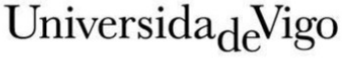


Regresión de Dirichlet en datos composicionales: distribución de tareas de consultoría

Rosa Caeiro Sánchez

Índice general

Resumen

Los datos composicionales son los que aparecen en una proporción, lo más frecuente son los que están en proporción 1 o 100, aunque pueden estar en cualquiera otra. Estos datos presentan un gran número de problemas al intentar buscar un modelo estadístico, ya que la restricción de que la suma de sus componentes sea igual a una constante está presente. Aunque estos datos son muy frecuentes en la práctica (ya que los podemos encontrar en la composición de un material de cualquier área o mismo en la proporción de horas dedicadas a alguna tarea), no existe mucha teoría sobre su modelización. En este trabajo consideramos los datos composicionales con sus respectivas restricciones. El enfoque será fundamentalmente el propuesto por Aitchison (1986). Trataremos la presencia de ceros, que resultan problemáticos para el ajuste de los modelos de Dirichlet (que utilizaremos). La motivación del trabajo viene de la práctica desarrollada en el seno de la empresa Azteca Consultig de Ingeniería S.L. En esta práctica, uno de los objetivos principales fue el planteamiento de un modelo explicativo de la distribución de horas entre distintos tipos de tareas (administrativas, comerciales...) en proyectos de consultoría de la citada empresa. Este trabajo finaliza proponiendo tal modelo, el cual permite identificar variables que tienen impacto significativo en la distribución de tareas.

Capítulo 1

Introducción

Este trabajo consiste en una memoria relativa a las prácticas realizadas en Azteca Consulting de Ingeniería S.L., una pequeña empresa de consultoría de Vigo. El objetivo de las prácticas (llevadas a cabo entre Junio y Agosto de 2012) fue diverso. Por una parte, se demandaba la construcción de una base de datos, donde se recogiera información relevante de la actividad de la empresa a partir de los partes de trabajo de cada trabajador. Por otra parte, se deseaba un análisis estadístico de los datos y la búsqueda de un modelo que permitiera relacionar las características de los proyectos realizados por la empresa con la distribución de horas entre los distintos tipos de tareas en cada proyecto.

La base de datos se creó en leguaje SQL, en el programa de sofware libre MySQL; para ello, en primer lugar, se siguió un aprendizaje sobre la creación y manipulación de este tipo de bases. Para crear la base con los datos de los que disponiamos tuvimos que crear distintas tablas y establecer relaciones entre ellas. En nuestro caso, se crearon tablas donde se introducirían los datos. Las tablas que se crearon fueron: las de trabajadores, departamentos, tareas primarias y secundarias (estas estaban relacionadas con las primarias y con el departamento que nos interesaba utilizar)... y una tabla donde estaban los datos que teniamos que introducir en la que aparecia la fecha, trabajador, horas, tarea secundaria. Puesto que la introducción de datos era muy compleja, debido a que este programa es con código, se decidió utilizar el programa Libre Base, que es un interfaz mucho más manejable visualmente. Pero como teniamos que mantener que la base estuviese en MySQL, se tuvieron que conectar estos dos programas. Y para finalizar, se tenían que poder exportar los datos desde MySQL a R (sofware utilizado para el análisis estadístico), para ello hubo que utilizar el paquete RODBC y crear un canal que conecte los dos programas (MySQL con R), despues de todo esto se podía empezar con el trabajo.

Por tanto, la práctica tuvo una dedicación importante en el apartado de gestión de

bases de datos y programación.

En el análisis estadístico, lo que se hizo fue en primer lugar, un estudio descriptivo de los datos, es decir, se realizaron los informes que resultaban más interesantes para la empresa, con los datos recogidos en la base de datos creada. Los informes recogían tanto la suma total de horas, como los porcentajes de ciertas combinaciones de los datos. Por ejemplo, uno de los informes era el que calculaba, según cada proyecto, cuántas horas se dedicaban a cada tarea y el porcentaje. Otro sería el que, a partir de los datos sobre proyectos y trabajadores, se calculaban las horas totales y porcentajes de cada tarea. Otro informe tuvo que ver con las tareas a las que se dedicaba cada trabajador, para poder así saber cúal era el perfil del trabajador. Así con diferentes casos. Estos informes, para que fueran más visibles y no hubiera que ejecutar de cada vez que se quisieran, se presentaron ya en la misma base datos.

En esta memoria, nos vamos a centrar en el segundo aspecto de las prácticas, que es la construcción de un modelo de regresión que nos permitiera, en un futuro, estimar la distribución de las horas del proyecto según la tarea; ésto permitiría a la empresa saber cuáles son los trabajadores que tendrían que desempeñar dicho proyecto, y con los informes ya tendrían un perfil de cada trabajador y así poder optimizar el rendimiento del proyecto.

Nosotros partimos de los datos que se recogen en la base de datos, estos son la fecha, el proyecto,el trabajador, las tareas primarias y secundarias y el tiempo dedicado a cada una de ellas.

Por otra parte, lo que tenemos que obtener es la distribución de las tareas sabiendo algunas características del proyecto. Para ello, se calcularon los porcentajes de las tareas dedicadas a cada proyecto. Por tanto, en este momento nos encontramos con que estamos con datos composicionales; por lo que no se pueden aplicar los modelos de regresión más habituales, sino que hay que buscar otros para el estudio de este tipo de datos. En primer lugar nos vamos a centrar en los datos composicionales, para poder así entender con más claridad el modelo de regresión que se utilizará en el análisis.

Antes de centrarnos en el modelo de Dirichlet, tenemos que centrarnos en los problemas con los ceros, ya que éste es uno de los problemas que tienen este tipo de modelos, pero no solo está presente en dichos modelos, tambien se pueden encontrar este tipo de problemas con datos composicionales.

Después de tener claro qué son los datos composicionales y los problemas con los ceros pasamos al modelo que hemos elegido para el estudio, es decir, nos centramos en los modelos de regresión de Dirichlet, aplicando posteriormente estos modelos a los datos que tenemos; para ello utilizaremos el sofware R con los paquetes compositions y DirichletReg.

Capítulo 2

Datos composicionales

El problema de los datos composicionales ha sido y es una fuente de preocupación para muchos científicos desde que en 1897 Karl Pearson pusiera de manifiesto la inadecuación de los métodos estadísticos clásicos para el estudio de los mismos. Los datos composicionales son realizaciones de vectores aleatorios de sumas constantes, es decir, cualquier vector x, cuyas componentes representan partes de un todo, está sujeto a la restricción de que la suma de sus componentes sea la unidad, o en el caso general, una constante.

Definición 2.1. Un dato composicional $x = (x_1, x_2, \ldots, x_D)$ es un vector con componentes estrictamente positivas, tal que la suma de todas ellas es igual a una constante k. Su espacio muestral es el símplex S^D, definido por

$$S^D = \left\{ (x_1, x_2, \ldots, x_D) / x_i > 0; \sum_{i=1}^{D} x_i = k \right\}.$$

En su formulación estadística, los datos composicionales son realizaciones de una composición, vector aleatorio cuyo recorrido está en S^D. En estos casos se plantea la necesidad de aplicar técnicas estadísticas para el estudio e interpretación de este tipo de datos.

Éstos aparecen en áreas muy diversas. Así, por ejemplo, en Geología al expresar la composición geoquímica de una roca. También es frecuente encontrar datos composicionales de naturaleza granulométrica provenientes de sedimentos marinos. En Economía encontramos datos de tipo composicional al estudiar, por ejemplo, la distribución de un presupuesto entre las diferentes partidas y por supuesto, tambien aparecen en nuestro caso, teniendo los porcenjes de horas dedicadas a cada departamento según el proyecto. También en Arqueometria aparecen datos composicionales al estudiar, por ejemplo, la

composición geoquímica de las cerámicas procedentes de excavaciones arqueológicas con el objetivo de determinar su origen.

