

Técnicas multivariadas

Análisis de correspondencias

Mario J. P. López

Departamento de Matemáticas
Universidad el Bosque

2020

Tipos de enfermedades respiratorias de acuerdo al IMC

Categoría IMC	Enfermedad Respiratoria				Total
	Obstructiva	Restrictiva	Combinada	Normal	
Bajo peso	13	40	40	101	194
Normal	22	16	18	57	113
Sobrepeso	10	12	5	62	89
Obeso	72	50	29	215	366
Total	117	118	92	435	762

Objetivos del AC

El AC Simple, o simplemente AC, es una herramienta para analizar las asociaciones entre las filas y columnas de una tabla de contingencia

$$\mathbf{X} = \begin{bmatrix} x_{11} & x_{12} & \cdots & x_{1j} & \cdots & x_{1q} \\ x_{21} & x_{22} & \cdots & x_{2j} & \cdots & x_{2q} \\ \vdots & \vdots & & \vdots & & \vdots \\ x_{i1} & x_{i2} & \cdots & x_{ij} & \cdots & x_{iq} \\ \vdots & \vdots & & \vdots & & \vdots \\ x_{p1} & x_{p2} & \cdots & x_{pj} & \cdots & x_{pq} \end{bmatrix}_{(p \times q)}$$

Estadística χ^2

Una forma de medir asociación entre las filas y columnas de una tabla de contingencias es a través de la estadística

$$\chi^2 = \sum_{i=1}^p \sum_{j=1}^q \left(\frac{x_{ij} - E_{ij}}{E_{ij}} \right)^2$$

donde

$$E_{ij} = \frac{x_{i.} \cdot x_{.j}}{x_{..}}$$

bajo la hipótesis de independencia $\chi^2 \sim \chi^2_{(p-1, q-1)}$.

Ejemplo

```
> X = matrix(c(13,40,40,101,  
               22,16,18, 57,  
               10,12, 5, 62,  
               72,50,29,215),4,byrow = T)  
  
> chisq.test(X)
```

Pearson's Chi-squared test

data: X

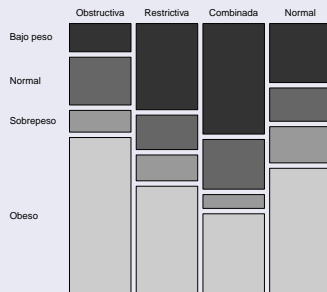
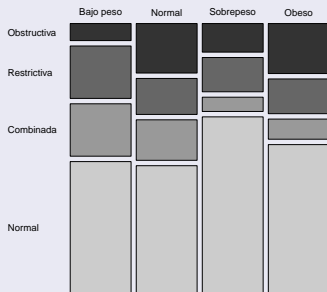
X-squared = 46.371, df = 9, p-value = 5.139e-07

Se rechaza entonces la hipótesis de independencia entre el peso y el tipo de enfermedad respiratoria.

Ejemplo

```
> X = as.table(X) # tabla de contingencia
> rownames(X) = c("Bajo peso", "Normal",
                  "Sobrepeso", "Obeso")
> colnames(X) = c("Obstructiva", "Restrictiva",
                  "Combinada", "Normal")
> plot(prop.table(X,1))
> plot(prop.table(t(X),1))
```

Ejemplo



Descomposición de la χ^2

Consiste en encontrar la descomposición en valores singulares de la matriz

$$C = \begin{bmatrix} c_{11} & c_{12} & \cdots & c_{1j} & \cdots & c_{1q} \\ c_{21} & c_{22} & \cdots & c_{2j} & \cdots & c_{2q} \\ \vdots & \vdots & & \vdots & & \vdots \\ c_{i1} & c_{i2} & \cdots & c_{ij} & \cdots & c_{iq} \\ \vdots & \vdots & & \vdots & & \vdots \\ c_{p1} & c_{p2} & \cdots & c_{pj} & \cdots & c_{pq} \end{bmatrix}_{(p \times q)}$$

con elementos

$$c_{ij} = \frac{x_{ij} - E_{ij}}{E_{ij}^{1/2}}$$

Descomposición de la χ^2

Los elementos c_{ij} pueden ser vistos como mediciones ponderadas de las discrepancias entre los valores observados x_{ij} y los valores esperados E_{ij} bajo independencia.

Descomposición de la χ^2

Considere las matrices

$$\begin{aligned} A &= \text{diag}(x_{i.}) \\ &= \begin{bmatrix} x_{1.} & 0 & 0 & 0 \\ 0 & x_{2.} & 0 & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & x_{p.} \end{bmatrix} \end{aligned} \quad \begin{aligned} B &= \text{diag}(x_{.j}) \\ &= \begin{bmatrix} x_{.1} & 0 & 0 & 0 \\ 0 & x_{.2} & 0 & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & x_{.q} \end{bmatrix} \end{aligned}$$

y sean

$$\mathbf{a} = A\mathbf{1}_p$$

$$\mathbf{b} = B\mathbf{1}_q$$

los vectores con las frecuencias marginales fila y columna, respectivamente.

