Desempeño educativo de los estudiantes de la Licenciatura en Economía

Daniel Czarnievicz & Mauro Gopar & Romina Quagliotti

2018

Índice

Abstract

En el presente trabajo se buscó estudiar la relación entre el desempeño educativo de los estudiantes de la Licenciatura en Economía y sus características sociodemográficas mediante la aplicación de técnicas de análisis factorial. Para ello se utilizaron datos provenientes del Sistema de Gestión de Bedelías y el Formulario Estadístico de ingreso a la Universidad correspondientes a las generaciones de ingreso 2012 y 2013. Se encontró una asociación entre que el máximo nivel educativo alcanzado por el padre y la madre sea bajo, que el estudiante haya cursado sexto año en un liceo público, y que lo haya hecho en el interior del país. No se halló una asociación entre la situación laboral del estudiante al momento de ingreso a la facultad y su desempeño en sus primeros cuatro años.

Introducción

En este trabajo se utilizan los mismos datos que en el anterior, pero se aplican otras técnicas de análisis multivariado para continuar con el estudio del desempeño y las características sociodemográficas de los estudiantes de la Licenciatura en Economía. Se realiza un análisis de componentes principales con las variables asociadas al desempeño de los estudiantes, y a partir de este se construye un índice de rendimiento. Luego, mediante análisis de correspondencia múltiple, se estudian las asociaciones entre el puntaje obtenido en este índice y las características sociodemográficas de los estudiantes.

El trabajo se estructura de la siguiente forma. En primer lugar, se explican las metodologías utilizadas. Luego, se detallan los resultados obtenidos. Finalmente, se desarrollan algunas conclusiones. Para obtener información acerca de la base de datos utilizada y un análisis descriptivo de la misma se puede recurrir al primer trabajo.

Metodología

Las técnicas que se utilizan corresponden a las englobadas dentro del análisis factorial. Consisten en el estudio de los datos, buscando encontrar factores que simplifiquen la visualización y la interpretación, y procurando capturar la mayor cantidad de información de la tabla de datos original posible. Más precisamente, se aplicarán técnicas de Análisis de Componentes Principales y Análisis de Correspondencia simple y múltiple.

El Análisis de Componentes Principales (ACP) es una técnica que tiene por objetivo encontrar similitudes entre individuos y variables de una tabla de datos. Con esta información, se pretende construir nuevos ejes donde la información contenida en la tabla original se traslade lo mejor posible a la proyección de los datos en ellos. Se realiza sobre matrices que contienen individuos en las filas y variables cuantitativas en las columnas.

Se realizan dos estudios, uno para la nube de las filas (individuos) y otro para la nube de columnas (variables), aunque están muy relacionados.

En lo que refiere a la nube de filas, se puede proyectar cada observación en un nuevo eje, multiplicando el total de la matriz de datos $\mathbf{X}$ por el vector $\mathbf{u}$, que define a aquél, ponderando por la influencia de cada variable en el cálculo de distancias entre filas (la matriz diagonal $\mathbf{M}$, usualmente igual a la matriz identidad $\mathbf{I}$). Haciendo el producto XMu_s, se obtiene un vector de dimensión $Nx1$, que determina las proyecciones de los N individuos en el nuevo eje, llamado $F_s, s = 1, 2, J$. La forma de elegir el vector u_s es pensando en maximizar la inercia proyectada sobre el mismo, definida como $\sum_{i=1}^{N} p_i[F_s(i)]^2$, donde p_i representa el peso de cada individuo, por lo general $1/N$.

Normalizando la norma de los vectores u_s a 1, llamando $\mathbf{D}$ a la matriz diagonal con los pesos p_i y considerando $M = I$ se tiene que maximizar la inercia proyectada en el eje definido por u_s es igual a maximizar $u_s' MX'DXMu_s$ sujeto a $u'u = 1$. Desarrollando se tiene que los ejes que resuelven el problema son los definidos por los vectores propios asociados a los valores propios λ_s, que verifican $X'DXu_s = \lambda_s u_s$, de donde también se desprende que que F_s es vector asociado a $XMX'D$ (multiplicando ambos lados por XM). Además, la inercia proyectada sobre el primer eje corresponderá al valor propio más alto y la suma de los valores propios, será la inercia de todos los ejes, es decir, la inercia total. Con tablas de datos estandarizadas, se corresponderá con el número de variables, ya que la inercia de la tabla es igual a la traza de la matriz de varianzas y covarianzas de las variables (todas iguales a 1 en el caso estandarizado).