Antes de indicar la problemática específica que comporta el análisis estadístico de dicho datos introduciremos dos definiciones de gran importancia: el operador clausura y la noción de subcomposición.

A partir de un vector con componentes positivas siempre podemos obtener un dato composicional S^D. Basta con dividir cada una de sus componentes por la suma de todas ellas. Este hecho conduce a dar las siguientes definiciones.

Definición 2.2. El operador clausura C es una transformación que hace corresponder a cada vector $w = (w_1, w_2, \ldots, w_D)$ de $\mathbb{R}_+^D$ su dato composicional asociado $C(w) = kw/(w_1 + w_2 + \ldots + w_D)$ de S^D, con k la constante de clausura.

En alguno de los casos puede interesarnos analizar únicamente el valor de las magnitudes relativas de un subconjunto de partes de unos datos composicionales. Es pues necesario disponer de un procedimiento de formación de subcomposiciones.

Definición 2.3. Si S es un subconjunto cualquiera de las partes $1, 2, \ldots, D$ de un dato composicional $x \in S^D$ y x_S simboliza el subvector formado por las correspondientes partes de x, entonces $s = C(x_s)$ recibe el nombre de subcomposición de las S partes de x.

2.1. Principales problemas:

La restricción de la suma constante ha sido considerada la fuente de todos los problemas pues impide la aplicación de los procedimientos estadísticos habituales que se utilizan para datos que no presentan esta restricción. Nótese,por ejemplo, que el cambio en una de las partes provoca el cambio en como, mínimo, otra de las demás.

Una de las dificultades más relevantes es la imposibilidad de interpretar correctamente las covarianzas y los coeficientes de correlación. En 1897 Pearson ya puso de manifiesto esta dificultad, donde advertía de la existencia de una falsa correlación entre las partes de una composición. La matriz de correlaciones habitual no puede analizarse en el estudio de vectores de suma constante porque presenta necesariamente correlaciones negativas no nulas, determinadas precisamente por la mencionada restricción. Pearson clasificó estas correlaciones como espureas ya que falsean la imagen de las relaciones de dependencia y pueden conducir a interpretaciones erróneas. En particular, si analizamos la matriz de covarianzas usual entre las partes de la composición, $K = \{cov(x_i, x_j)/i, j = 1, , 2, \ldots D\}$, obtenemos que

$$cov(x_i, x_1) + cov(x_i, x_2) + \cdots + cov(x_i, x_D) = 0 \qquad i = 1, 2, \ldots, D$$

a causa de la restricción $\sum_{i=1}^{D} x_i = k$. Sabemos que $cov(x_i, x_i) = var(x_i) > 0$,excepto en la situación trivial que la parte x_i sea una constante. Este hecho provoca que necesariamente deba haber una covarianza $cov(x_i, x_j) < 0$ $(i \neq j)$. Veamos que estas covarianzas no son libres de tomar cualquier valor. Esto invalida la interpretación habitual de las covarianzas, y en consecuencia de las correlaciones, puesto que a priori suponemos que deberían poder adquirir libremente valores nulos, positivos o negativos. Por el mismo motivo, el hecho que el coeficiente de correlación entre dos partes cualesquiera de una composición sea igual a 0 no puede interpretarse, como es habitual, como indicio de independencia entre ambas partes.

Encontramos otra incoherencia en relación a las subcomposiciones. Intuitivamente esperaríamos encontrar una cierta relación entre la matriz de covarianzas de una subcomposición y la de la composición de procedencia. Sin embargo, no existe ninguna relación. Es posible, incluso, que dos partes estén correlacionadas positivamente en el seno de una composición y en cambio pasen a tener correlación negativa al analizarlas como partes integrantes de una subcomposición.

En general tampoco es correcto aplicar las operaciones clásicas del espacio real vectorial a los datos composicionales. Esto tiene consecuencias estadísticas importantes porque existen multitud de conceptos y técnicas estadísticas que se fundamentan de forma más o menos explícita en la distancia euclideana.

Otra de las dificultades importantes es la falta de familias paramétricas suficientemente flexibles para modelar los conjuntos de datos composicionales. Las distribuciones de Dirichlet y sus generalizaciones se obtienen mediante la clausura de vectores aleatorios con componentes independientes. Como consecuencia, sus partes son prácticamente independientes, puesto que su correlación está únicamente motivada por el hecho de haber dividido todas sus componentes por la suma de éstas. Esto impide su uso en la modelización de fenómenos con relaciones de dependencia no inducidas por la suma constante.

Todas estas dificultades ponen de relieve la necesidad de replantear el análisis estadístico de los datos composicionales.

2.2. Principios fundamentales:

Muchos han sido los autores que han intentado afrontar los problemas del análisis estádistico de los datos composicionales. La solución no aparece hasta 1982 cuando Aitchison presenta, por primera vez, una forma de evitar la restricción de la suma constante.

Aitchison argumenta que las dificultades de la interpretación vienen motivadas por centrar la atención en las magnitudes absolutas de las partes $x_1, x_2, \ldots, x_D$ de una composición. La atención debe centrarse en la magnitud relativa de las partes, es decir, en los cocientes x_i/x_j $(i, j = 1, 2, \ldots, D; i \neq j)$. Por lo tanto, diremos que el problema es composicional cuando reconozcamos que el valor en términos absolutos de las partes es irrelevante. Esto es un principio fundamental del análisis de datos composicionales que Aitchison denomina *invarianza por cambios de escala.* Una importante consecuencia que se deduce de este principio es que: "cualquier función aplicada sobre los datos composicionales debe poder expresarse en términos de cocientes entre todas sus parte".

Trabajando con los cocientes desaparecen los problemas de las correlaciones espureas. Por otra parte, la magnitud relativa entre las partes de la composición original, es decir, x_i/x_j.

La metodología de Aitchison se basa en la transformación de los datos composicionales al espacio real multivariante. Notemos que el espacio muestral para los cocientes entre las partes es $\mathbb{R}^{D-1}$ y por tanto podemos aplicar cualquier técnica estadística clásica.

Tenemos diversas posibilidades de transformación de los datos, todas ellas basadas en los logaritmos de cocientes entre las partes de un dato composicional.

Definición 2.4. La transformación log-cociente aditiva, alr, de $x \in S^D$ a $y \in \mathbb{R}^{D-1}$ se define como $y = alr(x) = (ln(x_2/x_D), ln(x_3/x_D), \ldots, ln(x_{D-1}/x_D))$

Esta transformación es biyectiva pero no es simétrica en las partes de x ya que la parte del denominador adquiere un protagonismo especial respecto al resto. Este hecho condujo a Aitchison a introducir la trasformación log-cociente centrada

Definición 2.5. La transformación log-cociente centrada (clr) de $x \in S^D$ a $z \in \mathbb{R}^D$ se define como $z = clr(x) = (ln\,(x_1/g\,(x))\,, ln\,(x_2/g\,(x))\,, \ldots, ln\,(x_D/g\,(x)))$, donde $g(x)$ es la media geométrica de las D partes de x.

Esta trasformación es biyectiva, y simétrica entre las partes. Su imagen es el hiperplano de $\mathbb{R}^D$ que pasa por el origen y es ortogonal al vector de unidad, es decir, la suma de las componentes del vector transformado es igual a cero. Nos encontramos ante una nueva dificultad ya que la matriz de covarianzas del vector trasformado será singular.