Descomposición de la χ^2

Considere la descomposición en valores singulares

$$C = \Gamma \Lambda \Delta^T$$

donde Γ contiene los vectores propios de CC^T , Δ los vectores propios de $C^T C$ y

$$\Lambda = \text{diag} \left\{ \lambda_1^{1/2}, \lambda_2^{1/2}, \dots, \lambda_s^{1/2} \right\}$$

con $\lambda_1 \geq \lambda_2 \geq \dots \geq \lambda_s$ los valores propios de CC^T diferentes de cero.

Ejemplo

```
n = sum(X)
xi. = margin.table(X,1)
x.j = margin.table(X,2)
E = xi.%%t(x.j)/n ; E
C = (X-E)/sqrt(E)
dvs = svd(C)
```

dvs\$d	dvs\$u				dvs\$v			
$\lambda_k^{1/2}$	γ_1	γ_2	γ_3	γ_4	δ_1	δ_2	δ_3	δ_4
6.099	-0.82	0.14	0.22	0.51	0.56	-0.73	0.10	0.39
2.978	-0.05	-0.62	-0.68	0.39	-0.34	0.08	0.85	0.39
0.557	0.20	0.76	-0.51	0.34	-0.73	-0.43	-0.41	0.35
~0.00	0.53	-0.14	0.47	0.69	0.22	0.53	-0.31	0.76

Descomposición de la χ^2

De esta forma

$$c_{ij} = \sum_{k=1}^s \lambda_k^{1/2} \gamma_{ik} \delta_{jk}$$

y

$$\text{tr}(CC^T) = \sum_{k=1}^s \lambda_k = \sum_{i=1}^p \sum_{j=1}^q c_{ij}^2 = \chi^2$$

Descomposición de la χ^2

Suponga ahora que

$$c_{ij} = \lambda_1^{1/2} \gamma_{i1} \delta_{j1}$$

- Cuando γ_{i1} y δ_{j1} son ambas grandes y con el mismo signo en relación a otras coordenadas i, j , entonces c_{ij} será grande también, indicando una asociación positiva entre las categorías i y j .
- Cuando γ_{i1} y δ_{j1} son ambas grandes y con signos opuestos en relación a otras coordenadas i, j , entonces c_{ij} será grande también, indicando una asociación negativa entre las categorías i y j .

Descomposición de la χ^2

- En muchas aplicaciones los dos primeros valores propios, λ_1 y λ_2 , son dominantes, y el porcentaje del total de la χ^2 explicado por los vectores propios γ_1 , γ_2 y δ_1 , δ_2 es grande.
- De esta forma se pueden emplear para obtener una representación gráfica entre las filas o las columnas de la tabla de contingencia.

Proyecciones

Las proyecciones de las filas y columnas de C están dadas por

$$\mathbf{y}_k = C\delta_k \quad \text{y} \quad \mathbf{z}_k = C^\top \gamma_k, \quad k = 1, \dots, s$$

Pero en análisis de correspondencias usamos las proyecciones de las filas y las columnas ponderadas de C

$$\mathbf{r}_k = A^{-1/2} C\delta_k$$

$$\mathbf{s}_k = B^{-1/2} C^\top \gamma_k$$

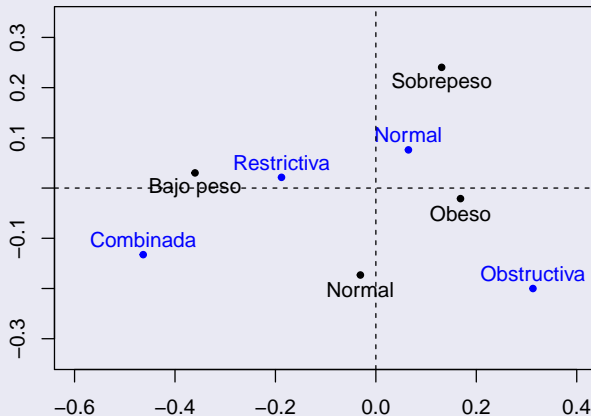
$$k = 1, \dots, s.$$

Ejemplo

```
A = diag(xi.)
B = diag(x.j)
rk = diag(xi.^(-.5))%*%C%*%dvs$v
sk = diag(x.j^(-.5))%*%t(C)%*%dvs$u

# Biplot
plot(rk[,1],rk[,2],asp = 1, pch = 20,
      xlim=c(-0.6,0.4), ylim=c(-0.1,0.1))
points(sk[,1],sk[,2],asp = 1, col = 2, pch = 20)
text(rk[,1],rk[,2]+.03, rownames(X))
text(sk[,1],sk[,2]+.03, colnames(X), col = 2)
abline(h=0,v=0,lty=2)
```

