Trabajo Práctico 1 Análisis Multivariado

Mariana Vargas V.

June 28, 2017

1 Introducción

Trabajamos realizando un análisis estadístico de una base de datos que consiste en datos recopilados acerca de ciertas variables económicas de algunos países, extraídos de la página del Banco Mundial.

2 Ejercicio 1

Realizamos un análisis exploratorio de los datos a partir de medidas descriptivas y gráficos. Veamos por ejemplo el box-plot de las variables de la figura 1.

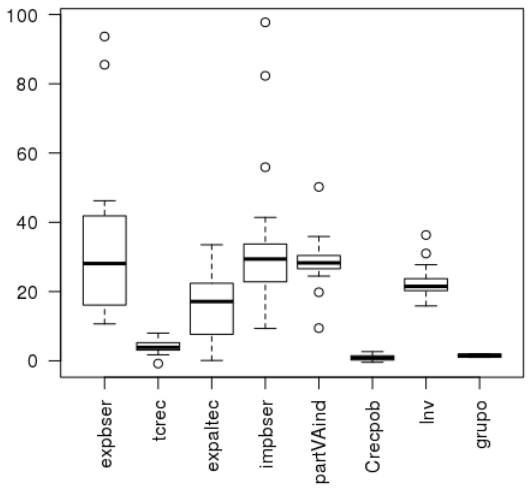

Figure 1: Boxplots para todas las variables.

Notar que las variables que corresponden a importación y exportación de servicios, y participación del valor agregado industrial tienen las medias más altas, mientras que la tasa de crecimiento del PBI y el crecimiento poblacional tienen las más bajas. Podemos confirmarlo en la tabla 1.

Podemos también observar los histogramas para cada variable, lo cual se detalla en la figura 2. Podemos notar como muchas de las variables son bastante sesgadas0.

	expbser	tcrec	expaltec	impbser	partVAind	Crecpob	Inv
1	Min. :10.66	-0.790	0.100	9.34	9.44	-0.440	15.85
2	1st Qu.:17.13	3.095	7.705	22.91	26.63	0.250	20.34
3	Median :28.09	3.865	17.130	29.39	28.27	0.875	21.50
4	Mean :31.41	4.131	15.649	32.69	28.77	1.000	22.25
5	3rd Qu.:39.84	5.070	22.332	33.63	30.36	1.450	23.46
6	Max. :93.65	8.000	33.520	97.74	50.22	2.640	36.33

Table 1: Medidas descriptivas de las variables.

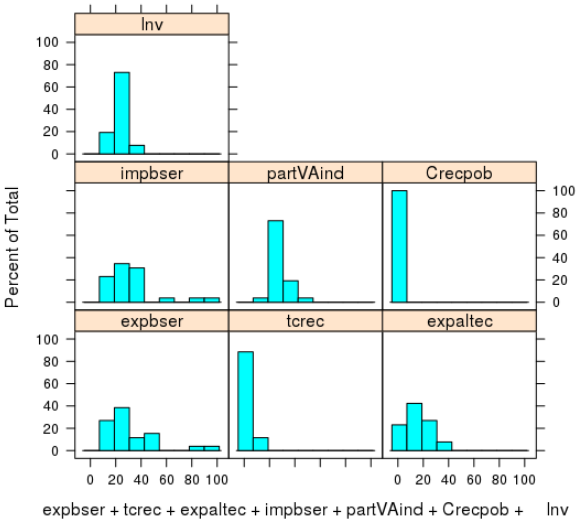

Figure 2: Histogramas para todas las variables.

Analizando las matrices de varianzas y covarianzas, y la de correlaciones entre variables de las tablas 2 y 3 podemos extraer algunas conclusiones interesantes. En primer lugar, las exportaciones e importaciones tienen una alta varianza, incluyendo la exportación de alta tecnología (en menor medida) y una correlación positiva entre sí, mientras que variables como el crecimiento poblacional y la exportación de servicios o tecnología mantienen una correlación negativa. Esto sugiere que países con industrias de alto valor agregado que exportan productos de alta tecnología no tienen un crecimiento poblacional marcado, y viceversa.

