también utilizando un valor δ_j para cada parte x_j. El problema es que este reemplazo aditivo no es coherente con las operaciones básicas en el espacio S^D y, en consecuencia, distorsiona la estructura de covarianzas del conjunto de datos. Pueden consultarse más detalles en Martín-Fernández y otros (2000).

Otra estrategia de reemplazo de ceros por redondeo, el **reemplazo simple**, consiste en reemplazar los ceros por una cantidad pequeña δ_j para obtener un vector de componentes positivas, $\mathbf{w} \in S^D$ y luego aplicar el operador clausura, $\mathbf{r} = \mathcal{C}(\mathbf{w})$. Cuando se aplica este procedimiento a una composición con ceros, el valor δ_j imputado inicialmente queda modificado como consecuencia de la clausura que se realiza a posteriori. Esta particularidad implica una pérdida de naturalidad en el tratamiento del valor cero puesto que el valor que finalmente reemplaza al cero no es el valor δ_j decidido por el investigador.

Martín-Fernández y otros (2003) proponen una estrategia de **reemplazo multiplicativo**, que preserva la estructura de covarianzas y es coherente con las propiedades de los datos composicionales. Consiste en imputar de forma multiplicativa los valores nulos con un valor pequeño prefijado:

$$r_j = \begin{cases} \delta_j, & \text{si} \quad x_j = 0 \\ \left(1 - \frac{\sum_{k|x_k=0} \delta_k}{c}\right) x_j, & \text{si} \quad x_j > 0 \end{cases} \tag{3.14}$$

donde δ_j es el valor introducido en la parte x_j y c es la constante (habitualmente 1 ó 100) de la restricción de suma-constante que caracteriza los vectores composicionales (3.1). La modificación en los valores que no son cero es multiplicativa, como en el método anterior, pero el método multiplicativo es coherente con las operaciones básicas en el símplex y con la estructura de los vectores definidos en S^D. Además se conservan los ratios $r_j/r_k = x_j/x_k$ para todos los valores no nulos x_j, x_k, lo que implica que se conserva su estructura de covarianzas. Por todo ello, el reemplazo multiplicativo es más adecuado que el aditivo propuesto por Aitchison para sustituir ceros por redondeo en datos composicionales. Sin embargo, si $\delta_j = \delta \quad \forall j$, el reemplazo multiplicativo asigna exactamente el mismo valor (δ) a todos los ceros de la composición, lo que introduce una correlación artificial entre componentes que contengan ceros en la misma posición.

Un elemento clave del método de reemplazo multiplicativo es la decisión sobre qué valor δ_j escoger. El valor utilizado debe ser un valor pequeño y no superior al umbral de detección. En Martín-Fernández y otros (2003) se realizaron estudios de sensibilidad en función del valor δ_j utilizado, y se mostró que los mejores resultados se obtienen utilizando δ_j igual al 65 % del valor del umbral de detección. De esta forma, surge la pregunta de hasta qué punto el valor de medidas utilizadas en análisis multivariante depende de los valores δ_j utilizados en el tratamiento previo de los ceros. Esta cuestión debe analizarse a través de un análisis de sensibilidad. En Aitchison (1986) se sugiere que es suficiente realizar un análisis de sensibilidad de los valores δ_j en el rango $\varepsilon_{ij}/10 \leq \delta \leq \varepsilon_{ij}$, donde ε_{ij} es el umbral asociado al cero analizado.

Debido a que la distorsión que el reemplazo multiplicativo provoca en la estructura de covarianzas se incrementa al aumentar la cantidad de ceros presentes en el conjunto de datos, en muchas situaciones se hace más recomendable utilizar el **método alr-EM** propuesto en Palarea-Albaladejo y Martín-Fernández (2008). Esta alternativa paramétrica se basa en la distribución

de probabilidad normal logística aditiva y es aplicable cuando se dispone de un conjunto de datos relativamente grande. Estos autores introdujeron la idea de aplicar el algoritmo EM (de sus siglas en inglés *Expectation and Maximization*) al conjunto de datos obtenidos al aplicar la transformación log-cociente alr a los datos originales. El algoritmo EM es un procedimiento iterativo comúnmente aplicado en problemas de estimación máximo-verosímil a partir de conjuntos de datos reales multidimensionales con valores perdidos. Este procedimiento imputa los ceros teniendo en cuenta la información suministrada por el resto de componentes y corrige sustancialmente la tendencia a la subestimación de la variabilidad del reemplazo multiplicativo.

Capítulo 4

Modelos de regresión

4.1. Modelo de regresión multivariante

La regresión lineal tiene por objetivo identificar y estimar un modelo lineal a partir de una variable respuesta que depende linealmente de una o más covariables o variables explicativas. Se supone que las variables respuesta se ven afectadas por errores o desviaciones aleatorias de la media del modelo.

Transformar los vectores composicionales en vectores log-cociente permite, como se ha comentado, afrontar el problema de ajustar un modelo lineal multivariante procediendo de la forma habitual. La regresión lineal multivariante representa la generalización del modelo de regresión lineal al caso en que disponemos de más de una variable dependiente y queda formulado de la siguiente manera:

$$\mathbf{Y}_{n\times m} = \mathbf{X}_{n\times p}\mathbf{B}_{p\times m} + \mathbf{E}_{n\times m} \tag{4.1}$$

donde $\mathbf{Y}$ es una matriz de n observaciones en m variables respuesta; $\mathbf{X}$ es la matriz de diseño con columnas para $(p-1)$ variables explicativas más una columna inicial de unos para la constante de la regresión; $\mathbf{B}$ es la matriz de coeficientes de la regresión, con una columna para cada variable respuesta; y $\mathbf{E}$ es una matriz de errores. La matriz de diseño, como en el modelo de regresión lineal univariante, puede contener variables *dummy* representando factores.

$$\begin{pmatrix} y_{11} & \cdots & y_{1m} \\ \vdots & \ddots & \vdots \\ y_{n1} & \cdots & y_{nm} \end{pmatrix} = \begin{pmatrix} 1 & x_{11} & \cdots & x_{1p-1} \\ \vdots & \ddots & \vdots & \\ 1 & x_{n1} & \cdots & x_{np-1} \end{pmatrix} \cdot \begin{pmatrix} \beta_{01} & \cdots & \beta_{0m} \\ \beta_{11} & \cdots & \beta_{1m} \\ \vdots & \ddots & \vdots \\ \beta_{p-11} & \cdots & \beta_{p-1m} \end{pmatrix} + \begin{pmatrix} \varepsilon_{11} & \cdots & \varepsilon_{1m} \\ \vdots & \ddots & \vdots \\ \varepsilon_{n1} & \cdots & \varepsilon_{nm} \end{pmatrix} \tag{4.2}$$

Las hipótesis del modelo lineal multivariante afectan al comportamiento de los errores. Sea ϵ_i' la fila i-ésima de $\mathbf{E}$, entonces $\epsilon_i' \sim \mathbf{N}_m(\mathbf{0}, \mathbf{\Sigma})$, donde $\mathbf{\Sigma}$ es la matriz de covarianzas de los

errores, constante para todas las observaciones, y ϵ_i' y $\epsilon_{i'}'$ son independientes para $i \neq i'$.

El estimador de máxima verosimilitud de $\mathbf{B}$ en el modelo lineal multivariante es equivalente al estimador mínimos cuadrados para cada columna de la matriz $\mathbf{Y}$. Es decir, el problema multivariante se puede resolver a través de los m problemas univariantes.

$$\hat{\mathbf{B}} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{Y} \tag{4.3}$$

Sin embargo, los procedimientos de inferencia estadística en el modelo lineal multivariante tienen en cuenta la correlación entre las variables respuesta.

De forma paralela a la descomposición de la suma de cuadrados totales en suma de cuadrados de la regresión y de los residuos en el modelo linear univariante, en el modelo lineal multivariante existe la descomposición de la matriz de suma de cuadrados y productos cruzados (SSP, de sus siglas en inglés *sum-of-squares-and-cross-products*) totales en las matrices SSP de la regresión y de los residuos:

$$\mathbf{SSP}_T = \mathbf{Y}'\mathbf{Y} - n\bar{\mathbf{y}}\bar{\mathbf{y}}' = \hat{\mathbf{E}}'\hat{\mathbf{E}} + \left(\hat{\mathbf{Y}}'\hat{\mathbf{Y}} - n\bar{\mathbf{y}}\bar{\mathbf{y}}'\right) = \mathbf{SSP}_R + \mathbf{SSP}_{Reg} \tag{4.4}$$

donde $\bar{\mathbf{y}}$ es el vector $(m \times 1)$ de medias de las variables respuesta; $\hat{\mathbf{Y}} = \mathbf{X}\hat{\mathbf{B}}$ es la matriz de los valores ajustados; y $\hat{\mathbf{E}} = \mathbf{Y} - \hat{\mathbf{Y}}$ es la matriz de residuos. Se pueden desarrollar varios contrastes de hipótesis a partir de (4.4).