Por esta razón Aitchison aplica una estrategia doble en sus trabajos. En las aplicaciones que exigen simetría en el tratamiento de sus componentes utiliza la transformación clr. Para la modelización de conjuntos de datos composicionales con distribuciones multivariantes, utiliza la transformación alr. De esta forma evitamos trabajar con distribuciones degeneradas.

2.3. Interpretación y visualización de los parámetros

Interpretación de μ como una composición

El parámetro de localización de la distribución de la normal logística, μ, puede ser expresado como una composición a través de la transformación logística aditiva. Esto es

$$alr^{-1} = \xi \text{ donde } \xi \in S^D$$

Interpretar ξ es más simple que μ en la escala logit multivariante. Sin embargo, algunas de las propiedades estadísticas de la media se pierden con la transformación del simplex. Especialemente, μ es la media y la moda de una distribución normal logística. La transformación $alr^{-1}(\cdot)$ no conserva dicha propiedad. Sin embargo, dicha transformación es monótonona en cada D componentes de μ. Como consecuencia, los valores ordenados se mantienen bajo esta transformación. Entonces, $\xi = alr^{-1}(\mu)$ puede ser interpretada como la media multivariante aconsejable. Como se muestra acontinuación, esta interpretación es usualmente caracterizada por puntos estimados como parámetros, y como el centro de la asimetría de la distribución normal logística.

Covariables

Para incorporar el efecto de las covariables, los parámetros de localización, μ, puede depender de variables explicativas. Para una variable escalar x_i, indicada por $j = 1, \ldots, D$ observaciones, puede ser reemplazado en la expresión de la densidad por $\beta_0 + \beta_1(x_j - \bar{x})$. Entonces, β_0 y β_1 son vectores en $\mathbb{R}^D$, y $\bar{x}$ es la media de los valores de las covariables. Esta parametrización permite interpretar el β_0 como la localización, y el β_1 como el cambio en la localización por la unidad en cambios en x.

Equivalentemente, la expresión de regresión $\mu_j = \beta_0 + \beta_1(x_j - \bar{x})$ puede ser escrita como la perturbación de las composiciones. Esto se puede lograr tomando la transformación logística aditiva en ambos casos,

$$alr^{-1}(\mu_j) = alr^{-1}(\beta^0) \oplus alr^{-1}(\beta_1)^{(x_j - \bar{x})} \Leftrightarrow \xi_j = \xi \oplus \gamma^{\mu_j}$$

El escalar μ_j es el centro de la covarianza. En este caso el ξ es la localización en el simplex. Además, el papel de la regresión de los parámetros de la composición, γ es claro: el parámetro de localización es la totalidad de la localización (ξ) perturbado por γ. Este γ puede ser interpretado direcatamente como una composición. Es la cantidad por la que la localización es desplazada por una unidad de cambio en la covariable, debido a una perturbación

Capítulo 3

Problemas con los ceros

Los ceros en datos composicionales están clasificadas en ceros redondeados y ceros esenciales. Los ceros redondeados corresponden a pequeñas proporciones o por debajo del límite de detección, mientras que los ceros esenciales son una indicación de la ausencia completa del componente en la composición. En análisis de datos composicionales, la presencia de componenetes ceros respresenta uno de los mayores obstáculos frente a la aplicación del análisis del "logratio" y la regresión de Dirichlet. En el análisis del "logratio", no podemos coger el logaritmo de ceros cuando aplicamos la transformación aditiva. En el modelo de Dirichlet, la presencia de ceros hace que la función de desidad desaparezca.

3.1. Reemplazamiento de los ceros en datos composicionales

Aitchison (1986) clasifica los ceros en los datos composicionales dentro de los ceros redondeados y los ceros esenciales. Los ceros redondeados son una medida del proceso artificial, donde las observaciones son consideradadas cero cuando están por debajo del límite detección. Tales ceros serán valores que tratarán como valores erróneos y serán reemplazados. Por otra parte, a veces las observaciones recuerdan que los ceros son la indicación de la ausencia de dicho componente en la composición. Estos ceros son los llamados ceros esenciales o ceros verdaderos y su patrón de incidencia debe ser investigada y modelada.

3.1.1. Métodos comunes de reemplazamiento de los ceros

Como solución del redondeo de los ceros, Aitchison (1986) sugiere la reducción del número de componentes en la composición por fusión. Esto es, eliminar las componentes con observaciones cero por combinaciones con alguna de las otras componentes. Esta aproximación no es apropiada cuando el objetivo es modelar la composición original o los modelos que incluyan solo tres componentes. A veces, es más lógico aproximar este cero por un valor muy pequeño que no distorsione seriamente la estructura de la covarianza de los datos. El primer método para esta sustitución lo propusoo Aitchison (1986), el método del reemplazamiento aditivo, es simplemente reemplazar los ceros por un valor muy pequeño de δ y luego normalizar los imputaciones de la composición. Se mostró que los reemplazamientos aditivos no son subcomposciones coherentes y consecuentemente, distorsionan la estructura de la covarianza de los datos. Palarea-Albaladejo et al.(2007-2008) propusieron un método alternativo usando los reemplazamientos multiplicativos que conserva los valores distintos de cero. Sea $x = (x_1, \ldots, x_D) \in S^D$ una composición con ceros redondeados. El método de reemplazamientos multiplicativo de la composición x contienen c ceros con una composición cero libre $r \in S^D$ con la siguiente regla de reemplazamiento,

$$r_j = \begin{cases} \delta_j & \text{si } x_j = 0 \\ (1 - c\delta)x_j & \text{si } x_j > 0 \end{cases}$$

En la adición, Palarea-Albaladejo et al.(2007-2008) enfatizó que el mejor resultado era el obtenido cuando δ está cerca del 65 % del límite de detección. Sin embargo, cuando los reemplazos multiplicativos imputan exactamente el mismo valor en todos los ceros de la composición, estes reemplazos indroducen correlaciones artificiales entre componentes que tienen valores ceros en alguna de sus composiciones.

Además de esta aproximación no paramétrica, una aproximación paramétrica es la basada en la aplicación de una modificanción del algoritmo EM en las transformaciones "logratio" propuestas. Sin embargo, ninguno de estos métodos puede ser aplicado cuando los datos composicionales surgen de una distribución de Dirichlet.

Hijazi propuso un nuevo método basado en en la beta regresión para reemplazar los ceros cuando los datos surgen de una distribución de Dirichlet. Sin embargo, el método propuesto raramente podría reemplazar los ceros redondeados por valores que sobrepasasen los límites de detección.

3.2. Método basado en algoritmos EM

Los algoritmos EM (Expectation-Maximization) son algoritmos iterativos extensamente utilizados para estimación paramétrica por métodos de máxima verosimilitud cuando parte de dichos datos son datos erróneos o censudados. Este algoritmo consiste en dos partes. El paso E encuentra las extración condicional de los datos errróneos dadas por las observaciones y por los parámetros estimados, y posteriormente cambia los datos erróneos por los valores estimados. El paso M realiza la estimación de máxima verosimilitud de los parámetros usados en la estimación como si fueran valores correctos.

En el contexto de los datos composicionales, sea $X = [x_{ij}]$ la muestra aleatoria de n observaciones con D componentes de la composición de tal forma que $X = (X_{obs}, X_{error})$ donde X_{obs} y X_{error} son las observaciones y la parte de errores respectivamente. Supongamos que X sigue una distribución de Dirichlet con parámetros $\Lambda = (\lambda_1, \ldots, \lambda_D)$ y los ceros redondeados se producen cuando $x_{ij} < \gamma$ donde γ es el límite de detección. El logaritmo de máxima verosimilitud de los datos completos para Λ basada en la muestra X viene dada por

$$l = \log L = n \left\{ \log \Gamma(\lambda) - \sum_{j=1}^{D} \log \Gamma(\lambda_j) \right\} + \sum_{i=1}^{n} \sum_{j=1}^{D} (\lambda_j - 1) \log(x_{ij})$$

donde $\lambda = \sum_j^D \lambda_j$. Una reparametrización habitual de la desidad de Dirichlet es la construida por $\mu_j = \lambda_j / \lambda$ para $j = 1, 2, \ldots, D-1$ mientras que λ será el parámetro de dispersión del modelo.