Para un análisis más preciso hay que tener en cuenta que Estonia y China producen valores atípicos detectados a través del cómputo de las distancias de Mahalanobis, que para estos países arroja valores de 14.39 y 14.42 respectivamente.

	expbser	tcrec	expaltec	impbser	partVAind	Crecpob	Inv
expbser	422.28	12.92	32.20	395.67	18.01	-8.15	19.33
tcrec	12.92	3.68	3.60	14.65	1.95	0.05	4.78
expaltec	32.20	3.60	99.66	13.59	-1.78	-4.93	-2.49
impbser	395.67	14.65	13.59	400.17	0.42	-4.70	20.40
partVAind	18.01	1.95	-1.78	0.42	44.94	-1.82	20.61
Crecpob	-8.15	0.05	-4.93	-4.70	-1.82	0.72	-0.17
Inv	19.33	4.78	-2.49	20.40	20.61	-0.17	19.88

Table 2: Matriz de varianzas y covarianzas.

	expbser	tcrec	expaltec	impbser	partVAind	Crecpob	Inv
expbser	1.00	0.33	0.16	0.96	0.13	-0.47	0.21
tcrec	0.33	1.00	0.19	0.38	0.15	0.03	0.56
expaltec	0.16	0.19	1.00	0.07	-0.03	-0.58	-0.06
impbser	0.96	0.38	0.07	1.00	0.00	-0.28	0.23
partVAind	0.13	0.15	-0.03	0.00	1.00	-0.32	0.69
Crecpob	-0.47	0.03	-0.58	-0.28	-0.32	1.00	-0.04
Inv	0.21	0.56	-0.06	0.23	0.69	-0.04	1.00

Table 3: Matriz de correlaciones.

3 Ejercicio 2

Nos concentramos ahora en estudiar la distribución de estas variables. Nuestra pregunta es si son variables normales, lo que abordamos a través de pruebas de hipótesis y pruebas gráficas, tanto univariadas como multivariadas. Entre las primeras podemos trabajar con qqplots para cada una de las variables, lo que esta sumarizado en la figura 4. A simple vista podemos notar que las importaciones y exportaciones de bienes y servicios, las inversiones, y la participación del valor agregado no tienen una distribución normal. Para confirmar esto realizamos un test de Shapiro-Wilks: para las exportaciones e importaciones de bienes y servicios obtenemos p-valores de 0.0002 y 0.0001 respectivamente, para la participación del VA obtenemos 0.001 y, por último, 0.0068 para las inversiones, confirmando así nuestra hipótesis. Las demás variables tienen una distribución normal.

Entre las pruebas multivariadas podemos usar un qqplot de las distancias de Mahalanobis ordenadas (figura 3), que bajo una distribución normal se distribuyen como una χ^2. El gráfico de los cuantiles empíricos versus los de la χ^2 con siete grados de libertad sugiere que la distribución es normal multivariada. Para completar nuestro análisis calculamos la asimetría y kurtosis, obteniendo 0.34 y 1.586 respectivamente, confirmando así que la distribución es una normal multivariada.

4 Ejercicio 3

Se nos pide realizar una prueba de hipótesis para verificar si el vector dado $mu_h0 = (30, 4, 15, 30, 27, 0.8, 20)$ es efectivamente el vector de medias de la población. Calculamos entonces el estadístico T^2 de Hotelling obteniendo $T^2 =$

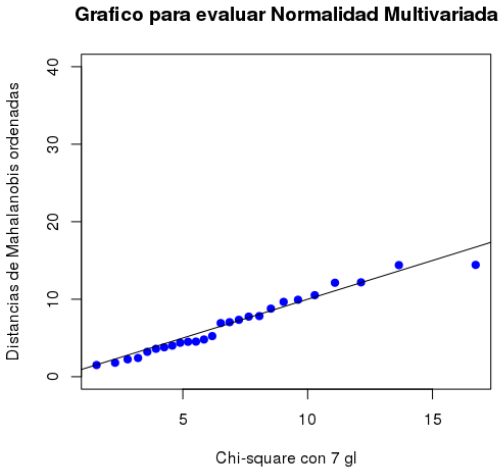

Figure 3: qq-plot de las distancias de Mahalanobis ordenadas.

18.57 mientras que el valor crítico para una confianza del 95% es 23.42. Luego aceptamos la hipótesis nula de que mu_h0 es el vector verdadero de medias.

5 Ejercicio 4

Realizamos el análisis en dos grupos separados: países desarrollados y países subdesarrollados.