Denotamos por $\mathbf{SSP}_H$ a la matriz SSP incremental para una hipótesis, es decir, la diferencia entre el $\mathbf{SSP}_{Reg}$ para el modelo sin restricciones y el $\mathbf{SSP}_{Reg}$ para el modelo con restricciones impuestas por la hipótesis. Los contrastes multivariantes se basan en los m autovalores λ_j de $\mathbf{SSP}_H\mathbf{SSP}_R^{-1}$ (la matriz SSP de la hipótesis "dividida entre" la matriz SSP residual), esto es, los valores de λ para los cuales:

$$\det(\mathbf{SSP}_H\mathbf{SSP}_R^{-1} - \lambda\mathbf{I}_m) = 0. \tag{4.5}$$

Los estadísticos para los contrastes multivariantes comúnmente utilizados son funciones de estos autovalores:

$$\begin{aligned}
&\text{Pillai-Bartlett Trace,} \quad T_{PB} = \sum_{j=1}^{m} \frac{\lambda_j}{1-\lambda_j} \\
&\text{Hotelling-Lawley Trace,} \quad T_{HL} = \sum_{j=1}^{m} \lambda_j \\
&\text{Wilks's Lambda,} \quad \Lambda = \prod_{j=1}^{m} \frac{1}{1+\lambda_j} \\
&\text{Roy's Maximum Root,} \quad \lambda_1
\end{aligned} \tag{4.6}$$

Por convención los autovalores de $\mathbf{SSP}_H\mathbf{SSP}_R^{-1}$ se ordenan en orden descendente, por lo que λ_1 es el mayor autovalor. Normalmente se utilizan aproximaciones de estos estadísticos a la F de Snedecor.

Supongamos que queremos contrastar la hipótesis lineal:

$$H_0 : \mathbf{L}_{q\times p}\mathbf{B}_{p\times m} = \mathbf{C}_{q\times m} \tag{4.7}$$

donde $\mathbf{L}$ es una matriz conocida de rango $q < p$, y la matriz $\mathbf{C}$ está compuesta por constantes, normalmente ceros. En este caso la matriz SSP para la hipótesis es:

$$\mathbf{SSP}_H = \left(\hat{\mathbf{B}}'\mathbf{L}' - \mathbf{C}'\right)\left[\mathbf{L}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{L}'\right]\left(\mathbf{L}\hat{\mathbf{B}} - \mathbf{C}\right) \tag{4.8}$$

Los diferentes estadísticos de contraste se basan en los $k = \text{mín}(q, m)$ autovalores no nulos de $\mathbf{SSP}_H\mathbf{SSP}_R^{-1}$.

4.2. Regresión composicional

Cuando la variable respuesta es composicional disponemos de una muestra en S^D que denotamos por $\mathbf{w}_1, \mathbf{w}_2, \ldots, \mathbf{w}_n$. El dato i-ésimo $\mathbf{w}_i$ se asocia con una o más variables externas o covariables agrupadas en el vector $\mathbf{t}_i = (t_{i0}, t_{i1}, \ldots, t_{ir})$, donde $t_{i0} = 1$. El objetivo es estimar los coeficientes de una curva o superficie en S^D cuya ecuación es:

$$\hat{\mathbf{w}}(\mathbf{t}) = \boldsymbol{\beta}_0 \oplus (t_1 \otimes \boldsymbol{\beta}_1) \oplus \cdots \oplus (t_r \otimes \boldsymbol{\beta}_r) = \bigoplus_{j=0}^{r} \left(t_j \otimes \boldsymbol{\beta}_j\right) \tag{4.9}$$

donde $\mathbf{t} = (t_0, t_1, \ldots, t_r)$ son las covariables cuyo primer elemento se define como la constante $t_0 = 1$. Los coeficientes composicionales del modelo, $\boldsymbol{\beta}_j \in S^D$, se estiman a partir de los datos. El modelo (4.9) es muy general y adopta diversas formas dependiendo de cómo se definan las

covariables t_j. Por ejemplo, si $t_j = t^j$, el modelo es polinomial.

El método más popular para ajustar el modelo (4.9) es mediante el criterio de desviaciones de mínimos cuadrados. Como la variable respuesta $\mathbf{w}(\mathbf{t})$ es composicional, se miden las desviaciones en el símplex utilizando los conceptos de la geometría de Aitchison. Las desviaciones del modelo (4.9) a los datos se mide mediante $\|\hat{\mathbf{w}}(\mathbf{t}_i) \ominus \mathbf{w}_i\|_a^2 = d_a^2(\hat{\mathbf{w}}(\mathbf{t}_i), \mathbf{w}_i)$. La *suma de errores al cuadrado* (SSE en sus siglas en inglés) es:

$$SSE = \sum_{i=1}^{n} \|\hat{\mathbf{w}}(\mathbf{t}_i) \ominus \mathbf{w}_i\|_a^2 = \sum_{i=1}^{n} \|e_i\|_a^2 = \sum_{i=1}^{n} d_a^2(\hat{\mathbf{w}}(\mathbf{t}_i), \mathbf{w}_i) \tag{4.10}$$

donde e_i son los residuos. Hay que minimizar la función objetivo (4.10) como función de los coeficientes composicionales $\boldsymbol{\beta}_j$. El número de coeficientes estimados en este modelo lineal es $(r+1)\cdot(D-1)$.

Este problema de mínimos cuadrados se reduce a un problema ordinario de mínimos cuadrados en $D-1$ cuando transformamos los datos composicionales en vectores log-cociente. Consideremos, por simplicidad, el siguiente modelo de regresión, con constante $\boldsymbol{\beta}_0$ y un coeficiente de regresión $\boldsymbol{\beta}_1$:

$$\mathbf{w}_i = \boldsymbol{\beta}_0 \oplus (t_i \otimes \boldsymbol{\beta_1}) \otimes \varepsilon_i \qquad (i = 1, \ldots, n) \tag{4.11}$$

donde ε_i son los errores del modelo. Tomando la transformación log-cociente aditiva alr, podemos expresar el modelo (4.11) de la siguiente forma:

$$alr\mathbf{w}_i = alr\boldsymbol{\beta}_0 + t_i alr\boldsymbol{\beta_1} + alr\varepsilon_i \qquad (i = 1, \ldots, n) \tag{4.12}$$

Que puede ser parametrizado como:

$$alr\mathbf{w}_i = \boldsymbol{\alpha}_0 + t_i\boldsymbol{\alpha_1} + \epsilon_i \qquad (i = 1, \ldots, n) \tag{4.13}$$

Por ejemplo, en el caso de una única covariable o variable explicativa t:

$$\begin{aligned} log\left(\tfrac{w_1}{w_D}\right) &= \alpha_{01} + t\cdot\alpha_{11} + \epsilon_1 \\ &\vdots \\ log\left(\tfrac{w_{D-1}}{w_D}\right) &= \alpha_{0D-1} + t\cdot\alpha_{1D-1} + \epsilon_{D-1} \end{aligned} \tag{4.14}$$

y

$$\begin{aligned} \boldsymbol{\alpha}_0 = (\alpha_{01}, \ldots, \alpha_{0D-1}) = alr(\boldsymbol{\beta}_0) = \left(log\left(\frac{\beta_{01}}{\beta_{0D}}\right), \cdots, log\left(\frac{\beta_{0D-1}}{\beta_{0D}}\right)\right), \\ \boldsymbol{\alpha}_1 = (\alpha_{11}, \ldots, \alpha_{1D-1}) = alr(\boldsymbol{\beta}_1) = \left(log\left(\frac{\beta_{11}}{\beta_{1D}}\right), \cdots, log\left(\frac{\beta_{1D-1}}{\beta_{1D}}\right)\right). \end{aligned} \tag{4.15}$$

Las estimaciones $\widehat{\boldsymbol{\alpha}_0}$ y $\widehat{\boldsymbol{\alpha}_1}$ se obtienen como se ha visto por el método de mínimos cuadrados, ya que ahora estamos ante un problema de regresión multivariante con $D-1$ variables respuesta. Posteriormente obtenemos $\widehat{\boldsymbol{\beta}_0} = alr^{-1}\widehat{\boldsymbol{\alpha}_0}$ y $\widehat{\boldsymbol{\beta}_1} = alr^{-1}\widehat{\boldsymbol{\alpha}_1}$, siendo alr^{-1} la inversa de la transformación alr:

$$\begin{aligned}
&\widehat{\boldsymbol{\beta}_0} = (\widehat{\beta_{01}}, \ldots, \widehat{\beta_{0D}}) = \\
&= \left(\frac{e^{\alpha_{01}}}{e^{\alpha_{01}} + \cdots + e^{\alpha_{0D-1}} + 1}, \ldots, \frac{e^{\alpha_{0D-1}}}{e^{\alpha_{01}} + \cdots + e^{\alpha_{0D-1}} + 1}, \frac{1}{e^{\alpha_{01}} + \cdots + e^{\alpha_{0D-1}} + 1} \right), \\
&\widehat{\boldsymbol{\beta}_1} = (\widehat{\beta_{11}}, \ldots, \widehat{\beta_{1D}}) = \\
&= \left(\frac{e^{\alpha_{11}}}{e^{\alpha_{11}} + \cdots + e^{\alpha_{1D-1}} + 1}, \ldots, \frac{e^{\alpha_{1D-1}}}{e^{\alpha_{11}} + \cdots + e^{\alpha_{1D-1}} + 1}, \frac{1}{e^{\alpha_{11}} + \cdots + e^{\alpha_{1D-1}} + 1} \right).
\end{aligned} \tag{4.16}$$

Análogamente al modelo de regresión, se puede sugerir una medida para evaluar el comportamiento del modelo, como la habitual medida R^2, para estimar el porcentaje de variabilidad explicada por el modelo propuesto.

$$R^2 = \frac{totvar(\hat{\mathbf{w}})}{totvar(\mathbf{w})} = 1 - \frac{SSE}{totvar(\mathbf{w})} \tag{4.17}$$

donde $\mathbf{w}$ son los datos observados y $\hat{\mathbf{w}}$ los datos ajustados del modelo.

4.3. Regresión de Dirichlet

La regresión de Dirichlet es adecuada para el análisis de datos composicionales y es una alternativa al enfoque de Aitchison (1986), pero requiere el conocimiento de la función de densidad condicional conjunta de las variables respuesta.