Ya que la distribución de Dirichlet pertenece a la familia exponencial, los datos completos de log-verosimilitud es lineal en las no observaciones. Por tanto, en la t-ésima iteración del modelo del algoritmo EM, la computación de la esperanza condicional de los datos en el paso E reducidos por la computación del valor condicionado de la parte errónea, $E[X_{error}|X_{obs}; \theta^{(t)}]$, dada por la estimación de los parámetros $\theta^{(t)} = \Lambda^{(t)}$.

El paso E en los algoritmos EM debe ser modificada para incorporar el límite de detección. Por tanto, en la t-ésima iteración en el paso E, los valores de x_{ij} son reemplazados de acuerdo con la siguiente regla:

$$x_{ij}^{(t)} = \begin{cases} x_{ij} & \text{si } x_{ij} \geq \gamma \\ E(x_{ij}|x_{i,-j}, x_{ij} < \gamma, \Lambda^{(t)}) & \text{si } x_{ij} < \gamma \end{cases}$$

para $i = 1, \ldots, n$ y $j = 1, \ldots, D$, donde $x_{i,-j}$ denota el conjunto de observaciones de

la i-ésima composición de los datos. La esperanza condicionada en el método propuesto debe ser computada con la distribución de Dirichlet. La esperanza se escribe como

$$E(x_j|x_{-j}, x_j < \gamma) = \frac{1}{P(x_j < \gamma)} \int_0^{\gamma} \frac{\Gamma(\lambda)}{\Gamma(\lambda_j)\Gamma(\lambda - \lambda_j)} x_j^{\lambda_j - 1}(1 - x_j)^{\lambda - \lambda_j - 1} dx_j$$

donde λ y λ_j están obtenidos por la regresión Beta de x_j en x_{-j}. La expresión anterior se reduce a

$$E(x_j|x_{-j}, x_j < \gamma) = \frac{\lambda_j F_1(\gamma)}{\lambda F_2(\gamma)}$$

donde F_1 y F_2 son las distribuciones acumuladas de las variables beta con parámetros $(\lambda_j + 1, \lambda - \lambda_j)$ y $(\lambda_j, \lambda - \lambda_j)$

Capítulo 4

Modelos de regresión de Dirichlet

El modelo de Dirichlet con parámetros constates es un modelo que puede adaptarse a algunas formas de datos composicionales. Veremos que el modelo de Dirichlet puede incluir regresores o variables explicativas, dando lugar a una clase más amplia de aplicaciones.

Este modelo es bastante flexible para explicar las tendencias y estructuras de covarianzas directamente en los datos composicionales, y que no requieren la independencia de los objetos de Aitchison. La estructura de la covarianza de la distribución es necesariamente negativa. En la función de densidad de Dirichlet, los datos deben introducirse solamente en proporciones. Esto significa que, dados dos vectores, x y $x* = cx$, ambos son tratados como $y = C(x) = C(x*)$ donde C es el operador clausura. Por lo tanto, los modelos de Dirichlet tienen propiedades de invarianza de escala.

4.1. Modelos de Dirichlet:

Dado $X = (x_1, \ldots, x_D)$ un vector positivo (de proporciones) de dimensión D y función de densidad $f(x)$ cuyos parámetros son $(\lambda_1, \ldots, \lambda_D)$, $\lambda_i > 0$ para cualquier i, se dice que siguen un modelo de Dirichlet si

$$f(x) = \frac{\Gamma(\lambda)}{\prod_{i=1}^{D} \Gamma(\lambda_i)} \prod_{i=1}^{D} x_i^{\lambda_{i-1}}$$

donde

$$\sum_{i=1}^{D} x_i = 1 \ , \ \lambda = \sum_{i=1}^{D} \lambda_i \text{ y } \Gamma\text{es la función gamma.}$$

Los modelos de regresión de Dirichlet son fáciles de estimar si suponemos que la distribución de Dirichlet no cambia con la covariable. Para introducir posibles efectos de una covariable s, los parámetros de la distribución de Dirichlet $\lambda_1, \ldots, \lambda_D$ se pueden escribir como una función $\lambda_j(s)$ de la covariable s que toma valores positivos. En la práctica, además de la familia de polinomios (con restricción de positividad), se ha sugerido la familia exponencial ($\lambda_j(s_i) = e^{\beta s_i}$) como una función natural en este contexto (Gueorguieva et al., 2008).

Observación 4.1. En el caso donde $D = 2$, estamos en una distribución beta con parámetros (λ_1, λ_2). Por lo que el modelo de Dirichlet es una generalización de la distribución beta.

Algunos parámetros que podemos calcular en el modelo de Dirichlet son:

- Media

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

- Varianza

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

- Covarianza

$$Cov(x_i, x_j) = \frac{-\lambda_i \lambda_j}{\lambda^2(\lambda + 1)} \text{ para todo } i \neq j$$

donde $\lambda = \sum_{i=1}^{D} \lambda_i$

Observación 4.2. La covarianza siempre es negativa, ya que los valores de λ_i y λ son siempre positivos.

En el caso particular de que todos los parámetros son iguales, es decir, $\lambda_1 = \ldots = \lambda_D = \alpha$, la función de densidad sería,

$$f(x) = \frac{\Gamma(D\alpha)}{\Gamma(\alpha)^D} \prod_{i=1}^{D} x_i^{\alpha-1}$$

En este caso la media y la varianza coinciden para cualquier x_i, y la covarianza es la misma para cualquier x_i, x_j con $i \neq j$.

El concepto tradicional de correlación busca una relación lineal entre las componentes (variables). Si existe una fuerte relación lineal, entonces puede haber información sobrante que empeora la calidad de la información.

En teoría de la información, la cantidad de información se mide por la incertidumbre de la distribución. La distribución gaussiana con varianzas grandes tiene mayor incertidumbre. Para la distribución de Dirichlet, un pequeño α indica mayor incertidumbre, ya que

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

y α pequeño indica varianza más grandes.

Por otra parte la situación con α pequeña indican menos covarianza,

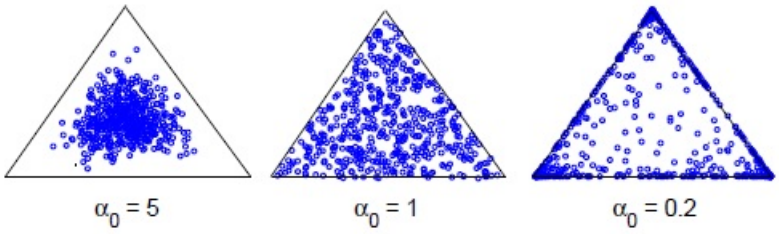
$$Cov(x_i, x_j) = \frac{-\lambda_i\lambda_j}{\lambda^2(\lambda + 1)} = \frac{-\alpha^2}{D^2\alpha^2(D\alpha + 1)} = \frac{-1}{D^2(D\alpha + 1)}$$

Y la correlación no depende de los valores de α, ya que

$$Cor(x_i, x_j) = \frac{Cov(x_i, x_j)}{\sqrt{Var(x_i)}\sqrt{Var(x_j)}} = \frac{-1}{D - 1}$$

Wang et al. (2008) definen la correlación para datos composicionales como el valor (estimado) de α.