Vemos las medidas descriptivas para los países del primer grupo en la tabla 4 y las del segundo en la tabla 5.

	expbser	tcrec	expaltec	impbser	partVAind	Crecpob	Inv
1	Min. :10.76	1.750	7.64	9.34	24.45	0.0800	17.31
2	1st Qu.:27.90	3.080	15.22	27.31	26.71	0.1500	20.59
3	Median :30.14	3.720	19.31	32.41	28.96	0.2500	20.87
4	Mean :35.40	3.593	20.39	33.53	28.73	0.4785	21.27
5	3rd Qu.:45.83	4.180	27.33	40.20	30.23	0.7400	21.55
6	Max. :85.50	5.530	33.52	82.29	34.49	1.2700	26.15

Table 4: Medidas descriptivas de países desarrollados.

Notar como las medias de las tasas de crecimiento son sistemáticamente más bajas en los países subdesarrollados (lo podemos ver prestando atención a las medianas, que son una medida más robusta con respecto a los valores atípicos), así también como los niveles de exportación tanto de bienes y servicios como de tecnología. En cambio, el crecimiento poblacional tiende a ser mayor.

A partir de las matrices de covarianzas y correlaciones de las tablas 6, 7, 8 y 9 podemos ver cómo en los países desarrollados las variables presentan mayor varianza. También observamos que en los países desarrollados la exportación de bienes y servicios tiene una relación negativa respecto de la exportación de alta tecnología mientras que en los países subdesarrollados esta relación es positiva.

	expbser	tcrec	expaltec	impbser	partVAind	Crecpob	Inv
1	Min. :10.66	-0.790	0.10	11.52	9.44	-0.440	15.85
2	1st Qu.:15.08	3.610	3.61	18.02	26.60	1.230	20.26
3	Median :23.60	4.400	7.90	28.76	28.01	1.460	22.49
4	Mean :27.43	4.669	10.91	31.86	28.81	1.522	23.24
5	3rd Qu.:29.75	6.570	18.58	31.53	32.08	2.240	24.51
6	Max. :93.65	8.000	29.84	97.74	50.22	2.640	36.33

Table 5: Medidas descriptivas de países subdesarrollados.

	expbser	tcrec	expaltec	impbser	partVAind	Crecpob	Inv
expbser	369.49	6.37	-74.90	331.07	4.60	-2.35	-10.28
tcrec	6.37	0.91	1.02	4.47	1.13	-0.06	-0.29
expaltec	-74.90	1.02	72.87	-73.75	-0.36	0.21	-6.16
impbser	331.07	4.47	-73.75	305.99	-0.44	-1.74	-8.59
partVAind	4.60	1.13	-0.36	-0.44	7.72	-0.71	1.85
Crecpob	-2.35	-0.06	0.21	-1.74	-0.71	0.17	0.11
Inv	-10.28	-0.29	-6.16	-8.59	1.85	0.11	5.82

Table 6: Matriz de covarianzas de países desarrollados.

En ambos casos la exportación de tecnología está asociado positivamente a una tasa de crecimiento, lo que sugiere que la industria de alto valor agregado es un motor de crecimiento. Para enriquecer nuestro análisis observamos los gráficos de las figuras 5, 6, 7 y 8. Notamos que las tasas de crecimiento no son muy distintas, excepto por el hecho de que en los países desarrollados es seis veces menos variable. Lo que es interesante respecto de esta variable es cómo se relaciona con las demás acorde al tipo de país.

6 Ejercicio 5

Hicimos un test de verosimilitud para corroborar que la diferencia entre matrices de covarianza por grupo es significativamente distinta. Planteamos H0 = "las matrices de covarianza son iguales" contra H1 = "las matrices de covarianza son distintas". El p-valor es de $1.3 * e^{-6}$, con lo que tenemos evidencia suficiente para rechazar la hipótesis nula y afirmar que las matrices son significativamente distintas para países desarrollados y subdesarrollados.