Sea $\mathbf{x} = (x_1, \ldots, x_D)$ un vector positivo con $\sum_{i=1}^{D} x_i = 1$ y distribución de Dirichlet con parámetros positivos $(\lambda_1, \ldots, \lambda_D)$ y función de densidad:

$$f(\mathbf{x}) = \left(\frac{\Gamma(\lambda)}{\prod_{i=1}^{D} \Gamma(\lambda_i)} \right) \prod_{i=1}^{D} x_i^{\lambda_i - 1} \tag{4.18}$$

donde $\lambda = \sum_{i=1}^{D} \lambda_i$. Cuando $D = 2$ la expresión (4.18) se reduce a la distribución beta. La media, varianza y covarianza de las variables son:

$$E(x_i) = \frac{\lambda_i}{\lambda} \tag{4.19}$$

$$Var(x_i) = \frac{\lambda_i(\lambda - \lambda_i)}{\lambda^2(\lambda + 1)} \tag{4.20}$$

$$Cov(y_i, y_j) = \frac{-\lambda_i \lambda_j}{\lambda^2(\lambda + 1)}; \quad i \neq j \tag{4.21}$$

Con una adecuada elección de parámetros, la distribución de Dirichlet permite una gran flexibilidad. Para posibilitar relaciones entre vectores aleatorios que siguen una distribución de Dirichlet y un conjunto de variables explicativas, se pueden introducir covariables en λ_i

En este trabajo no se ajusta un modelo de regresión de Dirichlet, cuyos coeficientes son de difícil interpretación y que ya fue objeto de análisis en el trabajo de fin de máster de Rosa Caeiro Sánchez.

Capítulo 5

Aplicación a datos reales

5.1. Información disponible

La información disponible para este trabajo, obtenida tras la realización de las prácticas en Azteca Ingeniería Consulting, se refiere a una serie de proyectos en los que ha trabajado la empresa. En cada proyecto, el total de horas que se le ha dedicado se reparte entre los cuatro departamentos de la empresa: administrativo, técnico, comercial e I+D+i. Por lo tanto, en este caso, los datos composicionales nos informan del porcentaje de horas de cada proyecto en cada uno de los departamentos.

Se dispone de información sobre 53 proyectos (ver Tabla A.1 en el Anexo A). Estos datos se representan en un gráfico tridimensional en la Figura 5.1, en donde se observa una mayor concentración de los proyectos en el vértice del departamento técnico, lo que significa que en una gran cantidad de proyectos el porcentaje de horas de contenido técnico es elevado. La media geométrica composicional es 5,87 % en el departamento administrativo, 78,19 % en el técnico, 6,13 % en el comercial y 9,81 % en el departamento de I+D+i.

Una representación bidimensional más sencilla se muestra en las Figuras 5.2 y 5.3, en las que se ha suprimido en cada una de ellas uno de los departamentos. En la Figura 5.2 se comprueba que todos los proyectos presentan un porcentaje elevado de horas dedicadas o bien al departamento técnico o bien al departamento de I+D+i, lo que se corresponde con las dos principales tipologías de proyectos: los de ingeniería y los de consultoría. En la Figura 5.3, donde no se tiene en cuenta el efecto de la parte del departamento técnico, el reparto entre el resto de los departamentos está más equilibrado, aunque en general la dedicación a las tareas comerciales presenta bajos porcentajes en la mayoría de los proyectos. Esto se debe a que, al ser la principal actividad de esta empresa la consultoría en áreas tecnológicas, sólo hay un trabajador dedicado a tareas comerciales.

Por otro lado, para cada uno de los proyectos se dispone también de una serie de variables

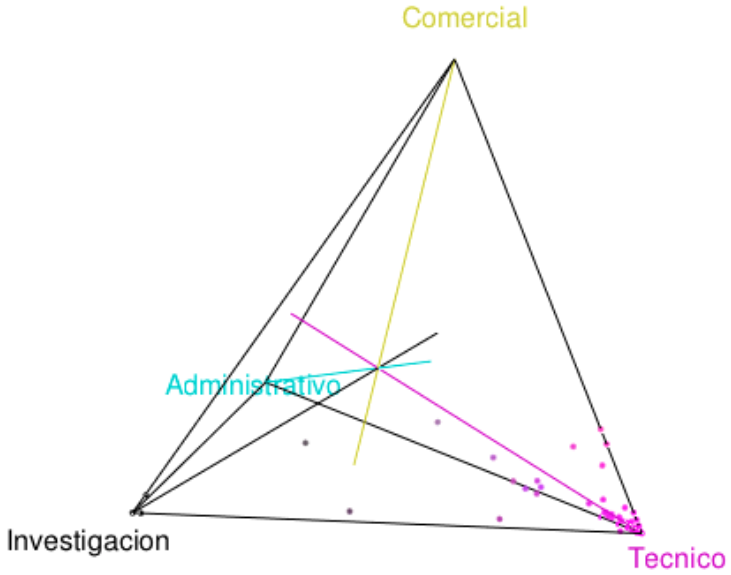

Figura 5.1: *Representación de los datos composicionales*

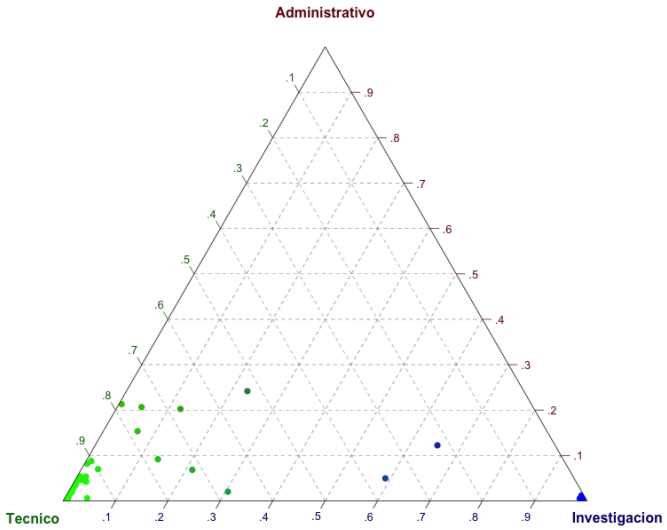

Figura 5.2: *Gráfico ternario para los departamentos técnico, administrativo e I+D+i*

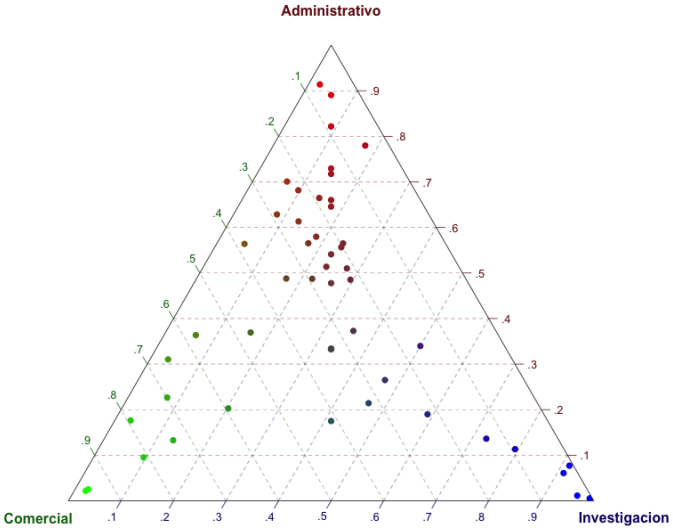

Figura 5.3: *Gráfico ternario para los departamentos comercial, administrativo e I+D+i*

que los caracterizan. Estas variables son:

1. **Servicio**: indica qué servicio se presta al cliente con ese proyecto. El servicio puede ser de ingeniería, de consultoría o mixto. En la muestra de proyectos, 36 son de ingeniería, 7 de consultoría y 10 mixtos.

2. **Tipo de proyecto**: variable que indica el alcance del proyecto. Distingue entre:

   – Microproyecto: proyectos de corta duración dedicados a un trabajo específico.

   – Proyecto: proyectos ordinarios.

   – Macroproyecto: proyectos o conjuntos de proyectos de larga duración y que implican una gran variedad de tareas y dedicación de horas.

   Se dispone de 14 microproyectos, 32 proyectos y 7 macroproyectos.

3. **Cliente**: esta variable distingue si el proyecto es para el principal cliente de la empresa (38 proyectos) o no (15 proyectos).

4. **Trabajadores técnicos**: número de trabajadores con perfil técnico que dedican horas al proyecto.

5. **Trabajadores consultores**: número de trabajadores consultores que dedican horas al proyecto.

6. **Trabajadores de soporte**: número de trabajadores de perfil administrativo o comercial que dedican horas al proyecto.

7. **Trabajadores totales**: número total de trabajadores que dedican horas al proyecto.