Ejemplo



Resultado

Los vectores r_k y s_k tienen media y varianza ponderada:

$$\bar{r}_k = \frac{\sum_{i=1}^p r_{ik} a_i}{\sum_{i=1}^p a_i} = \frac{\mathbf{r}_k^\top \mathbf{a}}{\mathbf{1}_p^\top \mathbf{a}} = 0$$

$$\bar{s}_k = \frac{\sum_{j=1}^q s_{jk} b_j}{\sum_{j=1}^q b_j} = \frac{\mathbf{s}_k^\top \mathbf{b}}{\mathbf{1}_q^\top \mathbf{b}} = 0$$

y

$$\text{Var}(r_k) = \frac{\sum_{i=1}^p r_{ik}^2 a_i}{\sum_{i=1}^p a_i} = \frac{\mathbf{r}_k^\top \mathbf{A} \mathbf{r}_k}{\mathbf{1}_p^\top \mathbf{a}} = \frac{\lambda_k}{n}$$

$$\text{Var}(s_k) = \frac{\sum_{j=1}^q s_{jk}^2 b_j}{\sum_{j=1}^q b_j} = \frac{\mathbf{s}_k^\top \mathbf{B} \mathbf{s}_k}{\mathbf{1}_q^\top \mathbf{b}} = \frac{\gamma_k}{n}$$

Medidas de calidad

- A partir de las expresiones de $Var(r_k)$ y $Var(s_k)$,

$$I_k = \frac{\lambda_k}{\sum_{k=1}^s \lambda_k}$$

se puede interpretar como la proporción de la varianza explicada por \mathbf{r}_k y \mathbf{s}_k .

- Las proporciones

$$C_a(i, \mathbf{r}_k) = \frac{x_{i \cdot} r_{ki}^2}{\lambda_k}, \quad C_b(j, \mathbf{s}_k) = \frac{x_{\cdot j} s_{kj}^2}{\lambda_k}$$

son la contribuciones absolutas de la fila i y la columna j a la varianza de \mathbf{r}_k y \mathbf{s}_k , respectivamente. Estas son medidas de la importancia de las categorías fila y columna.

Ejemplo

```
Ik = dvs$d^2/sum(dvs$d^2) ; round(Ik,3)
```

I_k	0.802	0.191	0.007	0.000
-------	-------	-------	-------	-------

```
Ca = A%*%rk^2%*%diag(dvs$d^(-2)) ; round(Ca,2)
```

	Ca			
	$C_a(i, r_1)$	$C_a(i, r_2)$	$C_a(i, r_3)$	$C_a(i, r_4)$
Bajo peso	0.68	0.02	0.05	- -
Normal	0.00	0.38	0.47	- -
Sobrepeso	0.04	0.58	0.26	- -
Obeso	0.28	0.02	0.22	- -

Ejemplo

```
Cb = B%%sk^2%%diag(dvs$d^(-2)) ; round(Cb,2)
```

	Cb			
	$C_b(j, s_1)$	$C_b(j, s_2)$	$C_b(j, s_3)$	$C_b(j, s_4)$
Obstruativa	0.31	0.53	0.01	- -
Restringida	0.11	0.01	0.73	- -
Combinada	0.53	0.18	0.17	- -
Normal	0.05	0.28	0.10	- -

ACM

- Extensión del Análisis de Correspondencias a más de dos variables categóricas
- Considere un conjunto de n individuos y p variables. Cada variable con c_j modalidades, $j = 1, \dots, p$.
- El método consiste de un AC a una matriz indicadora \mathbf{X} con elementos

$$x_{ik} = \begin{cases} 1, & \text{si el individuo } i \text{ seleccionó la categoría } k \\ 0, & \text{en otro caso} \end{cases}$$

con $i = 1, \dots, n$, $k = 1, \dots, m$ y $m = \sum_{j=1}^p c_j$




ACM

Matriz indicadora

$$\mathbf{X} = \left[\begin{array}{ccc|ccc|c|c|c} x_{11} & \cdots & x_{1c_1} & x_{1(c_1+1)} & \cdots & x_{1(c_1+c_2)} & \cdots & \cdots & x_{1m} \\ x_{21} & \cdots & x_{2c_1} & x_{2(c_1+1)} & \cdots & x_{2(c_1+c_2)} & \cdots & \cdots & x_{2m} \\ \vdots & & \vdots & \vdots & & \vdots & & & \vdots \\ x_{n1} & \cdots & x_{nc_1} & x_{n(c_1+1)} & \cdots & x_{n(c_1+c_2)} & \cdots & \cdots & x_{nm} \end{array} \right]$$

ACM

- El análisis de correspondencias múltiple consistirá de un análisis de correspondencias sobre la matriz \mathbf{X} o sobre las matrices $\mathbf{X}^T \mathbf{X}$ o $\mathbf{X} \mathbf{X}^T$.
- La matriz $\mathbf{X}^T \mathbf{X}$ se conoce como matriz de Burt y fuera de su bloque diagonal contiene las tablas de contingencia entre pares de variables.
- Un análisis de correspondencias sobre \mathbf{X} es equivalente a un análisis de correspondencias sobre $\mathbf{X}^T \mathbf{X}$ o $\mathbf{X} \mathbf{X}^T$.

-  Johnson, R., Wichern, D.
Applied Multivariate Statistical Analysis.
6th ed. Pearson, 2007.
-  Mardia, K., Kent, J., Bibby, J.
Multivariate Analysis.
Academic Press, 1995.
-  Rencher, A.
Methods of Multivariate Analysis.
2nd ed. Willey, 2002.