7 Ejercicio 6

La descomposición espectral de la matriz arrojó el siguiente vector de valores propios:

$$D = (810.3787144, 100.5699330, 58.3808456, 17.0778631, 3.2166448, 1.5382517, 0.1624454)$$

	expbser	tcrec	expaltec	impbser	partVAind	Crecpob	Inv
expbser	475.92	25.20	101.07	486.04	33.24	-10.13	59.04
tcrec	25.20	6.13	12.00	27.01	2.88	-0.44	9.10
expaltec	101.07	12.00	86.02	93.50	-2.96	-5.12	11.09
impbser	486.04	27.01	93.50	526.20	1.38	-7.10	52.86
partVAind	33.24	2.88	-2.96	1.38	85.91	-3.12	41.02
Crecpob	-10.13	-0.44	-5.12	-7.10	-3.12	0.74	-1.57
Inv	59.04	9.10	11.09	52.86	41.02	-1.57	33.49

Table 7: Matriz de covarianzas de países subdesarrollados.

	expbser	tcrec	expaltec	impbser	partVAind	Crecpob	Inv
expbser	1.00	0.35	-0.46	0.98	0.09	-0.30	-0.22
tcrec	0.35	1.00	0.12	0.27	0.43	-0.15	-0.13
expaltec	-0.46	0.12	1.00	-0.49	-0.02	0.06	-0.30
impbser	0.98	0.27	-0.49	1.00	-0.01	-0.24	-0.20
partVAind	0.09	0.43	-0.02	-0.01	1.00	-0.63	0.28
Crecpob	-0.30	-0.15	0.06	-0.24	-0.63	1.00	0.12
Inv	-0.22	-0.13	-0.30	-0.20	0.28	0.12	1.00

Table 8: Matriz de correlaciones de países desarrollados.

8 Ejercicio 7

Para estandarizar la matriz las matrices D y U involucradas en la descomposición espectral de S, calculadas con el comando `svd`. La ecuación es

$$X_1 = XUD^{-1/2}U'$$

La matriz resultante se muestra en la tabla 10.

	expbser	tcrec	expaltec	impbser	partVAind	Crecpob	Inv
expbser	1.00	0.47	0.50	0.97	0.16	-0.54	0.47
tcrec	0.47	1.00	0.52	0.48	0.13	-0.21	0.64
expaltec	0.50	0.52	1.00	0.44	-0.03	-0.64	0.21
impbser	0.97	0.48	0.44	1.00	0.01	-0.36	0.40
partVAind	0.16	0.13	-0.03	0.01	1.00	-0.39	0.76
Crecpob	-0.54	-0.21	-0.64	-0.36	-0.39	1.00	-0.32
Inv	0.47	0.64	0.21	0.40	0.76	-0.32	1.00

Table 9: Matriz de correlaciones de países subdesarrollados.

	1	2	3	4	5	6	7
1	-0.89	-2.83	1.44	1.64	4.41	4.25	2.09
2	0.15	-2.39	1.99	1.33	3.16	5.64	4.15
3	2.52	-1.62	1.10	3.30	3.58	5.25	3.08
4	-0.77	-0.32	2.36	1.38	3.90	5.05	3.41
5	1.36	-0.50	2.14	1.53	3.82	6.37	3.08
6	0.13	-0.03	0.71	1.81	5.06	5.03	2.70
7	-0.38	0.62	2.38	1.75	6.77	5.15	5.44
8	1.31	-0.76	3.25	5.20	3.91	3.97	4.10
9	-1.65	0.62	2.19	3.12	1.55	4.98	2.63
10	1.17	0.42	3.00	1.30	4.99	5.71	2.30
11	0.19	-1.07	2.81	1.57	3.28	4.73	3.64
12	-0.05	-1.56	2.17	2.25	4.30	3.73	3.26
13	-0.86	-0.77	1.35	2.54	3.04	6.64	2.63
14	-1.04	-1.76	0.66	2.87	5.83	6.36	2.26
15	-0.99	-1.55	0.86	4.32	4.42	6.34	5.08
16	0.04	-0.57	0.90	0.81	3.16	5.35	4.20
17	0.03	-0.73	1.21	1.78	4.06	2.62	2.85
18	-0.65	-2.17	3.39	1.09	3.97	4.29	5.11
19	-0.07	0.38	2.69	2.10	3.82	6.12	3.68
20	-0.37	-0.93	0.79	1.47	3.84	4.70	3.10
21	-0.01	-0.98	1.14	2.03	3.78	3.70	4.46
22	0.99	-0.21	2.47	1.87	4.33	4.87	2.08
23	-1.13	-1.13	3.76	3.01	4.95	4.35	1.52
24	-1.22	-1.31	4.01	1.95	3.56	6.10	3.54
25	-0.58	1.29	0.86	2.34	3.30	4.06	3.90
26	1.42	-1.60	2.21	1.50	3.24	5.83	3.67

Table 10: Matriz de datos estandarizada.