Cuando se trabaja con la nube de columnas, se sigue un procedimiento similar, considerando que los pesos asociados a las observaciones en el análisis por filas, ahora se corresponden con la métrica a utilizar (la matriz $\mathbf{D}$, en general, diagonal que toma siempre el valor $1/N$) y la métrica del análisis anterior es el peso de cada variable (la matriz $\mathbf{M}$, en general la identidad $\mathbf{I}$), por lo que el problema consiste en maximizar $v_s' DXMX'Dv_s = \mu_s v_s$. El resultado es equivalente al anterior, por lo que los ejes que maximizarán la inercia proyectada de las variables serán aquéllos definidos por los vectores v_s asociados a los valores propios μ_s, que verifican $XMX'v_s = \mu_s v_s$.

Con esto se observa que F_s, es colineal con v_s y los valores μ_s y λ_s coinciden hasta que $s = J$, a partir del cual los valores propios son nulos.

Haciendo el producto $X'Dv_s$ se obtiene el vector de factores G_s asociado a las columnas.

Aplicando esta técnica para las filas y las columnas de la matriz de datos cuantitativos, se obtienen los nuevos ejes, donde cada uno captura una inercia igual al valor propio que lo define y son ortogonales entre sí. Dependiendo de la calidad de representación de cada variable y cada observación, se elige la cantidad de ejes.

En cuanto al análisis de correspondencia simple y múltiple, consiste en aplicar técnicas similares a tablas de contingencia o disjuntas completas (según el caso). Se procede creando los perfiles fila o columna de la tabla, calculando las frecuencias relativas de las modalidades de una variable dada la otra, y se calcula la distancia entre filas o columnas usando la distancia χ^2 definida como

$$d^2_{\chi^2}(i,l) = \sum_{j=1}^{J} \frac{1}{f_j} \left(\frac{f_{ij}}{f_{i.}} - \frac{f_{lj}}{f_{l.}} \right)^2$$

donde f_{ij} corresponde a la frecuencia relativa de fila i y columna j con respecto al total de la fila, $f_{i.}$ es la frecuencia relativa de la fila i en el total, y f_j es el peso de la columna j en el total, por lo que se considera en mayor medida las distancias entre filas en modalidades raras, es decir, cuando la frecuencia relativa de la j es muy baja. La suma de f_{ij} para cada i es 1, por lo que la nube de las filas está en un hiperplano.

Al hacerse el análisis de correspondencia simple sobre una tabla de contingencia, las filas y las columnas corresponden ambas a modalidades y el desarrollo es el mismo.

Para proyectar la nube de las filas (o las columnas), se asigna un peso a cada una según la frecuencia marginal y la inercia de cada fila es la distancia χ^2 multiplicada por $f_{i.}$. La inercia total es la suma de las inercias y es χ^2/n, siendo n el número de observaciones y χ^2 el estadístico que representa la distancia de las observaciones hacia observaciones teóricas independientes:

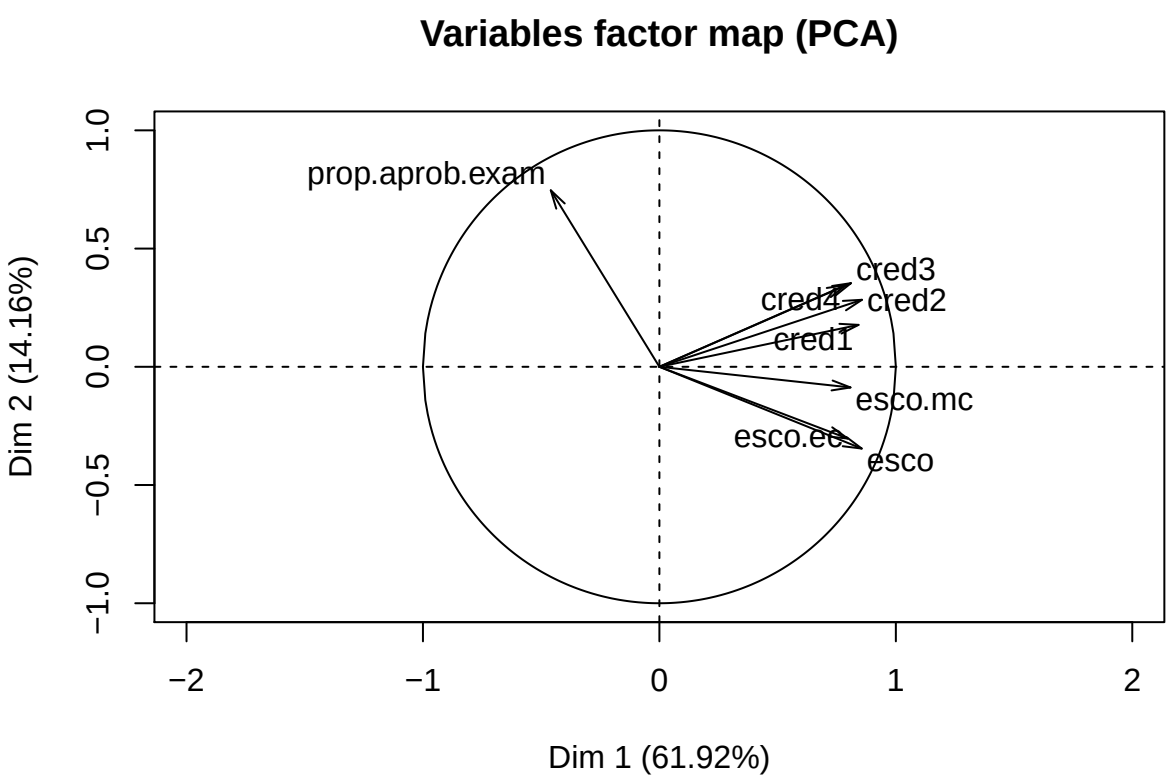
$$\chi^2 = \sum_{ij} \frac{(nf_{ij} - nf_{i.}f_j)^2}{nf_{i.}f_j}$$