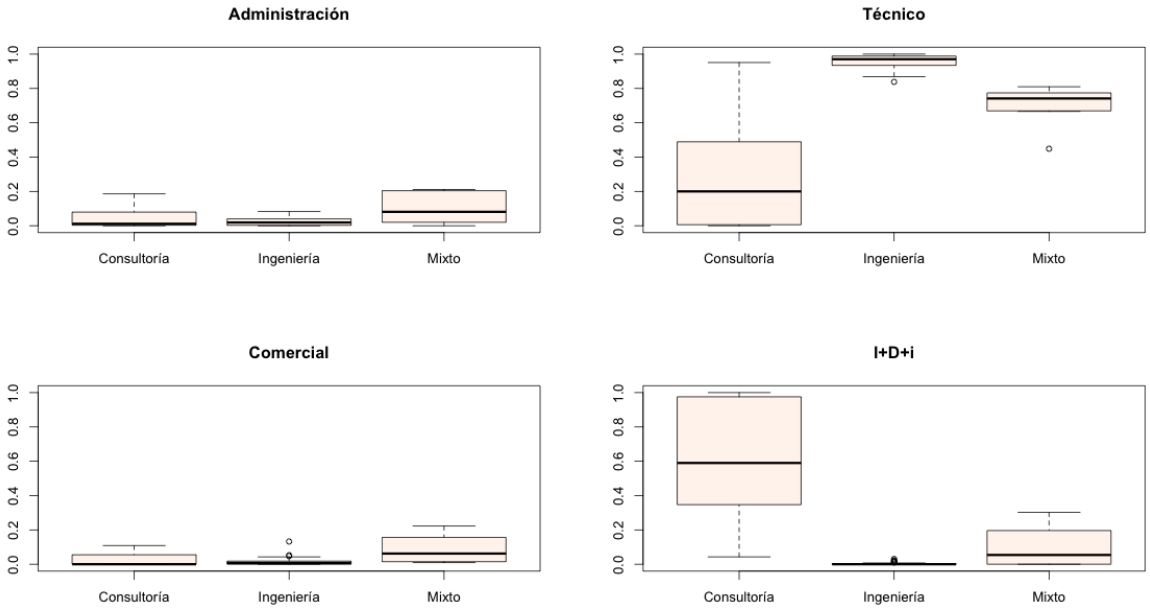

Figura 5.4: *Boxplots para la proporción de horas en cada departamento según la variable servicio*

En la Figura 5.4 se representan los boxplots para la proporción de horas dedicadas a cada uno de los cuatro departamentos según la variable servicio. En la Figura 5.5 los gráficos de dispersión para la proporción de horas según la variable número de trabajadores de soporte. Estas dos variables, como se verá más adelante, son las dos más influyentes en la distribución de las horas entre los diferentes departamentos.

En los gráficos se observa que los proyectos de ingeniería tienen un elevado porcentaje en horas del departamento técnico, mientras que los de consultoría presentan mayores porcentajes en horas de I+D+i. Además, los proyectos sin trabajadores de soporte tienen mayor proporción de horas técnicas.

5.2. Estudio de simulación

El primer problema con el que nos encontramos para ajustar un modelo de regresión a estos datos es la presencia de ceros, ya que en este caso no son aplicables las transformaciones log-cociente. Se considera que realmente a cada proyecto es necesario dedicar, en mayor o menor medida, horas de los cuatro departamentos. Sin embargo, por construcción de los partes de trabajo, que no permite detectar tiempos muy cortos en ciertas actividades, existen proyectos en los que no se ha registrado tiempo dedicado a algún departamento. Por tanto, estamos ante la problemática de los ceros por redondeo.

Se han comentado diferentes métodos de reemplazo de ceros, siendo el método multiplicativo

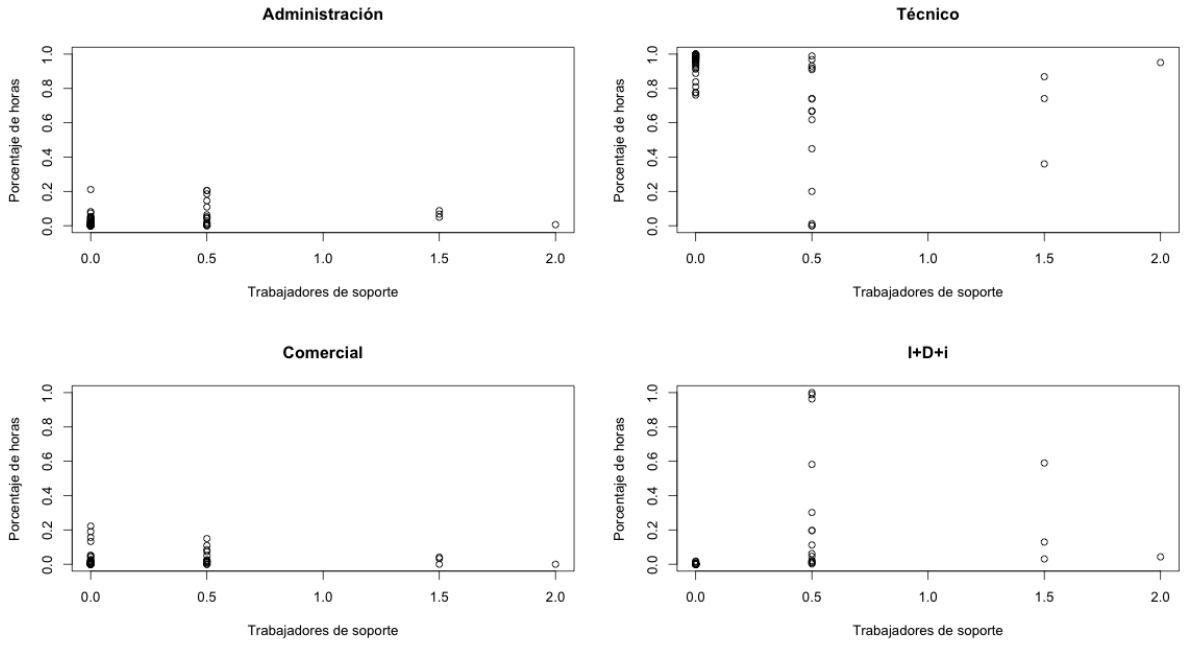

Figura 5.5: *Gráfico de dispersión para la proporción de horas según el número de trabajadores de soporte*

(3.14) el más coherente con las propiedades de los datos composicionales. Debido a que no se dispone de un tamaño de datos adecuado (Paralea-Albadalejo y otros (2007) sugieren $n > 70$), no se considera el método de reemplazo basado en el algoritmo EM.

Sin embargo, existe otra estrategia de reemplazo, propuesta por Smithson y Verkuilen (2006), que es la que se aplica en el paquete de R DirichletReg para el ajuste de un modelo de regresión de Dirichlet a un conjunto de datos composicionales. Con esta metodología, si existen componentes nulas en el conjunto de observaciones de vectores composicionales $\mathbf{y}$, se transforma cada componente y de $\mathbf{y}$ mediante

$$y* = \frac{y(n-1) + \frac{1}{d}}{n} \tag{5.1}$$

donde n es el número de observaciones en $\mathbf{y}$ y d el número de dimensiones. Este criterio conlleva que incluso las componentes no nulas de $\mathbf{y}$ serán reemplazadas por nuevos valores (si bien tal modificación será pequeña para n grande), y que los valores nulos se sustituirán por $\frac{1}{nd}$.

En este trabajo se propone un estudio de simulación para comprobar cuál de estos dos métodos de reemplazo de los ceros, el método multiplicativo o el de Smithson y Verkuilen, proporciona mejores resultados en el contexto de los modelos de regresión. Suponemos que tenemos una muestra de n vectores composicionales con 3 componentes: $\mathbf{y} = (y_1, y_2, y_3)$. El

modelo de regresión composicional visto en (4.11) puede ser parametrizado como

$$alr\mathbf{y}_i = \boldsymbol{\beta}_0 + \mathbf{x}_i\boldsymbol{\beta}_\mathbf{1} + \boldsymbol{\epsilon}_i \tag{5.2}$$

Hay que advertir en este punto un cambio de notación respecto al capítulo anterior, donde los coeficientes que en (4.13) se denotan por $\boldsymbol{\alpha}$ son en este caso los coeficientes $\boldsymbol{\beta}$.

Puesto que estamos en el caso de $D = 3$, se ha transformado en un problema de regresión multivariante con 2 variables dependientes. Podemos reescribir (5.2) como:

$$\begin{aligned} log\left(\frac{y_1}{y_3}\right) &= \beta_{01} + \beta_{11}x + \varepsilon_1 \\ log\left(\frac{y_2}{y_3}\right) &= \beta_{02} + \beta_{12}x + \varepsilon_2 \end{aligned} \tag{5.3}$$

donde$\begin{pmatrix}\varepsilon_1\\ \varepsilon_2\end{pmatrix} \in N_2\left(\begin{pmatrix}0\\0\end{pmatrix}, \Sigma\right)$. Podemos expresar (5.3) de la siguiente forma:

$$\begin{aligned} \frac{y_1}{y_3} &= e^{\beta_{01}+\beta_{11}x} \cdot e^{\varepsilon_1} \\ \frac{y_2}{y_3} &= e^{\beta_{02}+\beta_{12}x} \cdot e^{\varepsilon_2} \end{aligned} \tag{5.4}$$

Partiendo de estas ecuaciones se inicia el estudio de simulación, cuyos pasos son:

1. Suponemos un modelo con una única variable explicativa que sigue una distribución Uniforme en el intervalo unidad, por lo que se generan n observaciones de $X \in U[0, 1]$.

2. Se simulan los errores del modelo $\begin{pmatrix}\varepsilon_1\\ \varepsilon_2\end{pmatrix}$ generando n observaciones de una variable aleatoria $N_2\left(\begin{pmatrix}0\\0\end{pmatrix}, \Sigma\right)$.

3. Se generan las n variables composicionales $\mathbf{y} = (y_1, y_2, y_3)$ que cumplan las condiciones del modelo. Para ello en primer lugar se fijan unos valores para β_{01}, β_{11}, β_{02} y β_{12} y se despeja y_3 de (5.4), obteniendo:

$$y_3 = \frac{1}{1 + (e^{\beta_{01}+\beta_{11}x+\varepsilon_1} + e^{\beta_{02}+\beta_{12}x+\varepsilon_2})} \tag{5.5}$$

A continuación se pueden obtener y_1 e y_2 como:

$$y_1 = y_3 \cdot e^{\beta_{01}+\beta_{11}x} \cdot e^{\varepsilon_1} \tag{5.6}$$

$$y_2 = y_3 \cdot e^{\beta_{02}+\beta_{12}x} \cdot e^{\varepsilon_2} \tag{5.7}$$

4. Se introducen artificialmente valores nulos en $\mathbf{y}$ transformando en 0 las componentes menores a un ν considerado: $y_j = 0 \quad \forall y_j < \nu$, para $j = 1, 2, 3$. A continuación se normaliza el vector composicional mediante el operador clausura $\mathcal{C}$ visto en la definición 3.1.4.