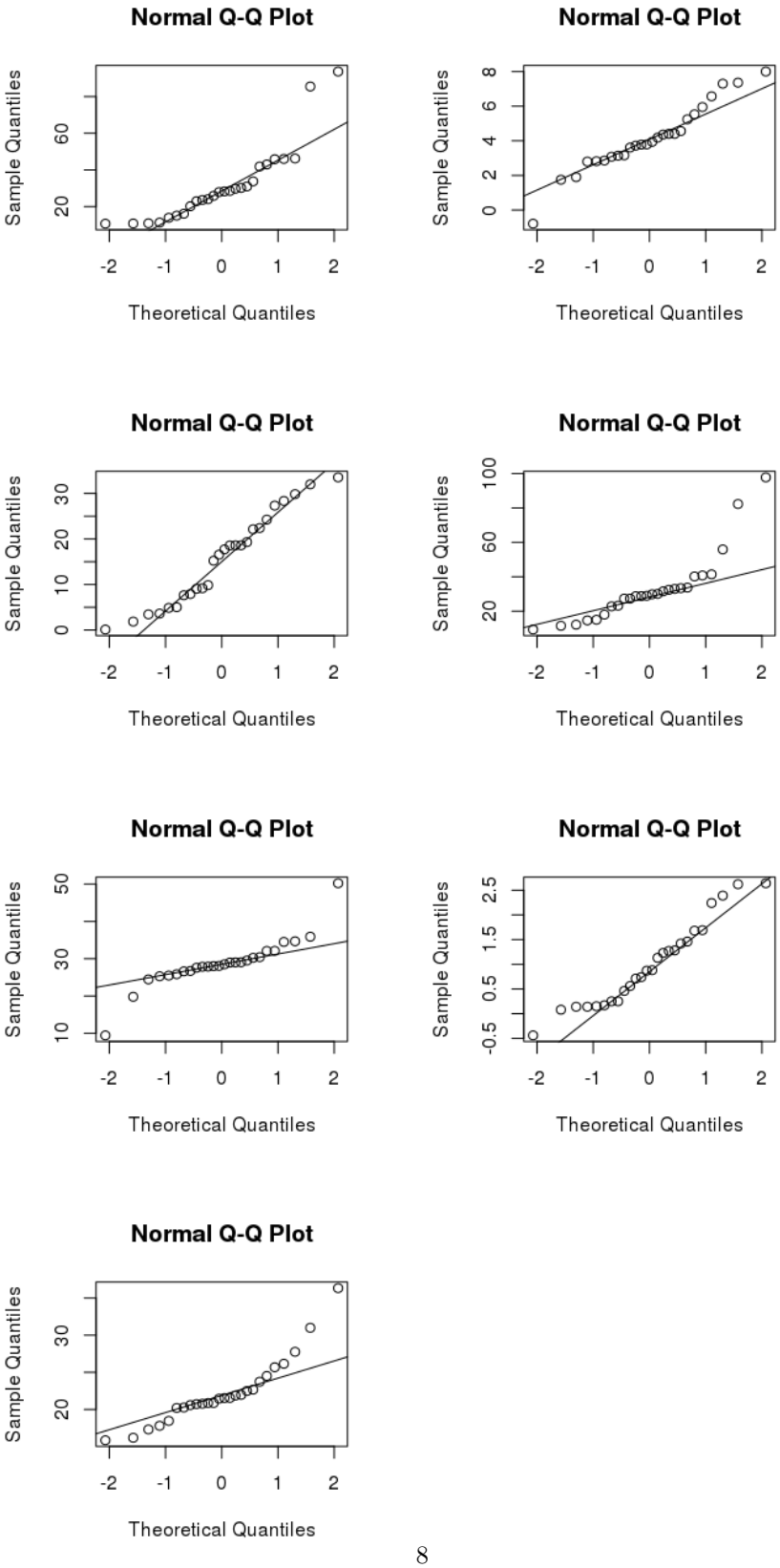

Figure 4: QQ-plots para todas las variables. De izquierda a derecha: exportaciones de bs y serv., tasa de crecimiento, exportación de tecnología, importaciones, participación del VA, crecimiento poblacional, e inversiones.