Resultados obtenidos

Análisis de componentes principales

Para este análisis se utilizan las siguientes variables relativas al desempeño de los estudiantes: escolaridad, escolaridad en el área métodos cuantitativos, escolaridad en el área economía, créditos aprobados en cada uno de los cuatro años, y proporción de aprobaciones rendidas por examen.

En primer lugar, se realiza un análisis de componentes principales con todas las variables mencionadas anteriormente, con el objetivo de estudiar la posibilidad de reducir las dimensiones del problema.



Cuadro 1: Eigenvalues

	eigenvalue	percentage of variance	cumulative percentage of variance
comp 1	4.953	61.916	61.916
comp 2	1.133	14.162	76.078
comp 3	0.543	6.788	82.866
comp 4	0.393	4.907	87.773
comp 5	0.320	4.005	91.778
comp 6	0.290	3.629	95.407
comp 7	0.205	2.566	97.972
comp 8	0.162	2.028	100.000

El primer eje captura un 61,92 % de la varianza, el segundo un 14,16 %, y el tercero un 6,79 %. A su vez, se observa que los dos primeros componentes son los únicos que tienen valor propio mayor a 1, y que entre ambos recogen un 76.07 % de la varianza total.

Cuadro 2: Contribuciones

	Dim.1	Dim.2
esco	14.787	10.554
esco.ec	12.816	8.127
esco.mc	13.189	0.675
cred1	14.324	2.774
cred2	14.824	7.089
cred3	13.259	11.052
cred4	12.540	10.541
prop.aprob.exam	4.261	49.188