5. Se reemplazan los valores nulos:

   - Mediante el método de Smithson y Verkuilen (5.1), obteniendo $\mathbf{y}^{*1}$. Esta estrategia modifica todas las observaciones, incluso las que no presentan componentes nulas.
   - Mediante el método multiplicativo (3.14), obteniendo $\mathbf{y}^{*2}$, que sólo modifica las observaciones con componentes nulas. Se toma $\delta_j = \delta = 0.005$ como valor de reemplazo de los ceros.

6. Para poder realizar un análisis comparativo se aplican también los dos métodos de reemplazo a los valores $\mathbf{y}$ originales:

   - El método de Smithson y Verkuilen afecta a todos los componentes de $\mathbf{y}$, a pesar de no contener valores nulos. Denotamos esta transformación por $\mathbf{y}^1$.
   - El método multiplicativo no tiene ningún efecto en $\mathbf{y}$, al no existir componentes nulas.

7. Se estiman los coeficientes de los siguientes modelos de regresión multivariante:

   - Modelo 1: $alr\mathbf{y}^1 = \boldsymbol{\beta}_0 + x\boldsymbol{\beta}_\mathbf{1} + \boldsymbol{\epsilon}$
   - Modelo 2: $alr\mathbf{y} = \boldsymbol{\beta}_0 + x\boldsymbol{\beta}_\mathbf{1} + \boldsymbol{\epsilon}$
   - Modelo 3: $alr\mathbf{y}^{*1} = \boldsymbol{\beta}_0 + x\boldsymbol{\beta}_\mathbf{1} + \boldsymbol{\epsilon}$
   - Modelo 4: $alr\mathbf{y}^{*2} = \boldsymbol{\beta}_0 + x\boldsymbol{\beta}_\mathbf{1} + \boldsymbol{\epsilon}$

Para el estudio de simulación se parte de los siguientes supuestos: $\beta_{01} = \beta_{02} = 0$, $\beta_{11} = \beta_{12} = 1$ y $\Sigma = \begin{pmatrix} 1.5 & 1 \\ 1 & 1.5 \end{pmatrix}$. Se consideran diferentes tamaños de la muestra n (25,50,100 y 250) y dos valores de ν, 0.05 y 0.10. Para cada escenario, se realizan 1000 simulaciones, recogiendo para cada una de ellas las estimaciones de los coeficientes de cada uno de los cuatro modelos. En las tablas 5.1 a 5.4 se recogen las medias, las desviaciones típicas y los errores cuadráticos medios (*MSE*, de sus siglas en inglés *Mean Square Error*) de los coeficientes estimados en cada escenario.

Se comprueba que es el modelo 1 el que presenta coeficientes con menores desviaciones típicas y menor MSE para todos los valores de n y ν considerados, pero a medida que aumenta n las diferencias entre el modelo 1 y el modelo 2 se van reduciendo. Las estimaciones de β_{11} y β_{12} en el modelo 1 presentan un sesgo mayor al del modelo 2, provocado por la modificación de los valores de $\mathbf{y}$ según la estrategia de reemplazo de Smithson y Verkuilen. Ambos modelos, el 1 y el 2, presentan mejores resultados que los modelos 3 y 4 debido a que disponen de información completa sobre la variable respuesta.

	$\nu = 0.05$				$\nu = 0.10$			
	Modelo 1 $(x, \mathbf{y}^1)$	Modelo 2 $(x, \mathbf{y})$	Modelo 3 $(x, \mathbf{y}^{*1})$	Modelo 4 $(x, \mathbf{y}^{*2})$	Modelo 1 $(x, \mathbf{y}^1)$	Modelo 2 $(x, \mathbf{y})$	Modelo 3 $(x, \mathbf{y}^{*1})$	Modelo 4 $(x, \mathbf{y}^{*2})$
	Medias							
β_{01}	-0.00355	-0.00255	0.00074	0.00386	-0.00257	-0.00132	0.03372	0.05202
β_{11}	0.90858	1.00273	1.07439	1.24721	0.91105	1.00483	1.38930	1.70353
β_{02}	-0.00008	0.00127	0.00474	0.00867	0.00325	0.00470	0.03826	0.05617
β_{12}	0.90731	1.00117	1.07261	1.24453	0.89659	0.98973	1.37124	1.68331
	Desviaciones típicas							
β_{01}	0.46780	0.50843	0.53966	0.62441	0.46890	0.50886	0.73240	0.90431
β_{11}	0.81099	0.88967	0.96138	1.13473	0.80645	0.88395	1.27950	1.59341
β_{02}	0.46683	0.50742	0.53892	0.62345	0.46941	0.50997	0.74012	0.91580
β_{12}	0.80513	0.88338	0.95537	1.12803	0.80708	0.88547	1.28933	1.60801
	MSE							
β_{01}	0.21885	0.25851	0.29123	0.38990	0.21988	0.25894	0.53755	0.82047
β_{11}	0.66606	0.79152	0.92979	1.34873	0.65827	0.78138	1.78868	3.03391
β_{02}	0.21793	0.25748	0.29045	0.38877	0.22036	0.26009	0.54924	0.84184
β_{12}	0.65682	0.78036	0.91801	1.33224	0.66208	0.78415	1.80020	3.05262

Tabla 5.1: Para $n = 25$: media, desviación típica y MSE de los coeficientes estimados para los diferentes modelos, con $\nu = 0.05$ (9.72 % de ceros) y $\nu = 0.10$ (32.37 % de ceros)

	$\nu = 0.05$				$\nu = 0.10$			
	Modelo 1 $(x, \mathbf{y}^1)$	Modelo 2 $(x, \mathbf{y})$	Modelo 3 $(x, \mathbf{y}^{*1})$	Modelo 4 $(x, \mathbf{y}^{*2})$	Modelo 1 $(x, \mathbf{y}^1)$	Modelo 2 $(x, \mathbf{y})$	Modelo 3 $(x, \mathbf{y}^{*1})$	Modelo 4 $(x, \mathbf{y}^{*2})$
	Medias							
β_{01}	0.00216	0.00299	0.00945	0.01121	-0.00273	-0.00214	0.04326	0.04820
β_{11}	0.94383	0.99402	1.17359	1.23251	0.95277	1.00333	1.61053	1.70573
β_{02}	-0.00047	0.00026	0.00615	0.00779	-0.00039	0.00034	0.04634	0.05145
β_{12}	0.94801	0.99834	1.17838	1.23745	0.94839	0.99874	1.60500	1.69993
	Desviaciones típicas							
β_{01}	0.33770	0.35270	0.41441	0.43506	0.33682	0.35179	0.59226	0.62842
β_{11}	0.58691	0.61657	0.75123	0.79180	0.58269	0.61193	1.04973	1.11571
β_{02}	0.33689	0.35194	0.41408	0.43478	0.33687	0.35189	0.59062	0.62667
β_{12}	0.58513	0.61476	0.74906	0.78959	0.57653	0.60570	1.03429	1.09942
	MSE							
β_{01}	0.11404	0.12441	0.17182	0.18941	0.11346	0.12376	0.35264	0.39723
β_{11}	0.34762	0.38020	0.59448	0.68101	0.34176	0.37447	1.47468	1.74287
β_{02}	0.11349	0.12386	0.17150	0.18910	0.11348	0.12383	0.35098	0.39537
β_{12}	0.34508	0.37794	0.59292	0.67983	0.33505	0.36687	1.43577	1.69863

Tabla 5.2: Para $n = 50$: media, desviación típica y MSE de los coeficientes estimados para los diferentes modelos, con $\nu = 0.05$ (9.63 % de ceros) y $\nu = 0.10$ (32.34 % de ceros)

	$\nu = 0.05$				$\nu = 0.10$			
	Modelo 1 $(x, \mathbf{y}^1)$	Modelo 2 $(x, \mathbf{y})$	Modelo 3 $(x, \mathbf{y}^{*1})$	Modelo 4 $(x, \mathbf{y}^{*2})$	Modelo 1 $(x, \mathbf{y}^1)$	Modelo 2 $(x, \mathbf{y})$	Modelo 3 $(x, \mathbf{y}^{*1})$	Modelo 4 $(x, \mathbf{y}^{*2})$
	Medias							
β_{01}	-0.00057	-0.00032	0.00878	0.00801	0.00049	0.00072	0.05926	0.05171
β_{11}	0.97473	1.00128	1.28608	1.24063	0.97154	0.99805	1.82046	1.69970
β_{02}	0.00101	0.00127	0.01016	0.00944	-0.00087	-0.00065	0.05756	0.05003
β_{12}	0.97251	0.99906	1.28446	1.23889	0.97057	0.99708	1.82051	1.69965
	Desviaciones típicas							
β_{01}	0.24229	0.24783	0.31810	0.30526	0.24182	0.24724	0.47147	0.43797
β_{11}	0.42225	0.43329	0.58467	0.55530	0.41529	0.42610	0.83154	0.76924
β_{02}	0.24555	0.25115	0.32115	0.30837	0.24253	0.24799	0.47318	0.43965
β_{12}	0.42197	0.43305	0.58465	0.55515	0.41852	0.42948	0.83859	0.77585
	MSE							
β_{01}	0.05870	0.06142	0.10126	0.09325	0.05848	0.06113	0.22579	0.19449
β_{11}	0.17894	0.18774	0.42368	0.36626	0.17328	0.18157	1.36461	1.08131
β_{02}	0.06030	0.06308	0.10324	0.09518	0.05882	0.06150	0.22722	0.19580
β_{12}	0.17882	0.18754	0.42273	0.36526	0.17603	0.18446	1.37648	1.09144