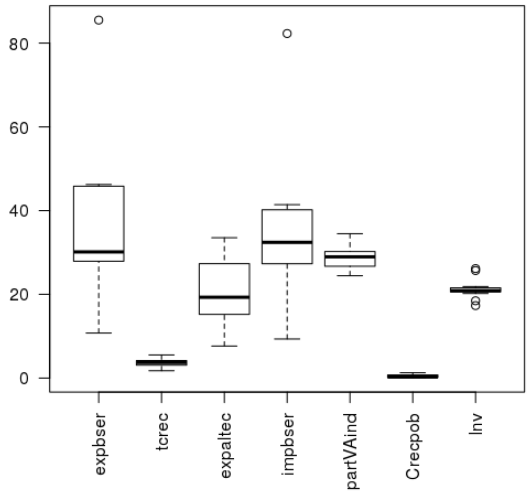

Figure 5: qq-plot de los países desarrollados.

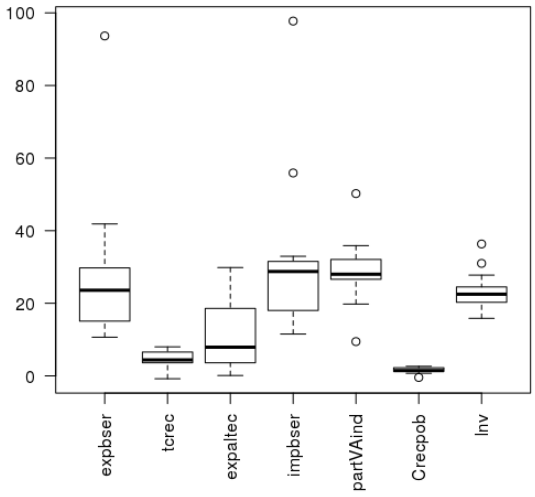

Figure 6: qq-plot de los países subdesarrollados.

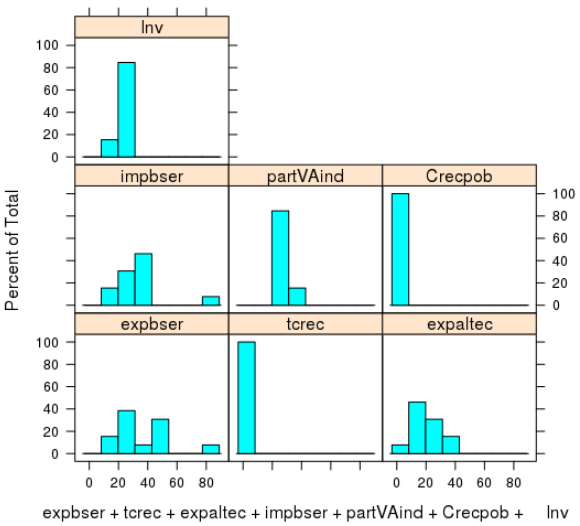

Figure 7: Histogramas de los países desarrollados.

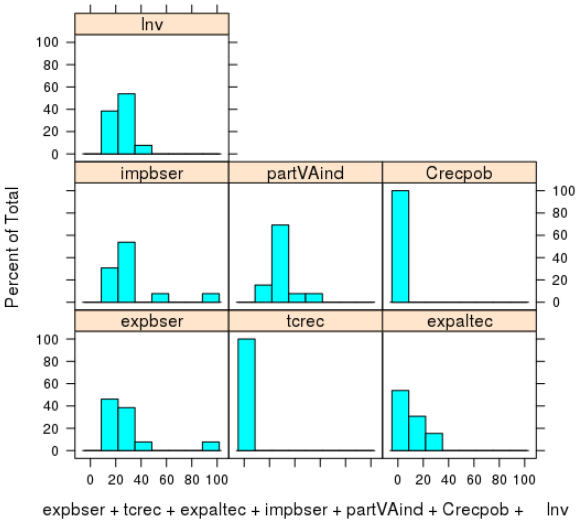

Figure 8: Histogramas de los países subdesarrollados.