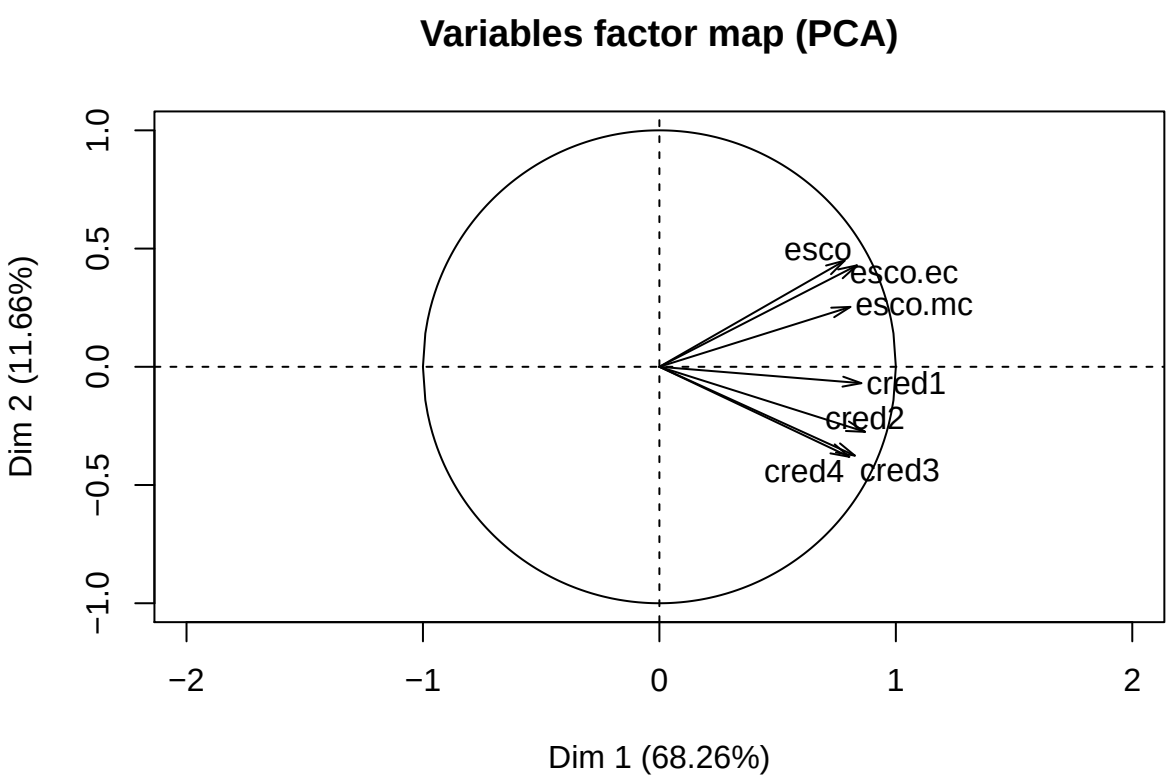
Cuadro 3: Cosenos cuadrados

	Dim.1	Dim.2
esco	0.732	0.120
esco.ec	0.635	0.092
esco.mc	0.653	0.008
cred1	0.710	0.031
cred2	0.734	0.080
cred3	0.657	0.125
cred4	0.621	0.119
prop.aprob.exam	0.211	0.557

Se observa que las variables correspondientes a las escolaridades y los créditos aprobados en cada año contribuyen en forma similar al primer componente. Por otro lado, la variable que considera la proporción de aprobaciones rendidas por examen contribuye poco al primer componente, pero en un 49,18 % de la inercia del segundo componente.

Asimismo, todas las variables ven representadas al menos un 75 % de su variabilidad entre los dos primeros componentes. Por lo tanto, en base a esta información, se encuentra que se podrían reducir las dimensiones del problema de ocho a dos.

En segundo lugar, se realiza un ACP con el objetivo de construir un índice de rendimiento de los estudiantes. Para esto, se utilizan las variables anteriores con excepción de la proporción de aprobaciones rendidas por examen, debido a que esta no indica acerca del rendimiento per se del estudiante, sino respecto a cómo el mismo avanzó.



Cuadro 4: Eigenvalues

	eigenvalue	percentage of variance	cumulative percentage of variance
comp 1	4.778	68.264	68.264
comp 2	0.816	11.658	79.923
comp 3	0.399	5.699	85.621
comp 4	0.335	4.782	90.403
comp 5	0.291	4.163	94.567
comp 6	0.206	2.950	97.516
comp 7	0.174	2.484	100.000

Cuadro 5: Cosenos cuadrados

	Dim.1	Dim.2	Dim.3	Dim.4	Dim.5
esco	0.698	0.184	0.000	0.017	0.002
esco.ec	0.614	0.202	0.134	0.001	0.000
esco.mc	0.652	0.064	0.257	0.003	0.001
cred1	0.730	0.005	0.001	0.223	0.023
cred2	0.757	0.076	0.000	0.009	0.033
cred3	0.683	0.141	0.002	0.015	0.094
cred4	0.645	0.145	0.004	0.067	0.139

El primer componente recoge un 68,26 % de la variabilidad y es el único con valor propio mayor a 1. A su vez,

se considera que todas la variables se ven representadas en forma satisfactoria en el primer componente, ya que la que presenta un menor coseno cuadrado es escolaridad de economía, con un 61,4 %. Por lo tanto, se considera que es correcta la construcción de un índice en base a estas variables utilizando el primer componente. Como ponderadores del índice se utilizan las contribuciones de cada variable al primer componente.