La interpretación intuitiva de la correlación de Dirichlet en este caso particular es (misma referencia):



- Si $\alpha > 1$, la distribución tiene forma de protuberancias, donde los puntos se mezclan entre ellos.
- Si $\alpha = 1$, la distribución es uniforme, y cualquier proporción de la mixtura es igualmente preferible.
- Si $\alpha < 1$, la distribución tiene forma de valle formando picos en los vértices del simplex, y los datos se van alejando para los vértices de los triángulos.

Aunque esto es solo para el caso particular de que los parámetros son iguales. Volvamos al caso general donde los paramétros no tienen porque ser iguales.

Observación 4.3. Si un parámetro λ_i baja de valor, entonces al menos otro valor de λ_j tiene que aumentar debido a que tienen que mantener que la suma tiene que ser constante. Por lo que existe una correlación entre los parámetros.

4.2. Estimación de los parámetros en la regresión de Dirichlet

Sea $X_i = (x_{i1}, \ldots, x_{iD})$ una observación de un vector de proporciones con $x_{ij} > 0$ y $\sum_{j=1}^{D} x_{ij} = 1$ y sea s_i la correscondiente observación de la covariable para $i = 1, \ldots, n$. Se asume que las distribuciones condicionales de X_i dada por s_i son mutuamente independientes con $X_i|s_i$ función de distribución de Dirichlet con parámetros conocidos $\lambda(s_i) = (\lambda_1(s_i), \ldots, \lambda_D(s_i))$. La función de probabilidad dadas las covariables es

$$L = \prod_{i=1}^{n} \left\{ \Gamma\left(\lambda(s_i)\right) \prod_{j=1}^{D} \frac{x_{ij}^{\lambda_j(s_i)-1}}{\Gamma(\lambda_j(s_i))} \right\}$$

donde $\lambda(s_i) = \sum_{j=1}^{D} \lambda_j(s_i)$

Las propiedades de los estimadores de máxima verosimilitud de los parámetros del modelo de regresión de Dirichlet (para distintas posibles parametrizaciones de $\lambda_j(s_i)$) fueron discutidas por Hijazi y Jernigan (2007) y Murteira et al. (2012), entre otros. En particular, se ha establecido la convergencia en distribución de los parámetros estimados a una normal para parametrizaciones genéricas de la forma $\lambda_j(s_i) = G_j(s_i;\beta)$ (e.g. Murteira et al., 2012).

4.2.1. Medidas de diagnóstico

Después de la estimación del modelo de Dirichlet y la investigación de la distribución de los estimaciones de máxima verosimilitud, nos centraremos en el ajuste del modelo a los datos composicionales. Los diferentes modelos pueden proporcionar ajustes adecuados y dar resultados razonables pero el mejor modelo depende de los criterios elegidos. La generalización de la razón de verosimilitud se puede utilizar para elegir entre los modelos constantes y los que incluyen covariables. Proporcionaremos métodos de ajuste para evaluar el modelo de Dirichlet ajustado.

Medida R^2

Análogamente a la regresión lineal, podemos elaborar una medida numérica para evaluar el modelo construyéndolo como la medida R^2 habitual. Esta medida puede utilizarse para la estimación de los porcentajes de variación explicada por el modelo. Diferentes medidas de R^2 se han propuesto para evaluar modelos no lineales tales como el logístico y el modelo de probabilidad. Estas pseudo medidas de R^2 miden la mejora proporcional en la función de verosimilitud debido a la variable explicativa en el modelo, en comparación con el modelo constante.

Cuando se introduce el análisis "logratio", Aitchison (1986) propone una medida de la variabilidad total basada en la matriz de variación de los datos "logratio" transformados, $T(x)$ se define como

$$T(x) = [\tau_{ij}] = \left[var\left\{ log\left(\frac{x_i}{x_j}\right)\right\}\right]$$

Obviamente, $T(x)$ es simétrica con ceros en los elementos de la diagonal. La medida de variabilidad total se define como

$$totvar = \frac{1}{d}\sum_{i<j}\left[var\left\{log\left(\frac{x_i}{x_j}\right)\right\}\right] = \frac{1}{2d}\sum T(x)$$

Se comparan la variabilidad total de las observaciones y los datos ajustados para obtener la medida de R^2, es decir,

$$R^2 = \frac{totvar(\hat{x})}{totvar(x)}$$

donde x es la observación de los datos y $\hat{x}$ son los datos ajustados.

Suma de errores composicionales

Otra medida propuesta de variabilidad explicada es la basada en la distancia composicional. La distancia de Aitchison se define como la media de distancia entre dos composiciones. Esta distancia esta dada por

$$\Delta_s(X,x) = \left[\sum_{i=1}^{D}\left\{log\frac{x_i}{g(x)} - log\frac{X_i}{g(X)}\right\}^2\right]^{\frac{1}{2}}$$

donde $g(y)$ es la media geométrica de la composición y. Definen una métrica en el simplex y que verifique todas las propiedades necesarias en el análisis de datos composicionales, tales como invarianza en perturbaciones, invarianza de escala y invarianza bajo permutaciones de los componentes de la composición. Sea $x_i = x_{i1}, \ldots, x_{iD}$ la observación de los datos y $\hat{x_i} = \hat{x_{i1}}, \ldots, \hat{x_{iD}}$ los datos ajustados para $i = 1, \ldots, n$. Entonces, $\Delta_S(x_i, \hat{x_i})$ puede servir como el i-ésimo residuo y por consiguiente, análogo a la suma de cuadrados de los errores en la regresión lineal, la suma de los errores composicionales

$$SCE = \sum_{i=1}^{n}\Delta_S(x_i, \hat{x_i})$$

sirve como una medida de variabilidad no explicada en los modelos de datos composicionales.

Capítulo 5

Modelo para los datos reales

En esta Sección analizamos los datos reales proporcionados por la empresa Azteca Consulting de Ingeniería S.L., relativas a proyectos desarrollados por la misma. Estos datos se corresponden al porcentaje de horas (del total de cada proyecto) dedicadas a tareas técnicas, administrativas, de I+D+i y comerciales, por parte del personal de la empresa; también conocemos unas determinadas características de dichos proyectos. Disponemos datos relativos a 38 proyectos, guardados en el `data.frame` de nombre `tabla1`. Los datos de la variable principal (proporciones, datos composicionales) son los siguientes (primeras cuatro variables en `tabla1`):

	Administrativo	Técnico	Comercial	I+D+i
2	0.011904762	0.98809524	0.000000000	0.000000000
4	0.002961500	0.98124383	0.015794669	0.000000000
5	0.000000000	0.51612903	0.000000000	0.483870968
7	0.000000000	1.00000000	0.000000000	0.000000000
10	0.213355049	0.73615635	0.013029316	0.037459283
12	0.016757117	0.96648577	0.011806151	0.004950966
13	0.000000000	0.77777778	0.222222222	0.000000000
14	0.085106383	0.00000000	0.914893617	0.000000000
15	0.365853659	0.12195122	0.000000000	0.512195122
16	0.000000000	1.00000000	0.000000000	0.000000000
18	0.047106326	0.62584118	0.017496635	0.309555855
20	0.000000000	1.00000000	0.000000000	0.000000000
23	0.060606061	0.22222222	0.000000000	0.717171717
25	0.526645768	0.13793103	0.150470219	0.184952978
26	0.139534884	0.23255814	0.000000000	0.627906977
28	0.078267477	0.80243161	0.024316109	0.094984802
29	0.025086505	0.81574394	0.050173010	0.108996540

```
30    0.000000000 1.00000000 0.000000000 0.000000000
33    0.023655266 0.94372457 0.015443725 0.017176435
35    0.223404255 0.38297872 0.021276596 0.372340426
36    0.049689441 0.36024845 0.000000000 0.590062112
37    0.490566038 0.00000000 0.000000000 0.509433962
39    0.372340426 0.30851064 0.191489362 0.127659574
40    0.205607477 0.44859813 0.149532710 0.196261682
41    0.020111732 0.66927374 0.008938547 0.301675978
42    0.326530612 0.48979592 0.000000000 0.183673469
43    0.242236025 0.70807453 0.037267081 0.012422360
44    0.047393365 0.95260664 0.000000000 0.000000000
45    0.063424947 0.66596195 0.073995772 0.196617336
47    0.311688312 0.06493506 0.480519481 0.142857143
48    0.125786164 0.69182390 0.094339623 0.088050314
50    0.230769231 0.07692308 0.000000000 0.692307692
51    0.000000000 0.50000000 0.000000000 0.500000000
53    0.000000000 0.00000000 0.666666667 0.333333333
54    0.000000000 1.00000000 0.000000000 0.000000000
55    0.017288444 0.96784956 0.009402487 0.005459509
59    0.005488851 0.99039451 0.000000000 0.004116638
60    0.027210884 0.97278912 0.000000000 0.000000000
```