Tabla 5.3: Para $n = 100$: media, desviación típica y MSE de los coeficientes estimados para los diferentes modelos, con $\nu = 0.05$ (9.64 % de ceros) y $\nu = 0.10$ (32.38 % de ceros)

	$\nu = 0.05$				$\nu = 0.10$			
	Modelo 1 $(x, \mathbf{y}^1)$	Modelo 2 $(x, \mathbf{y})$	Modelo 3 $(x, \mathbf{y}^{*1})$	Modelo 4 $(x, \mathbf{y}^{*2})$	Modelo 1 $(x, \mathbf{y}^1)$	Modelo 2 $(x, \mathbf{y})$	Modelo 3 $(x, \mathbf{y}^{*1})$	Modelo 4 $(x, \mathbf{y}^{*2})$
	Medias							
β_{01}	0.00229	0.00240	0.01415	0.01051	0.00039	0.00048	0.07615	0.05159
β_{11}	0.98708	0.99806	1.41781	1.23962	0.98833	0.99934	2.10642	1.70262
β_{02}	0.00125	0.00135	0.01290	0.00935	-0.00148	-0.00140	0.07075	0.04732
β_{12}	0.99029	1.00128	1.42128	1.24298	0.99120	1.00223	2.11525	1.70956
	Desviaciones típicas							
β_{01}	0.15350	0.15490	0.22046	0.18997	0.15447	0.15589	0.34855	0.27756
β_{11}	0.26581	0.26864	0.40870	0.34248	0.26934	0.27224	0.62170	0.49048
β_{02}	0.15278	0.15419	0.22020	0.18947	0.15307	0.15449	0.34752	0.27644
β_{12}	0.26385	0.26669	0.40852	0.34160	0.26770	0.27059	0.62073	0.48957
	MSE							
β_{01}	0.02357	0.02400	0.04880	0.03620	0.02386	0.02430	0.12728	0.07970
β_{11}	0.07082	0.07217	0.34160	0.17471	0.07268	0.07412	1.61067	0.73424
β_{02}	0.02334	0.02378	0.04866	0.03599	0.02343	0.02387	0.12578	0.07866
β_{12}	0.06971	0.07112	0.34437	0.17573	0.07174	0.07322	1.62908	0.74316

Tabla 5.4: Para $n = 250$: media, desviación típica y MSE de los coeficientes estimados para los diferentes modelos, con $\nu = 0.05$ (9.66 % de ceros) y $\nu = 0.10$ (32.38 % de ceros)

Tanto para $\nu = 0.05$ como para $\nu = 0.10$ la transformación $\mathbf{y}^{*1}$ presenta menor MSE que $\mathbf{y}^{*2}$ para $n = 25$ y $n = 50$, pero ocurre lo contrario para $n = 100$ y $n = 250$. El sesgo sistemático existe para las dos transformaciones, pero es mayor para el método de Smithson y Verkuilen cuando $n \geqslant 100$.

Al transformar $\mathbf{y}$ sustituimos los valores nulos (que originariamente eran valores que oscilaban entre 0 y ν) por un valor que los sintetiza. Este valor de sustitución en el método multiplicativo es 0.005, mientras que en el método de Smithson y Verkuilen el valor por el que se sustituyen los ceros es $\frac{1}{3n}$, que va disminuyendo a medida que aumenta el tamaño de la muestra: toma los valores 0.0133, 0.0067, 0.0033 y 0.0013 para $n = 25$, $n = 50$, $n = 100$ y $n = 250$ respectivamente. Es por esto que con la estrategia de Smithson y Verkuilen se obtienen peores resultados a medida que aumenta n, ya que el valor de reemplazo se aleja más de lo que sería la media de los verdaderos valores que se han sustituido por ceros. El centro del intervalo $[0, \nu]$ es 0.025 y 0.05 dependiendo del valor de ν.

Este estudio de simulación pone de manifiesto la importancia del valor de sustitución de los valores nulos cuando se trabaja con datos composicionales. Si este valor de sustitución se aleja mucho del límite de detección provocará un aumento en la variabilidad de los coeficientes estimados en un modelo de regresión, así como un sesgo sistemático en los mismos.

En el siguiente apartado, para el ajuste de un modelo de regresión con información obtenida tras la realización de las prácticas en Azteca Ingeniería Consulting, se han sustituido los valores nulos mediante el método multiplicativo. Se ha tomado como valor de reemplazo δ la mitad del valor del límite de detección, que ha sido definido como el menor valor de todas las observaciones de los datos composicionales.

5.3. Ajuste del modelo de regresión

Con los datos disponibles tras el periodo de prácticas se pretende conocer si la distribución de las horas entre los diferentes departamentos de la empresa depende de alguna o algunas de las variables que caracterizan a los proyectos. Conocer esta relación facilitará la realización de predicciones y cronogramas para futuros proyectos.

El vector composicional $\mathbf{y}$ está formado por los departamentos de la empresa (Técnico, Administración, Comercial, I+D+i). La transformación *alr* de la composición de 4 partes produce vectores de 3 dimensiones en el espacio real, lo que permite considerar un modelo de regresión multivariante. El modelo en términos de la transformación *alr* puede expresarse de la siguiente forma:

$$
\begin{aligned}
log\left(\tfrac{\text{Administración}}{\text{I+D+i}}\right) &= \beta_{01} + \boldsymbol{\beta}_{11}\cdot X + \varepsilon_1 \\
log\left(\tfrac{\text{Técnico}}{\text{I+D+i}}\right) &= \beta_{02} + \boldsymbol{\beta}_{12}\cdot X + \varepsilon_2 \\
log\left(\tfrac{\text{Comercial}}{\text{I+D+i}}\right) &= \beta_{03} + \boldsymbol{\beta}_{13}\cdot X + \varepsilon_3
\end{aligned}
\tag{5.8}
$$

donde $\boldsymbol{\beta}_{1i}$ con $i = 1, 2, 3$ son los vectores de los coeficientes que acompañan en cada ecuación a las variables explicativas.

Los modelos lineales multivariantes se ajustan en R con la función `lm`, con los mismos argumentos que en el caso de modelos lineales univariantes pero indicando en la parte izquierda de la fórmula la matriz de respuestas. Esta función proporciona los coeficientes de la regresión estimados para cada respuesta. Para la verificación del modelo, la función `anova` permite manejar múltiples variables respuesta y comparar diferentes modelos.

Para seleccionar las variables explicativas, se parte de un modelo de regresión con todas las variables disponibles y se van eliminando sucesivamente aquellas variables que no son significativas en ninguna de las ecuaciones, empezando por las que ofrecen un p-valor promedio mayor. El modelo obtenido según este procedimiento es el que tiene como variables explicativas la variable servicio y el número de trabajadores de soporte. Las rectas estimadas para cada una de las tres ecuaciones son:

$$
\begin{aligned}
log\left(\frac{\text{Administración}}{\text{I+D+i}}\right) &= -2.95 + 4.76\cdot \mathbb{I}(\text{Ingeniería}) + 3.76\cdot \mathbb{I}(\text{Mixto}) - 0.56\cdot \text{Nº trab. soporte} \\
log\left(\frac{\text{Técnico}}{\text{I+D+i}}\right) &= -1.46 + 7.82\cdot \mathbb{I}(\text{Ingeniería}) + 5.26\cdot \mathbb{I}(\text{Mixto}) - 0.82\cdot \text{Nº trab. soporte} \\
log\left(\frac{\text{Comercial}}{\text{I+D+i}}\right) &= -2.64 + 3.89\cdot \mathbb{I}(\text{Ingeniería}) + 4.31\cdot \mathbb{I}(\text{Mixto}) - 1.90\cdot \text{Nº trab. soporte}
\end{aligned}
\tag{5.9}
$$

La variable servicio es de tipo factor y posee tres categorías: consultoría, ingeniería, y mixto. La categoría de referencia es consultoría. por lo que se crea un coeficiente para ingeniería y otro para mixto. En cada una de las ecuaciones en (5.9) el intercepto sería la predicción para un proyecto de consultoría y sin trabajadores de soporte.

Los coeficientes asociados a las diferentes categorías de la variable servicio tienen significación estadística en todas las ecuaciones de regresión. El coeficiente para la variable número de trabajadores de soporte sólo es significativo para la tercera ecuación. Los p-valores para el contraste de que cada coeficientes es nulo se muestran en la Tabla 5.5.

En la ecuación para $log\left(\frac{\text{Administración}}{\text{I+D+i}}\right)$ se observa que en los proyectos de ingeniería, en

	$log\left(\frac{\text{Administración}}{\text{I+D+i}}\right)$	$log\left(\frac{\text{Técnico}}{\text{I+D+i}}\right)$	$log\left(\frac{\text{Comercial}}{\text{I+D+i}}\right)$
Intercepto	0.002058	0.149	0.007843
Servicio - Ingeniería	$3.66 \cdot 10^{-6}$	$4.11 \cdot 10^{-10}$	0.000178
Servicio - Mixto	0.000272	$8.08 \cdot 10^{-6}$	$8.54 \cdot 10^{-5}$
Trabajadores soporte	0.413282	0.280	0.010538

Tabla 5.5: P-valores del contraste de significación de los coeficientes del modelo de regresión multivariante

relación a los de consultoría, aumenta el porcentaje de horas dedicadas al departamento de administración en relación a las horas dedicadas a I+D+i. Lo mismo ocurre con los proyectos mixtos en relación a los de consultoría: se produce un desplazamiento de las horas dedicadas a investigación en favor de horas administrativas.