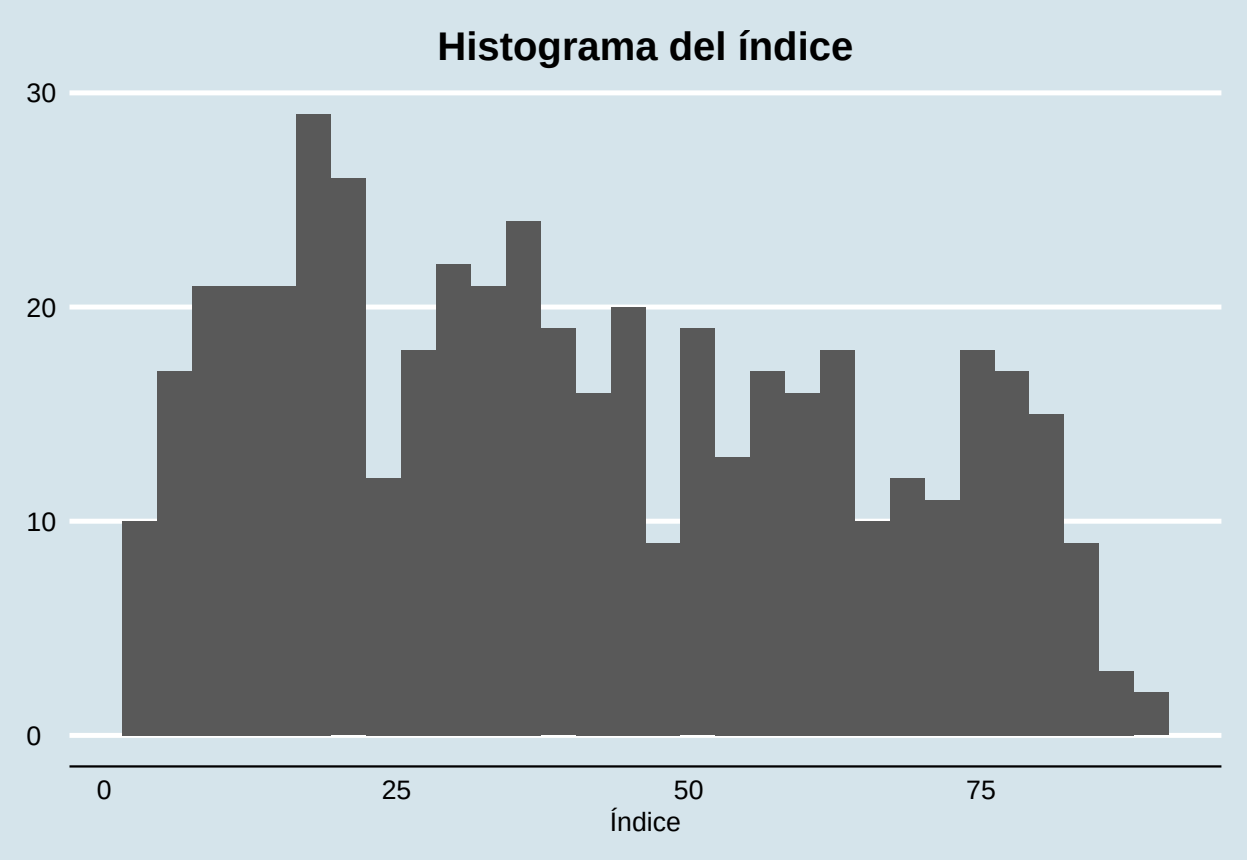
Cuadro 6: Contribuciones

esco	14.600
esco.ec	12.842
esco.mc	13.654
cred1	15.268
cred2	15.835
cred3	14.299
cred4	13.501

El índice se construyó de la siguiente forma:

$$indice_i = \sum_{j \in esco} \beta_j \frac{x_{ij}}{12} + \sum_{j \in creds} \gamma_j \frac{x_{ij}}{\text{máx}(90;\, x_{ij})}$$

donde x_{ij} es el valor de la variable x_j para el individuo i, β_j y γ_j son las contribuciones de dichas variables a la primer dimensión del PCA, y *esco* y *creds* representan el espacio de variables referentes a la escolaridad y los cráditos obtenidos, respectivamente. Por lo tanto, a cada individuo le corresponde un puntaje del índice entre 0 y 100, de acuerdo a sus valores de las variables escolaridades y créditos aprobados en cada año. Los divisores fueron escogidos de acuerdo a los máximos teóricos de cada variable.