Un resumen de estos valores viene dado por las proporciones medias de las distintas tareas: 11.5 % (administrativas), 60.8 % (técnicas), 8.3 % (comerciales), y 19.4 % (I+D+i). Estos valores cambian si consideramos grupos particulares de proyectos.

Debido a la presencia de ceros (tratados aquí como ceros redondeados), se hace necesaria una modificación de los datos tal y como se discutió en la Sección 3. Para ello se siguió el método de Aitchison (1986).

Por lo que respecta a las características medidas de los distintos proyectos, tenemos caracterizados los proyectos según la duración del proyecto (variable Tipo, con valores: microproyecto, preproyecto, proyecto o macroproyecto), el tipo de cliente (variable Cliente, binaria, con valores: oficina, no oficina), el servicio ofrecido (variable Servicio, ternaria, con valores: Ingenieria pura, Ingenería y Consultoría), una segunda variable descriptiva del cliente (variable cliente2, tenaria, con valores: cliente, colaborador, propio), y el número de trabajadores técnicos (variable Trab_tec), consultores (Trab_cons), de soporte técnico (Trab_sop) y el total (Trab_total) de los trabajadores dedicados a cada proyecto.

Respecto a Servicio, tenemos 14 proyectos de Consultoría, 16 de Ingeniería, y 8 de Ingeniería pura. Respecto al tipo de proyecto (Tipo), la distribución es 5 macroproyectos,

4 microproyectos, 9 preproyectos, y 13 proyectos. 23 proyectos fueron para oficina y 15 para no oficina (variable Cliente), mientras que la distribución de cliente2 fue la siguiente: 27 clientes, 7 colaboradores y 4 propios. Finalmente, señalamos que un 76 % de los proyectos involucraron tres trabajadores de la empresa o menos, con una media de 2.71 trabajadores involucrados en cada proyecto (desviación típica: 1.68).

El objetivo es plantear y ajustar un modelo de regresión que permita analizar la influencia de las características de los proyectos en la distribución de horas en las cuatro distintas tareas (proporciones). Teniendo en cuenta la naturaleza composicional de la variable respuesta, se utilizó a este fin el modelo de regresión de Dirichlet presentado en la Sección 4. Este modelo se ajustó a los datos utilizando la librería DirichletReg de R, que requiere una preparación preliminar de la variable respuesta. Para ello se utilizó la función

```
DR_data(tabla1[,1:4])
```

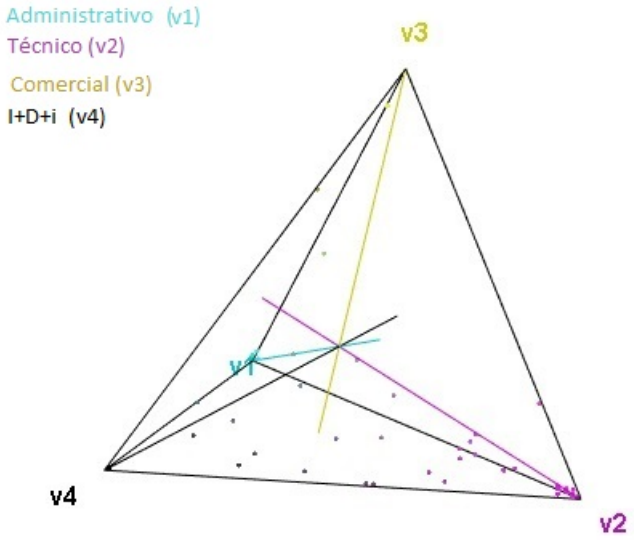
Esta función prepara una matriz con variables de composición para su posterior procesamiento en el paquete DirichletReg

Para la estimación del modelo de Dirichlet y el posterior estudio de la bondad del ajuste utilizaremos las librerias *compositions* y *DirichletReg*. La primera nos proporciona distintas medidas de ajuste del modelo a los datos composicionales, y la segunda ofrece la implementación de la regresión de Dirichlet en R.

Para la representación gráfica de los datos nos ayudamos de los paquetes antes mencionados. Para ello, una vez leido los datos e transformados adecuadamente para el ajsute del modelo, utilizamos el siguiente código

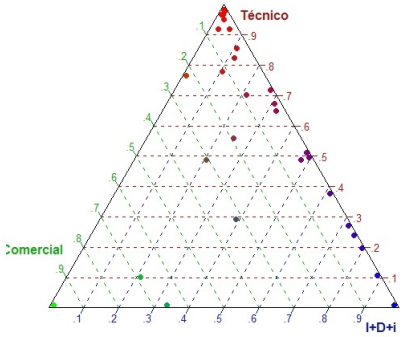
```
plot(DR_data(tabla1[,1:4]))
```

El gráfico que se obtiene es,



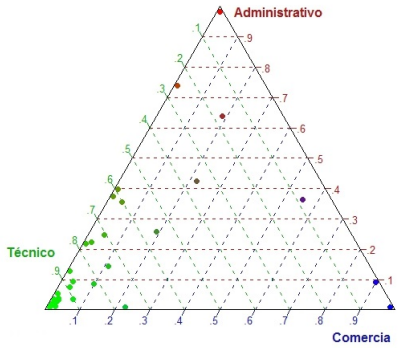
Este gráfico tridimensional muestra cómo la mayor parte de los 38 proyectos se sitúan en valores relativamente altos de las proporciones de tareas técnicas y de I+D+i (vértices correspondientes), de acuerdo con las medias proporcionadas arriba. Por otra parte, pocos proyectos se sitúan en valores altos para la proporción de horas en tareas comerciales o administrativas. La visualización de los datos puede ser más sencilla en gráficos bidimensionales, por lo que representaremos los diagramas ternarios de las diferentes combinaciones entre los tipos de tareas. Para ello utilizaremos las siguientes sentencias:

```
plot(DR_data(Y[,2:4]))
```

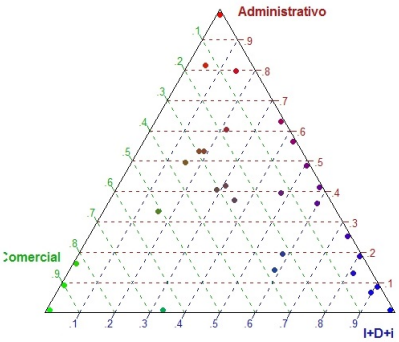


Lo que podemos observar en este gráfico, si eliminamos las tareas de administración, es que la mayoría de proyectos presentan más de un 60 % de tareas del departamento técnico frente a que menos del 25 % del departamento comercial. Respecto al departamento de I+D+i, podemos decir que hay una gran heterogeneidad de proporciones.