Similares conclusiones se extraen de las ecuaciones para $log\left(\frac{\text{Comercial}}{\text{I+D+i}}\right)$ y para $log\left(\frac{\text{Técnico}}{\text{I+D+i}}\right)$. En esta última el aumento en horas técnicas en detrimento de las de I+D+i para los proyectos de ingeniería y mixtos es mayor que en los otros casos, lo que concuerda con lo visto en la Figura 5.4. Además, con el aumento del número de trabajadores de soporte, las horas en el departamento comercial ceden espacio a las horas en I+D+i.

Se puede comprobar en la Figura 5.4 que los proyectos de ingeniería son los que menos horas de I+D+i requieren, presentando porcentajes muy elevados de horas técnicas. Para completar y facilitar el análisis, se representan los boxplots para la variable servicio en la Figura 5.6 y el diagrama de dispersión para la variable número de trabajadores de soporte en la Figura 5.7 en función de las transformaciones log-cociente. Estas Figuras muestran esencialmente los efectos detectados por el modelo ajustado en (5.9).

Aplicando la inversa de la transformación alr vista en (4.16) obtenemos la versión en el símplex de (5.9):

$$\begin{aligned}&[\text{Administrativo Técnico Comercial I+D+i}] = [0.039 \quad 0.171 \quad 0.053 \quad 0.738] \oplus \\ &\oplus \mathbb{I}(\text{Ingeniería}) \otimes [0.044 \quad 0.937 \quad 0.018 \quad 0.3 \cdot 10^{-3}] \oplus \mathbb{I}(\text{Mixto}) \otimes [0.138 \quad 0.619 \quad 0.240 \quad 0.003] \oplus \\ &\oplus \text{Nº trab. soporte} \otimes [0.264 \quad 0.204 \quad 0.069 \quad 0.463]\end{aligned} \tag{5.10}$$

Esta expresión es de más difícil interpretación que los coeficientes en (5.9). Los coeficientes transformados informan del tamaño relativo del efecto de cada variable en cada departamento, en relación a los demás.

Para contrastar si el efecto de la variable número de trabajadores de soporte es significativo, se ajusta el modelo sin esta variable y se aplica la función anova entre este modelo y el modelo con las dos variables. El resultado confirma que el efecto de los trabajadores de soporte es significativo a un nivel de significación del 10 %, ya que el estadístico del contraste de Pillai-Bartlet es 0.1294,

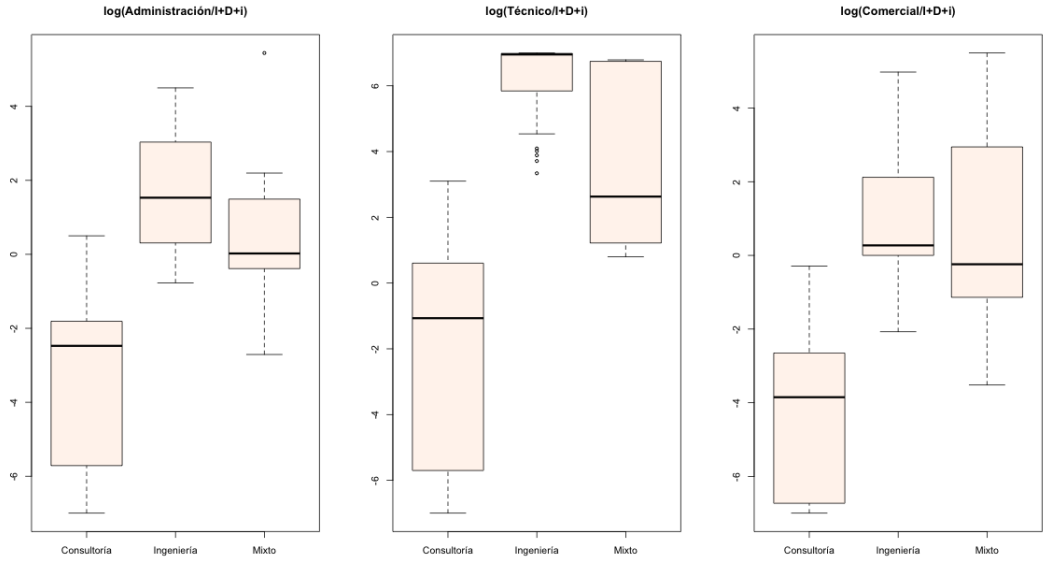

Figura 5.6: *Boxplots para las categorías de la variable servicio para las transformaciones alr*

y su p-valor asociado 0.0866.

Para valorar el comportamiento global del modelo disponemos de la medida R^2 vista en (4.17). Para el modelo de regresión composicional ajustado este valor es de 47.74, es decir, el modelo propuesto explica el 47.74 % de la variabilidad.

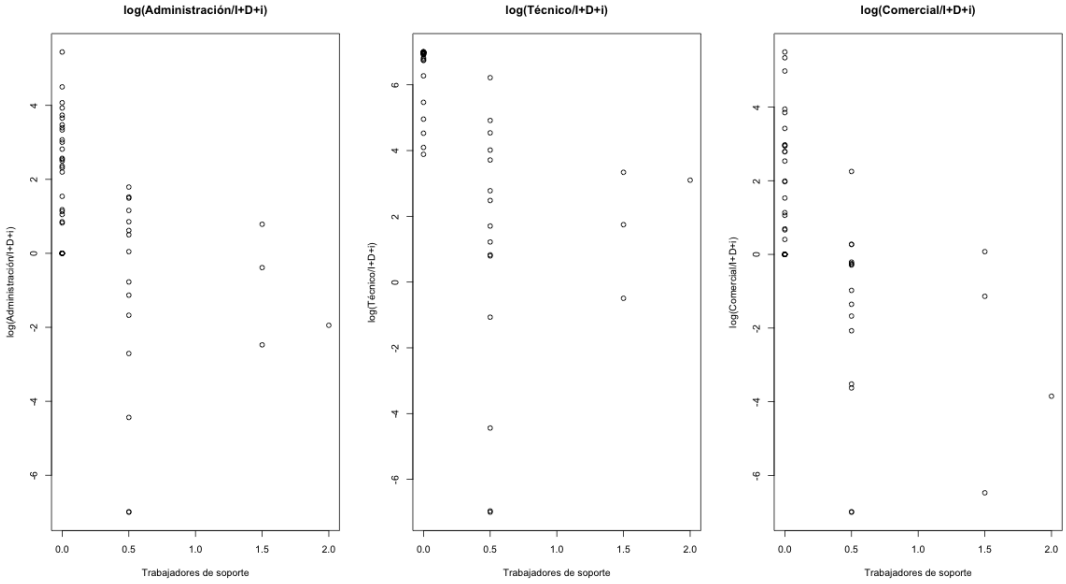

Figura 5.7: *Gráficos de dispersión para el número de trabajadores de soporte para las transformaciones alr*

Capítulo 6

Conclusiones

Al realizar un proceso de modelización y análisis de datos es necesario prestar especial atención a las características del espacio muestral sobre el que están definidos. Los datos que se refieren a partes de un total no se corresponden con valores continuos sobre el espacio real sino que se definen sobre un subconjunto de este último sujeto a las restricciones de no negatividad y de suma constante. El análisis con técnicas multivariantes estándar para datos reales no restringidos no es correcto en estos casos.

La metodología log-cociente permite trasladar el problema al espacio real y, por tanto, aplicar las técnicas estadísticas usuales. Sin embargo, la aplicación de esta estrategia no está exenta de problemas prácticos. En concreto, se ha analizado la problemática de la existencia de observaciones con partes nulas y se han visto algunas soluciones. Tras la realización de un estudio de simulación se concluye que la estrategia de reemplazo multiplicativa es adecuada en el contexto de estimación de modelos de regresión siempre y cuando se escoja un valor de sustitución para las componentes nulas apropiado.

Una vez analizado el problema de los ceros, se ha procedido a ajustar un modelo de regresión multivariante con la información disponible, la cual, en el contexto de las prácticas realizadas, consta de 53 proyectos de los que se conoce cómo se distribuyen las horas dedicadas a cada uno entre los diferentes departamentos de la empresa: administración, técnico, comercial e I+D+i. Se dispone también de una serie e variables que caracterizan a los proyectos.

Los vectores composicionales de 4 dimensiones han llevado a estimar 3 ecuaciones de regresión para la transformación *alr* de los datos. El modelo de regresión lineal multivariante ha seleccionado las variables servicio y número de trabajadores de soporte como variables explicativas. El modelo de regresión composicional ajustado explica un 47 % de la variabilidad de la respuesta.