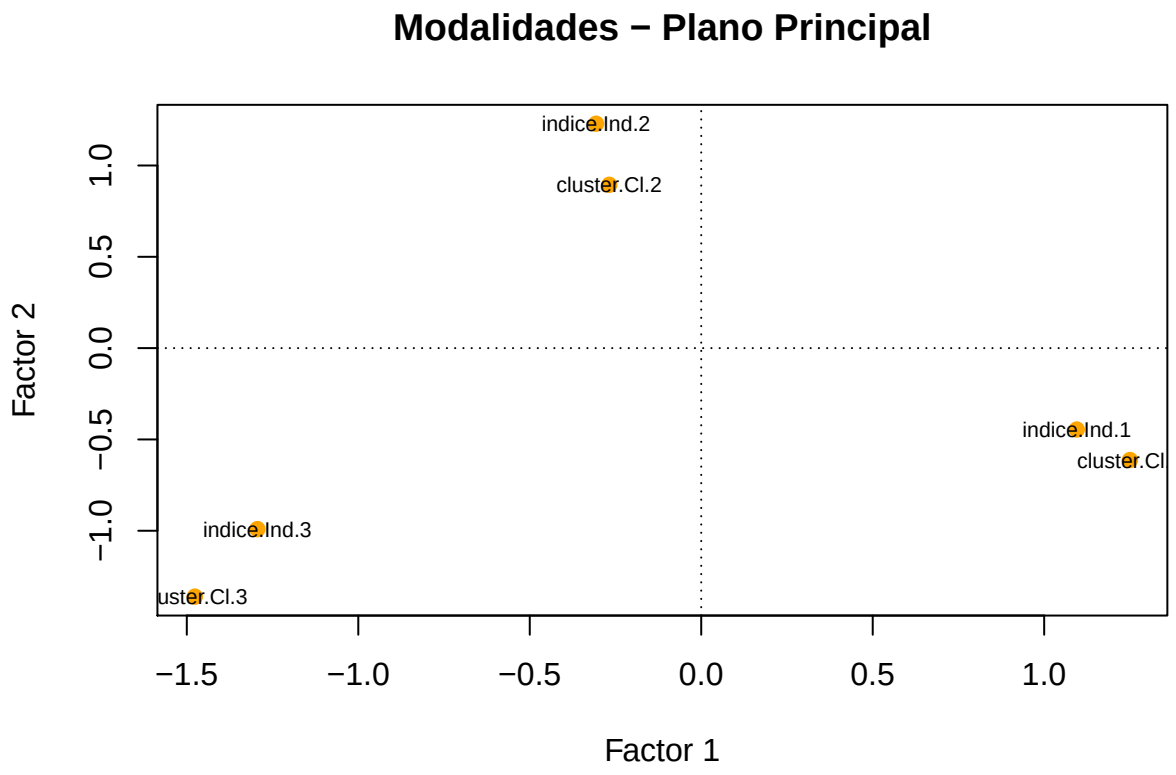


Análisis de correspondencias

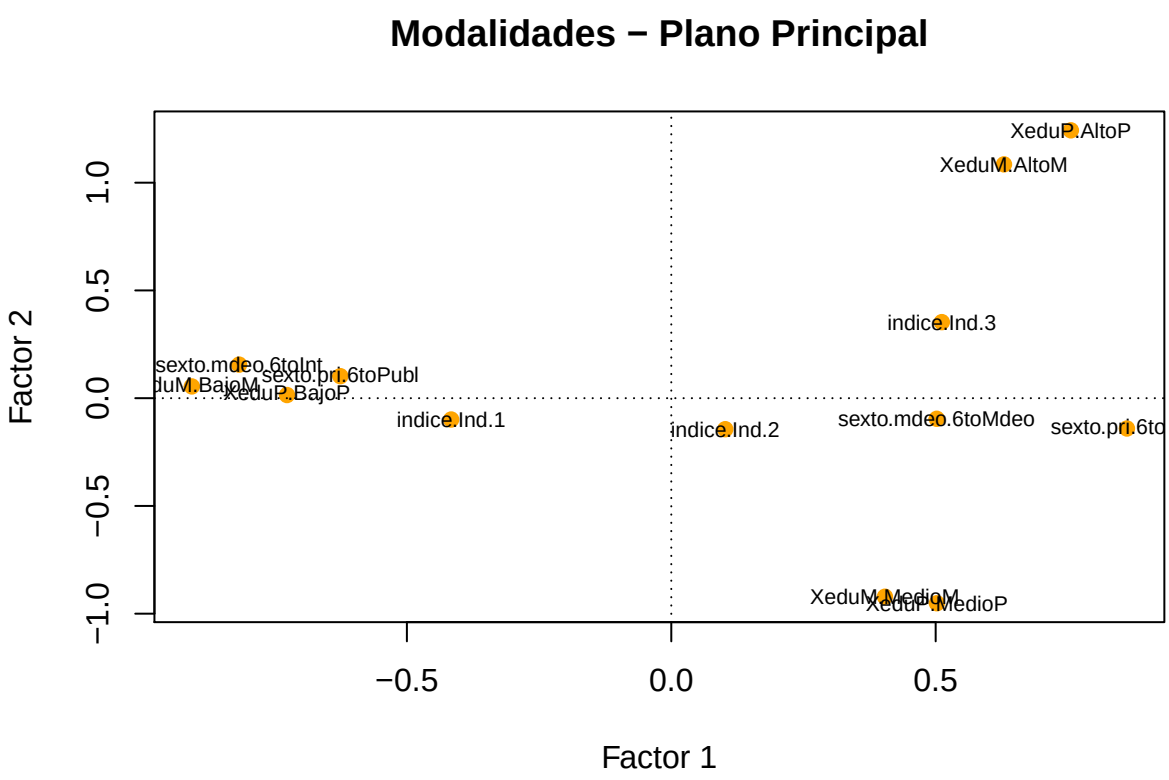
En esta parte, el objetivo es estudiar si existen asociaciones entre los distintos valores de las variables sociodemográficas y el rendimiento del estudiante medido por el índice. Para esto se utilizan las técnicas de análisis de correspondencia. Las variables sociodemográficas incluidas en el análisis son: máximo nivel educativo alcanzado por la madre, máximo nivel educativo alcanzado por el padre, si el estudiante realizó sexto año en una institución pública o privada, si lo realizó en Montevideo o en el Interior, si trabajaba al momento de su ingreso a la facultad, si buscaba trabajo en ese momento, su sexo, y su edad.

Debido a que la variable índice es cuantitativa continua, se discretizó utilizando un tercio del rango como umbrales para cada grupo. A los estudiantes que obtuvieron un puntaje en el índice de hasta 31 se les asignó la categoría "1", que se corresponde al rendimiento bajo. A los que obtuvieron entre 32 y 65 se les asignó la "2", que se corresponde con el rendimiento medio. Finalmente, los que obtuvieron más de 60 puntos fueron asignados a la categoría "3" de rendimiento alto. La categoría 1 quedó compuesta por 188 estudiantes, la 2 por 164, y la 3 por 120.