Análogamente, para eliminar la variable I+D+i utilizamos:



Este gráfico permite apreciar que las proporciones de tareas administrativas se sitúan en la mayor parte de los casos por debajo del 25 %. Se puede obtener un tercer gráfico eliminando las tareas técnicas, de la siguiente forma:



Este gráfico repite de alguna forma la información ya visualizada en los dos gráficos bidimensionales anteriores. El cuarto posible gráfico ya no es mostrado por la misma razón.

De estos tres gráficos, podemos destacar que los proyectos se carácterizan, especialmente, por altos porcentajes del departamento técnico frente al porcentaje bajo de las tareas comerciales. Esto tiene sentido, ya que la empresa Azteca es principalmente de

carácter técnico.

Para estimar el modelo de Dirichlet de regresión hemos utilizado el paquete DirichletReg, en concreto la función DirichReg de este paquete. Hemos tomado el valor por defecto del parámetro model ("common"), lo cual ajusta el modelo con link logarítmico $log\lambda_{ij} = \beta_j z_i$, donde λ_{ij} es el parámetro λ_j del modelo de Dirichlet (j indica tipo de tarea) para un proyecto de características z_i, y $beta_j$ es el vector de coeficientes del modelo (a estimar) para la tarea j.

Para la selección del modelo (es decir, de las variables explicativas), lo que se ajustaron los diferentes posibles modelos combinando las diferentes características de los proyectos y evaluando el AIC como medida de bondad en el ajuste. El modelo finalmente seleccionado (menor AIC) es el que tiene como variables explicativas Tipo, Servicio, Cliente y el número total de trabajadores dedicados al proyecto (Trab_total). Los resultados del modelo ajustado son los siguientes:

```
>modelo<-DirichReg(Y1_trans ~ Tipo + Servicio + Cliente+Trab_total, tabla1)
>summary(modelo)

DirichletReg(formula = Y1_trans ~ Tipo + Servicio + Cliente + Trab_total, data =
tabla1)

Standardized Residuals:
                    Min       1Q   Median      3Q      Max
Administrativo  -1.5569  -0.6796  -0.1225  0.4062   3.1595
Técnico         -5.6068  -0.8463   0.1903  0.7871   1.9915
Comercial       -2.5480  -0.7464  -0.3958  0.1338   3.4985
I+D+i           -1.6351  -0.7046  -0.0099  0.8110  15.4893

-------------------------------------------------------------
Beta-Coefficients for variable no. 1: Administrativo
                         Estimate Std. Error z-Value p-Value
(Intercept)               8.12722    1.15291   7.049 1.8e-12 ***
TipoMicroproyecto        -8.23678    0.97676  -8.433 < 2e-16 ***
TipoPreproyecto          -9.57261    1.03223  -9.274 < 2e-16 ***
TipoProyecto             -9.24170    0.91436 -10.107 < 2e-16 ***
TipoServicio técnico    -10.07046    0.80283 -12.544 < 2e-16 ***
ServicioIngeniería       -0.38921    0.62112  -0.627 0.53090
ServicioIngeniería pura   3.04261    0.83304   3.652 0.00026 ***
ClienteOficina            0.97033    0.53736   1.806 0.07096 .
Trab_total               -0.05535    0.17141  -0.323 0.74674
-------------------------------------------------------------
Beta-Coefficients for variable no. 2: Técnico
```

```
                         Estimate Std. Error z-Value  p-Value
(Intercept)              12.36003    1.11159  11.119  < 2e-16 ***
TipoMicroproyecto       -11.01608    1.00026 -11.013  < 2e-16 ***
TipoPreproyecto         -12.42931    0.99394 -12.505  < 2e-16 ***
TipoProyecto            -10.80056    0.85828 -12.584  < 2e-16 ***
TipoServicio técnico    -11.19702    0.81217 -13.787  < 2e-16 ***
ServicioIngeniería        0.86937    0.63561   1.368   0.1714
ServicioIngeniería pura   4.98089    0.85170   5.848 4.97e-09 ***
ClienteOficina           -0.09018    0.53552  -0.168   0.8663
Trab_total               -0.35056    0.17654  -1.986   0.0471 *
---------------------------------------------------------------
Beta-Coefficients for variable no. 3: Comercial
                        Estimate Std. Error z-Value  p-Value
(Intercept)              7.14468    1.09123   6.547 5.86e-11 ***
TipoMicroproyecto       -9.21300    0.95885  -9.608  < 2e-16 ***
TipoPreproyecto         -9.63293    0.99667  -9.665  < 2e-16 ***
TipoProyecto            -9.60419    0.85320 -11.257  < 2e-16 ***
TipoServicio técnico    -9.84894    0.78346 -12.571  < 2e-16 ***
ServicioIngeniería       0.51607    0.59527   0.867 0.385966
ServicioIngeniería pura  3.65670    0.81046   4.512 6.43e-06 ***
ClienteOficina           1.77768    0.53625   3.315 0.000916 ***
Trab_total              -0.09304    0.17745  -0.524 0.600077
---------------------------------------------------------------
Beta-Coefficients for variable no. 4: I+D+i
                        Estimate Std. Error z-Value  p-Value
(Intercept)               5.4506     1.0316   5.284 1.27e-07 ***
TipoMicroproyecto        -8.6010     0.9928  -8.663  < 2e-16 ***
TipoPreproyecto         -10.0953     0.9573 -10.546  < 2e-16 ***
TipoProyecto             -8.6941     0.8679 -10.017  < 2e-16 ***
TipoServicio técnico     -9.5683     0.7157 -13.368  < 2e-16 ***
ServicioIngeniería       -0.6504     0.6652  -0.978   0.3282
ServicioIngeniería pura   3.3231     0.8222   4.042 5.31e-05 ***
ClienteOficina            3.6787     0.5389   6.827 8.69e-12 ***
Trab_total                0.3242     0.1858   1.746   0.0809 .
---------------------------------------------------------------
Signif. codes: '***' < .001, '**' < 0.01, '*' < 0.05, '.' < 0.1

Log-likelihood: 245.7 on 36 df (198+3 iterations)
AIC: -419.4, BIC: -360.4797
Number of Observations: 38
Link: Log
Parametrization: common
```

En estos resultados se observa que todas las variables incluidas en el modelo final alcanzan significación estadística para alguna de las cuatro proporciones en el modelo. Excepción a esto es el indicador de servicio de ingeniería, que no es significativo; sin embargo, tiene sentido mantenerlo en el modelo al ser uno de los valores de la variable Servicio.

Si nos centramos en la primera de las proporciones (tareas administrativas), vemos que todas las variables son significativas excepto la mencionada "servicio de ingeniería", Cliente, y número total de trabajadores. Respecto de la segunda (tareas técnicas), ocurre lo mismo excepto en que Trab_total tiene significación estadística. En el caso de tareas comerciales, se produce un cambio de variables significativas (respecto a técnicas): Cliente tiene significación y Trab_total la pierde. Finalmente, para tareas de I+D+i identificamos las mismas variables significativas que para tareas comerciales.

A continuación analizamos la bondad en el ajuste del modelo, mediante el R^2 y el análisis de residuos. Para el cálculo del R^2 utilizamos el paquete `compositions`, que contiene las funciones `acomp` y `variation`. La función `acomp` permtie transformar los datos para que tengan estructura composicional de acuerdo con la filosofía de Aitchison. La función `variation` permite calcular la matriz de variación $T(x)$ definida en la Sección 4.