Anexos

Apéndice A

Tabla de datos

	Departamentos					Departamentos			
Proy.	Administrativo	Técnico	Comercial	I+D+i	Proy.	Administrativo	Técnico	Comercial	I+D+i
1	0.011	0.000	0.026	0.963	28	0.000	1.000	0.000	0.000
2	0.000	0.000	0.000	1.000	29	0.038	0.944	0.018	0.000
3	0.000	0.012	0.000	0.988	30	0.015	0.938	0.047	0.000
4	0.109	0.200	0.109	0.582	31	0.068	0.868	0.033	0.031
5	0.006	0.951	0.000	0.043	32	0.012	0.984	0.004	0.000
6	0.000	0.778	0.222	0.000	33	0.146	0.742	0.050	0.062
7	0.002	0.998	0.000	0.000	34	0.000	0.810	0.190	0.000
8	0.000	0.972	0.028	0.000	35	0.074	0.761	0.156	0.008
9	0.020	0.980	0.000	0.000	36	0.185	0.618	0.084	0.112
10	0.029	0.838	0.132	0.000	37	0.046	0.931	0.013	0.010
11	0.018	0.982	0.000	0.000	38	0.052	0.910	0.022	0.016
12	0.000	1.000	0.000	0.000	39	0.000	1.000	0.000	0.000
13	0.082	0.918	0.000	0.000	40	0.042	0.888	0.053	0.018
14	0.054	0.929	0.018	0.000	41	0.211	0.774	0.015	0.000
15	0.012	0.968	0.018	0.002	42	0.019	0.966	0.013	0.002
16	0.003	0.994	0.003	0.000	43	0.019	0.960	0.014	0.007
17	0.020	0.669	0.009	0.302	44	0.000	1.000	0.000	0.000
18	0.000	0.985	0.015	0.000	45	0.042	0.918	0.018	0.022
19	0.063	0.666	0.074	0.197	46	0.027	0.973	0.000	0.000
20	0.000	1.000	0.000	0.000	47	0.049	0.913	0.023	0.015
21	0.000	0.957	0.043	0.000	48	0.012	0.988	0.000	0.000
22	0.047	0.953	0.000	0.000	49	0.026	0.974	0.000	0.000
23	0.009	0.991	0.000	0.000	50	0.088	0.741	0.041	0.129
24	0.050	0.360	0.000	0.590	51	0.019	0.969	0.008	0.004
25	0.204	0.738	0.012	0.046	52	0.011	0.982	0.007	0.000
26	0.206	0.449	0.150	0.196	53	0.035	0.953	0.011	0.000
27	0.003	0.989	0.000	0.007					

Tabla A.1: Proporción de horas dedicadas en cada proyecto a los diferentes departamentos

Apéndice B

Código en R

Código para el estudio de simulación

```
# Parámetros fijados
# ------------------------------ #
beta01<-0
beta11<-1
beta02<-0
beta12<-1
sigma<-rbind(c(1.5,1),c(1,1.5))
M<-10000 #número de simulaciones
ceroReplace<-0.005 #valor por el que se sustituyen los ceros en el método multilicativo

# Función para generar el vector composicional de la variable respuesta y
# ------------------------------ #
generateY<-function(x,n){
  e<-rmvnorm(n, rep(0, 2), sigma)
  error1<-e[,1]
  error2<-e[,2]
  y<-matrix(0,nrow=n,ncol=3)
  y[,3]<-1/(1+(exp(beta01+beta11*x)*exp(error1)+exp(beta02+beta12*x)*exp(error2)))
  y[,2]<-y[,3]*exp(beta02+beta12*x)*exp(error2)
  y[,1]<-y[,3]*exp(beta01+beta11*x)*exp(error1)
  return (y)
}

# Función para normalizar un vector (aplicando operador clausura)
# ------------------------------ #
normalizeY<-function(y){
  rsums<-rowSums(y)
  y<-y/rsums
  return(y)
}

# Función para convertir en 0 los valores de Y menores que delta y normalizarla
# ------------------------------ #
putZeros<-function(y,delta){
  for (i in 1:nrow(y))
```

```
    for (j in 1:ncol(y))
      if (y[i,j]<delta){y[i,j]=0}
  return(normalizeY(y))
}

# Función del método multiplicativo para eliminar ceros
# ------------------------------ #
multiMethod<-function(y){
  for (i in 1:nrow(y)){
    c=0
    for (j in 1:ncol(y)){
      if (y[i,j]==0){
        c=c+1
        y[i,j]=ceroReplace
      }
    }
    for (k in 1:ncol(y)){
      if (y[i,k]!=ceroReplace){
        y[i,k]=y[i,k]*(1-c*ceroReplace)
      }
    }
  }
  return(y)
}


# ESTIMACIÓN DE LOS MODELOS
# ------------------------------ #

simReplaceZeros<-function(n,delta){
  # n -> tamaño de la muestra
  # se convierten en ceros las cantidades menores a delta
  beta01Est<-matrix(0,ncol=M,nrow=4)
  beta11Est<-matrix(0,ncol=M,nrow=4)
  beta02Est<-matrix(0,ncol=M,nrow=4)
  beta12Est<-matrix(0,ncol=M,nrow=4)
  numZeros<-0

  for (i in 1:M){
    x<-runif(n)
    y<-generateY(x,n)

    # (1) Vector composicional original, sin ceros por redondeo
    y1<-getdata(DR_data(y,trafo=T)) # transformar los Y por el método 1
    mod1<-lm(alr(acomp(y1))~x)

    mod2<-lm(alr(acomp(y))~x) # modelo del método 2 (sin impacto cuando no hay ceros)

    # (2) Vector composicional con ceros reemplazados
    y3<-getdata(DR_data(putZeros(y,delta))) # modelo del método 1
    mod3<-lm(alr(acomp(y3))~x)

    y4<-multiMethod(putZeros(y,delta))
    mod4<-lm(alr(acomp(y4))~x) # modelo del método 2
```

```
    # recoger las estimaciones de cada uno de los modelos
    beta01Est[1,i]<-mod1$coefficients[1,1]
    beta11Est[1,i]<-mod1$coefficients[2,1]
    beta02Est[1,i]<-mod1$coefficients[1,2]
    beta12Est[1,i]<-mod1$coefficients[2,2]

    beta01Est[2,i]<-mod2$coefficients[1,1]
    beta11Est[2,i]<-mod2$coefficients[2,1]
    beta02Est[2,i]<-mod2$coefficients[1,2]
    beta12Est[2,i]<-mod2$coefficients[2,2]

    beta01Est[3,i]<-mod3$coefficients[1,1]
    beta11Est[3,i]<-mod3$coefficients[2,1]
    beta02Est[3,i]<-mod3$coefficients[1,2]
    beta12Est[3,i]<-mod3$coefficients[2,2]

    beta01Est[4,i]<-mod4$coefficients[1,1]
    beta11Est[4,i]<-mod4$coefficients[2,1]
    beta02Est[4,i]<-mod4$coefficients[1,2]
    beta12Est[4,i]<-mod4$coefficients[2,2]

    # número de ceros en cada y
    numZeros<-numZeros+sum(y<delta)
  }

  # tablas de resultados
  coefMean<-rbind(rowMeans(beta01Est), rowMeans(beta11Est), rowMeans(beta02Est),
  + rowMeans(beta12Est))
  colnames(coefMean)<-c('(X,Y) Met 1','(X,Y) Met 2','(X,Y*) Met 1','(X,Y*) Met 2')
  rownames(coefMean)<-c('Beta01','Beta11','Beta02','Beta12')

  coefSd<-rbind(apply(beta01Est,1,sd),apply(beta11Est,1,sd),apply(beta02Est,1,sd),
  + apply(beta12Est,1,sd))
  colnames(coefSd)<-c('(X,Y) Met 1','(X,Y) Met 2','(X,Y*) Met 1','(X,Y*) Met 2')
  rownames(coefSd)<-c('Beta01','Beta11','Beta02','Beta12')

  MSE<-(coefMean-matrix(c(0,0,0,0,1,1,1,1,0,0,0,0,1,1,1,1),byrow=T,ncol=4,
  +nrow=4))^2+coefSd^2

  percZeros<-(numZeros/M)/n # porcentaje de la media de ceros

  result<-list(coefMean=coefMean, coefSd=coefSd, MSE=MSE, percZeros=percZeros)
  return(result)
}
```

Bibliografía

[1] Aitchison, J. (1982). *The statistical analysis of compositional data*. Journal of the Royal Statistical Society, Series B (Statistical Methodology) 44, 139-177.

[2] Aitchison, J. (1986). *The statistical analysis of compositional data*. Monographs on Statistics and Applied Probability. Chapman & Hall Ltd., London.

[3] Aitchison, J., Kay, J. W. (2003). *Possible solutions of some essential zero problems in compositional data analysis*, en Thió-Henestrosa, S., Martín-Fernández, J. A. (eds.), First Compositional Data Analysis Workshop - CoDaWork'03. Universitat de Girona.

[4] Egozcue, J. J., Pawlowsky-Glahn, V., Mateu-Figueras, G., Barceló-Vidal, C. (2003). *Isometric logratio transformation for compositional data analysis*. Mathematical Geology, 35, pp. 279-300.

[5] Fox, J., Friendly, M., Weisberg, S. (2013) *Hypothesis tests for Multivariate Linear Models Using the car package*. The R Journal, vol. 5, no. 1, 2013.

[6] Martín-Fernández, J. A., Barceló-Vidal, C., Pawlowsky-Glahn, V. (2000). *Zero replacement in compositional data sets*, in Kiers, H., Rasson, J., Groenen, P., and Shader, M., eds. Proceedings of the 7th Conference of the International Federation of Classification Societies (IFCS'2000), University of Namur, Namur: Springer-Verlag, Berlin, p. 155-160.

[7] Martín-Fernández, J. A., Barceló-Vidal, C., Pawlowsky-Glahn, V. (2003). *Dealing with zeros and missing values in compositional data sets*, Mathematical Geology, 35, pp. 253-278.

[8] Palarea-Albaladejo, J., Martín-Fernández, J. A. (2008). *A modified EM alr-algorithm for replacing rounded zeros in compositional data sets*. Computer & Geosciences 34(8), 902-917.

[9] Palarea-Albaladejo, J. A., Martín-Fernández, J. A., Gómez-García, J. (2007). *A parametric approach for dealing with compositional rounded zeros*, Mathematical Geology 39(7), 625-645.

[10] 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 measurements of organs*, Proc. R. Soc., 60, pp. 489-498.

[11] Smithson, M. y Verkuilen, J. (2006). *A Better Lemon Squeezer? Maximum-Likelihood Regression With Beta-Distributed Dependent Variables*. Psychological Methods, 11(1), 54-71.