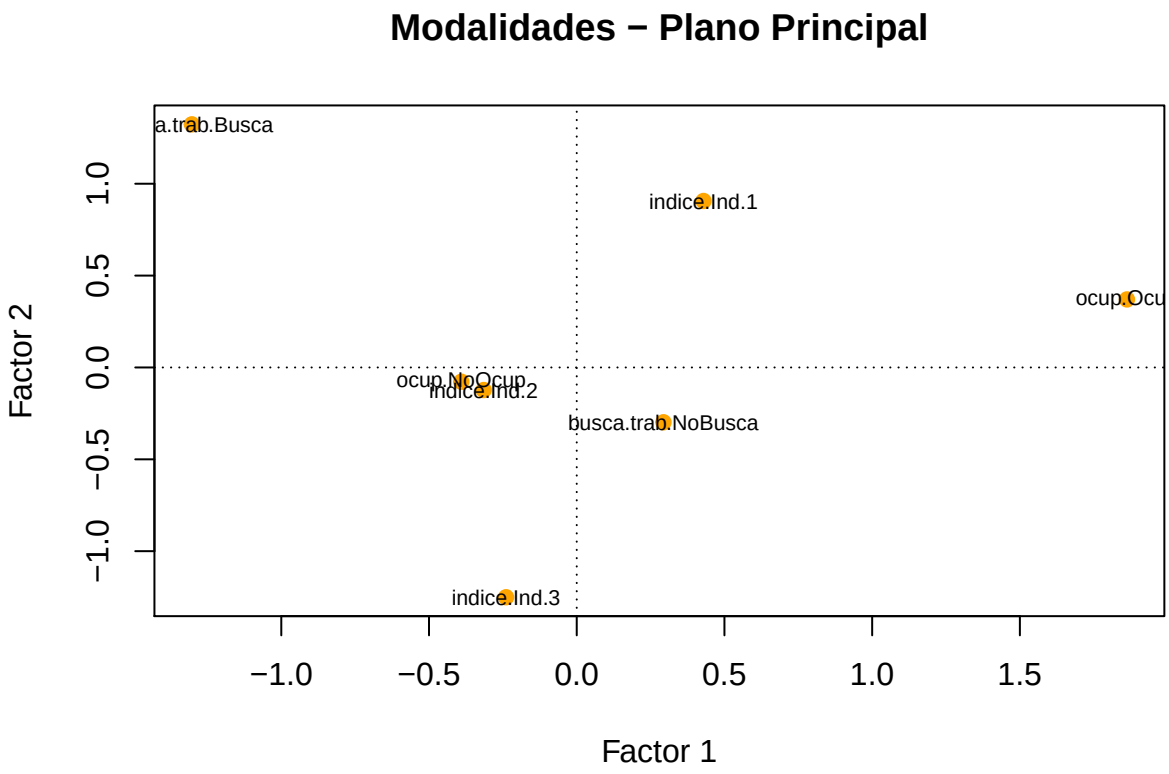
En primer lugar, se realiza un análisis de correspondencia simple entre las modalidades del índice y los clusters hallados en el trabajo anterior, con el objetivo de validar o no la categorización construida previamente. Se encuentra una fuerte asociación entre las modalidades de ambas variables, por lo que las categorizaciones halladas por ambos métodos son coincidentes. Se puede ver el resultado gráfico de esto en la siguiente imagen.