El código utilizado

```
Y1_orig<-tabla1[,1:4]
> (sum(variation(acomp(modelo$fitted$mu)))/8) / (sum(variation(acomp(Y1_orig)))/8)

[1] 0.909482
```

Puesto que el valor del R^2 es grande, las proporciones estimadas por el modelo construido y las proporciones observadas son muy similares, por lo que podemos considerar que el modelo ofrece un ajuste razonable.

Otra forma de comparar si el modelo ajustado se asimila al modelo real de los datos es hacer las comparaciones de las proporciones originales (observadas) con las proporciones ajustadas; para ello nos fijaremos en la media y mediana de los datos originales y de los ajustados.

```
> summary(modelo$d$Y1_trans[,])
 Administrativo        Técnico           Comercial             IDi
 Min.   :0.006579   Min.   :0.006579   Min.   :0.006579   Min.   :0.006579
 1st Qu.:0.010078   1st Qu.:0.319565   1st Qu.:0.006579   1st Qu.:0.006579
 Median :0.052585   Median :0.669219   Median :0.015508   Median :0.105886
```

```
 Mean   :0.118204   Mean   :0.598965   Mean   :0.087781   Mean   :0.195051
 3rd Qu.:0.212433   3rd Qu.:0.948627   3rd Qu.:0.052290   3rd Qu.:0.325352
 Max.   :0.519366   Max.   :0.980263   Max.   :0.897396   Max.   :0.704878
```

```
> summary(modelo$fitted.values$mu)
 Administrativo        Técnico          Comercial             IDi
 Min.   :0.008506   Min.   :0.08601   Min.   :0.006048   Min.   :0.001871
 1st Qu.:0.026839   1st Qu.:0.29369   1st Qu.:0.019403   1st Qu.:0.011871
 Median :0.072088   Median :0.52548   Median :0.071131   Median :0.223486
 Mean   :0.111795   Mean   :0.60984   Mean   :0.083999   Mean   :0.194371
 3rd Qu.:0.196381   3rd Qu.:0.93243   3rd Qu.:0.123421   3rd Qu.:0.329881
 Max.   :0.274262   Max.   :0.98255   Max.   :0.206997   Max.   :0.570815
```

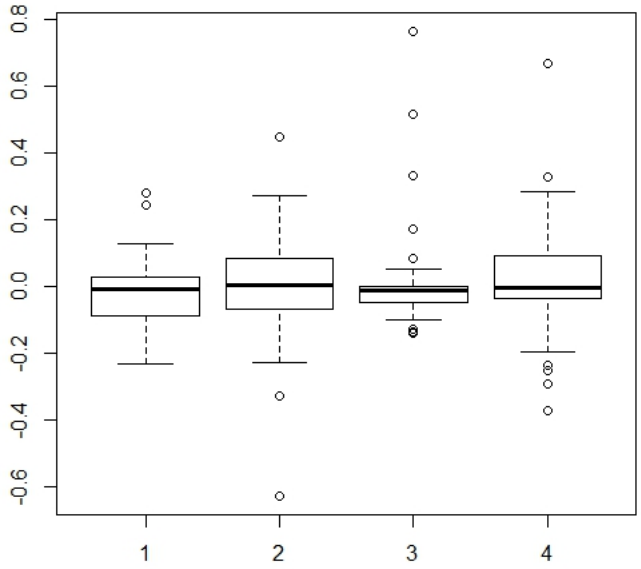
En estas comparaciones lo que se puede observar es que los datos originales (primer summary) y los datos ajustados (segundo summary) ofrecen valores muy similares en cuanto a media o mediana, indicando un buen ajuste del modelo.

El estudio de la bondad en el ajuste puede realizarse también a partir del análisis de residuos. Existen distintos tipos de residuos definidos para el modelo de Dirichlet. Gueorguieva et al. (2008) describen tres tipos de residuos. Los residuos crudos (raw residuals) se definen como las diferencias simples entre valores observados y predichos de las proporciones; los residuos tipificados, definidos como los residuos crudos divididos por la desviación típica estimada; y los residuos compuestos (composite residuals) que se definen como la suma de los cuadrados de los residuos tipificados.

Un análisis descriptivo de los residuos tipificados aparece incorporado al comienzo del output del modelo ajustado (arriba). En ese resumen se aprecia que los residuos tipificados están distribuidos en torno a cero, con un nivel de dispersión y asimetría que depende de la tarea considerada.

Hemos calculado también los residuos crudos, que son simplemente como se ha indicado $x_{ij} - \frac{\lambda_{ij}}{\lambda_j}$. El siguiente gráfico es un diagrama de caja (boxplot) de los residuos crudos:

```
> res.raw=residuals(modelo, "raw")
> boxplot(res.raw[,1],res.raw[,2],res.raw[,3],res.raw[,4])
```



De nuevo se puede observar que los residuos tienen pequeñas oscilaciones en torno a cero, por lo que el modelo estimado ajusta bien los datos. Por lo que podemos concluir que el modelo elegido es un buen modelo para explicar la variabilidad de la proporción de horas dedicadas a las cuatro tareas consideradas.

En resumen, el modelo de regresión de Dirichlet ha permitido identificar variables que afectan significativamente a la distribución de tareas en los proyectos de la empresa Azteca, y se ha mostrado como un modelo con un buen ajuste a los datos disponibles.

Capítulo 6

Conclusión

En esta memoria se explica que partiendo de unos datos composicionales, podemos utilizar los modelos de Dirichlet. Además se aplican al caso particular de los datos reales ofrecidos por la empresa Azteca Consulting de Ingeniería S.L, podiendo concluir que los modelos de Dirichlet son flexibles ante este tipo de datos y se ajustan bien al caso particular de la estimación de proporciones de horas dedicadas a las cuatro tareas.

Referencias

[1] Aitchison, J. (1982) , *The statistical analysis of compositional data. Journal of the Royal Statistical Society, Vol. 44, No. 2.*

[2] Aitchison, J. (1986) , *The Statistical Analysis of Compositional Data. The Statistical Analysis of Compositional Data. Chapman and Hall, New york, XII, 416 pp.*

[3] Aitchison, J., Barcelo-Vidal, C., Martin-Fernandez, J. y Pawlowsky-Glahn, V. (2000), *Logratio analysis and compositional distance. Mathematical Geology. Volume 33, Issue 7 , pp 845-848.*

[4] Gueorguieva R, Rosenheck R, Zelterman D (2008) *Dirichlet component regression and its applications to psychiatric data. Computational Statistics and Data Analysis 52, 5344-5355.*

[5] Hijazi,R. (2007), *Residuals and diagnostics in Dirichlet regression. ASA Proceeding of the Joint Statistical Meeting 2006, American Statistical Association.*

[6] Hijazi,R. (2009), *Dealing with Rounded Zeros in Compositional Data under Dirichlet Models. Proceeding of the 10th Islamic Countries Conference on Statistical Sciences 2009.*

[7] Hijazi RH, Jernigan RW (2007) *Modelling Compositional Data Using Dirichlet Regression Models. Preprint.*

[8] *http://cran.r-project.org/web/packages/compositions/compositions.pdf*

[9] *http://cran.r-project.org/web/packages/DirichletReg/DirichletReg.pdf*

[10] Hua-Yan Wang, Quiang Yang, Hong Qin, Hongbin Zha, *Dirichlet Component Analysis: Feature Extraction for Compositional Data, The 25th International*

Conference on Machine Learning (ICML 2008).

[11] Marco J. Maier, *DirichletReg: Dirichlet Regression for Compositional Data in R.*

[12] Murteira JMR, Ramalho JJS (2012), *Regression Analysis of Multivariate Fractional Data. Preprint.*

[13] Palarea-Albaladejo, J., Martín-Fernández, J. (2008), *A modified EM alr-algorithm for replacing rounded zeros in compositional data set. Computers & Geosciences.*

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