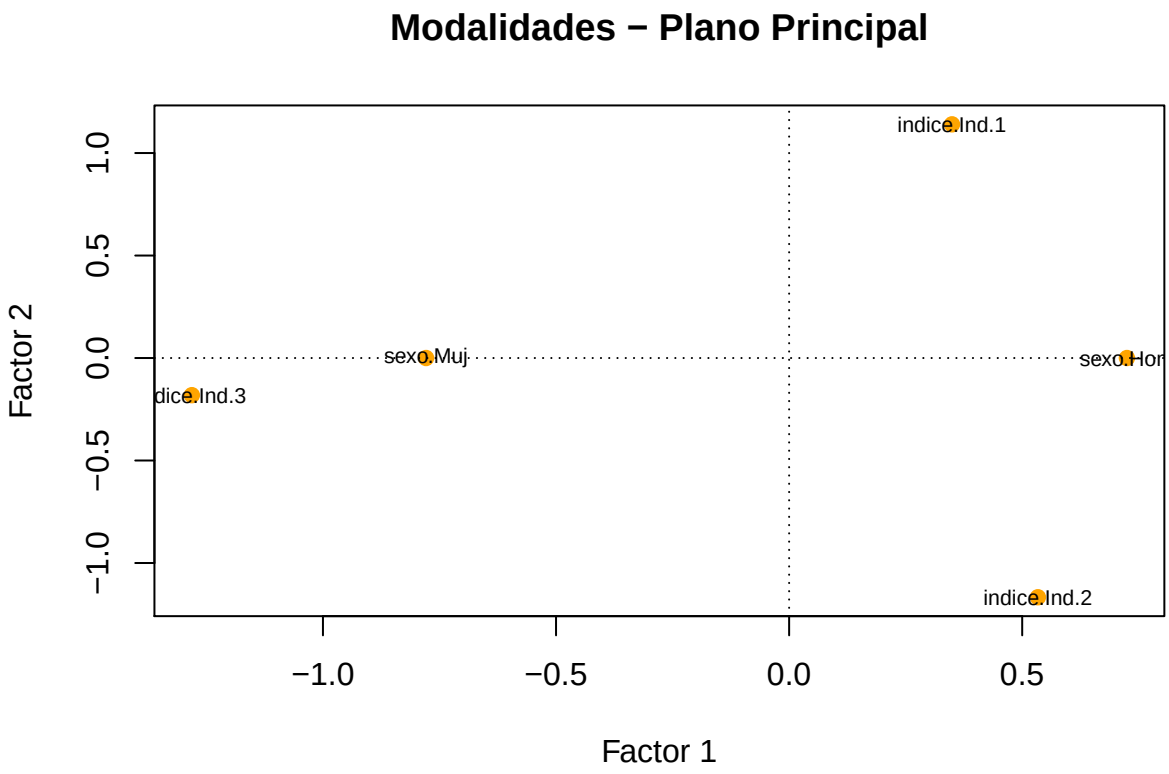
En segundo lugar, se analiza la asociación entre las modalidades del índice y de variables que indican características sociodemográficas de los estudiantes. A continuación se presentan los resultados hallados que se consideraron relevantes.



En primer lugar, se observa una clara asociación entre las modalidades de las variables máximo nivel educativo alcanzado por el padre y la análoga para la madre. A su vez, se encuentra un vínculo entre haber realizado sexto año en una institución pública, haberlo realizado en interior del país, y que el máximo nivel educativo alcanzado por el padre y la madre sea bajo. También, aunque en menor medida, las características anteriores parecen estar asociadas a obtener un puntaje bajo en el índice de rendimiento.



Al realizar un ACM con las modalidades del índice y de las variables correspondientes a la situación laboral del estudiante al momento de su ingreso a la facultad (si trabajaba o buscaba trabajo), no se encuentra ninguna asociación entre estas. Esto está en línea con una de las hipótesis presentadas al inicio del primer trabajo, la cual sostenía que no se esperaba encontrar una relación entre estas variables debido a que el desempeño considerado era en los primeros cuatro años de la carrera, mientras que las variables de índole laboral relevaban la situación del estudiante únicamente al ingreso.



Finalmente, no es claro que haya una asociación entre las categorías de las variables correspondientes al sexo del estudiante y su rendimiento.

Conclusiones

En primer lugar, en este trabajo se pudo validar la categorización construida en la primera entrega mediante el análisis de cluster, realizando un análisis de componentes principales.

En segundo lugar, respecto a las hipótesis planteadas en la primera entrega, se encuentran los siguientes resultados. Se halla una asociación entre que el máximo nivel educativo alcanzado por ambos padres sea bajo y que el estudiante tenga un mal desempeño. Sin embargo, no es posible arribar a una conclusión acerca de una asociación entre desempeño del estudiante y máximo nivel educativo alcanzado por sus padres para niveles medios y altos de ambas variables.

A su vez, se encuentra una asociación entre que el estudiante haya cursado sexto año en una institución pública, que lo haya hecho en el interior, y que su madre y padre tengan un nivel educativo bajo.

Por último, de acuerdo con lo planteado en las hipótesis, no se encuentran asociaciones entre las modalidades de las variables correspondientes a la situación laboral del estudiante al momento de su ingreso a la facultad y su desempeño en los primeros cuatro años.

Anexos

Cuadro 7: Estudiantes según clusters y categorías del índice

	Ind 1	Ind 2	Ind 3
Cl 1	151	1	0
Cl 2	37	163	35
Cl 3	0	1	85

Referencias

- Hadley Wickham (2017). tidyverse: Easily Install and Load the 'Tidyverse'. R package version 1.2.1. https://CRAN.R-project.org/package=tidyverse

- R Core Team (2017). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. URL https://www.R-project.org

- Sebastien Le, Julie Josse, Francois Husson (2008). FactoMineR: An R Package for Multivariate Analysis. Journal of Statistical Software, 25(1), 1-18. 10.18637/jss.v025.i01

- Jorge Blanco. 2006. «Introducción al Análisis Multivariado.» Instituto de Estadística. Universidad de la República.