

# **Caracterización de la pobreza en Chile mediante las variables indicadoras de la encuesta Casen 2017**

**EYP2435 - Análisis Multivariado**

Sebastián Celaya

Camila Echeverría

## **Introducción**

### **Descripción de los datos**

Para el desarrollo de nuestro análisis, utilizaremos las distintas variables disponibles en la categoría de índices de la encuesta socioeconómica Casen 2017, cuyos datos se pueden encontrar en [este link](#). Estas corresponden a distintos valores que buscan resumir y englobar las distintas características que resultan importantes para representar la situación socioeconómica de las personas y hogares chilenos.

Así, luego de guardar nuestros datos en el archivo `DatosP2.RData`, las variables con las que trabajaremos son las siguientes:

- `region`: el identificador de cada región.
- `comuna`: sirve para identificar a qué comuna del país pertenece el hogar.
- `pobreza`: corresponde a la situación de pobreza por ingresos. Puede ser extremo (1), no extremo (2) o no pobre(0).
- `numer`: número de personas (sin contar personal doméstico) que viven en el hogar.
- `esc`: corresponde al nivel de escolaridad alcanzado.
- `educ`: máximo nivel educacional alcanzado o nivel educacional actual.
- `depen`: dependencia del establecimiento educacional.
- `activ`: condición de actividad de la persona. Puede ser ocupado, desocupado o inactivo.
- `indmat`: indicador de materialidad de la vivienda.
- `indsan`: indicador de saneamiento en el hogar.
- `calglobviv`: calidad global de la vivienda.
- `iae`: el hogar presenta allegamiento externo.
- `iai`: el hogar presenta allegamiento interno.
- `hacinamiento`: existe hacinamiento en la vivienda.
- `ypchautcor`: ingrero individual por persona del hogar corregido.

Las variables `NA` presentes en la base de datos se dan por la manera en que fue diseñada la encuesta. De esta manera, representan la no respuesta o las respuestas en blanco. Es por esto que, en la mayoría de los casos, fueron reemplazadas por el valor 0 o algo similar para indicar que, en realidad, la pregunta no aplica a la persona en cuestión.

## **Metodologías**

Para poder corroborar o descartar nuestra hipótesis, es importante que, en primer lugar, realicemos un análisis exploratorio de nuestros datos. Como en su mayoría contamos con variables categóricas, veremos cómo se comportan éstas en términos de proporción poblacional en cada una de las regiones del país, para ver si resulta más conveniente trabajar con la variable `region` o eliminarla por completo de nuestro análisis.

Para poder tener una muestra más balanceada, realizaremos un muestreo aleatorio simple donde seleccionaremos a 20 mil personas en situación de pobreza (ya sea extrema o no) y 20 mil personas en situación de no pobreza.

Luego de esto, llevaremos a cabo un análisis factorial para reducir la dimensionalidad de nuestras variables y determinar cuántos factores necesitaremos para explicar entre un 70% y un 90% de la varianza total. Así, utilizaremos estos factores para modelar una regresión logística que sea capaz de caracterizar a la gente como no pobre (0) o pobre(1). Analizaremos los distintos estadísticos y elementos de esta regresión para determinar qué tan buen modelo es y si cumple con la suficiencia necesaria.

Finalmente, y con estos mismos factores, intentaremos llevar a cabo una regresión lineal donde la variable a predecir sea el ingreso autónomo per cápita del hogar corregido, donde calcularemos los distintos valores que describen la calidad de ajuste del modelo.

## **Resultados**

En el análisis exploratorio regional obtendremos los siguientes resultados:

### **Análisis de pobreza por Región:**

Como se puede observar en Figura 1a la mayor porcentaje de pobreza extrema es la región de la Araucanía (IX Región) seguida de la región de Ñuble (XVI Región). Además podemos observar que las regiones de Aysén (XI), Magallanes (XII), Metropolitana (XIII) y Antofagasta (II) poseen la menor porcentaje de pobres extremos.

Ahora analizando Figura 1b se muestra aquellas regiones con pobreza no extrema, y observamos que son las mismas mencionadas anteriormente, al igual que las regiones que poseen menor porcentaje de pobreza no extrema.

Después de observar las figuras anteriores, al analizar Figura 1c es ocurre lo que era de esperar sobre las regiones con mayor porcentaje de no pobreza, aquellas regiones que obtuvieron una baja porcentaje de pobreza, valga la redundancia, tienen alta porcentaje de no pobreza y viceversa.

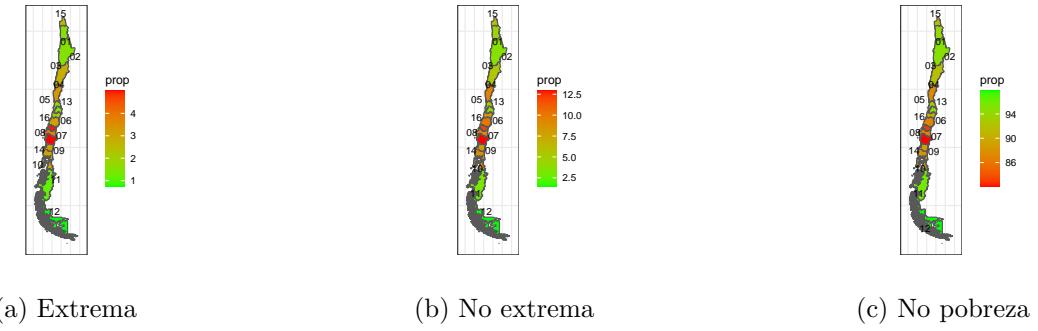


Figura 1: Pobreza por regiones

## Análisis de índice de actividad por Región:

De Figura 2a podemos observar que las regiones de Aysén (XI), Magallanes (XII) y Metropolitana (XIII) tiene mayor porcentaje de personas ocupadas, es decir, personas que durante la semana de referencia de la encuesta realizaron una actividad a cambio de remuneración o beneficios. De Figura 2b logramos ver que la región Arica y Parinacota (XV), Coquimbo (IV) y Biobio (VIII) presentan mayor porcentaje de desocupados, es decir, personas que no son ocupadas, pero durante las 4 últimas semanas han estado en búsqueda de un puesto de trabajo. Por último de Figura 2c podemos observar que la región de Coquimbo (IV) y Biobio (VII) muestran mayor proporción de personas inactivas, es decir, personas que no son ocupadas, ni desocupadas.

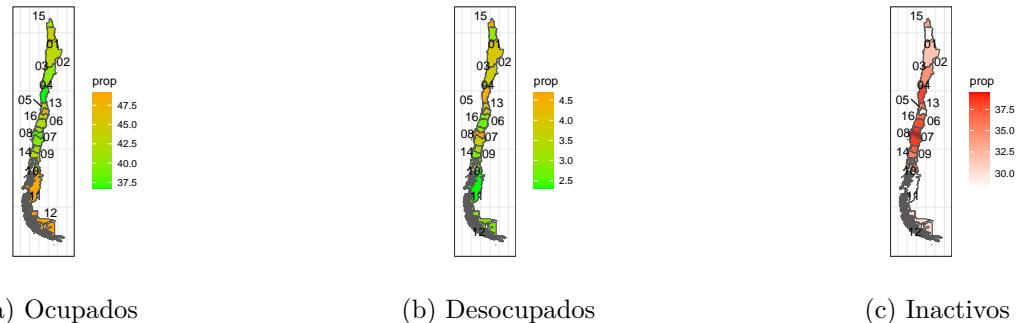
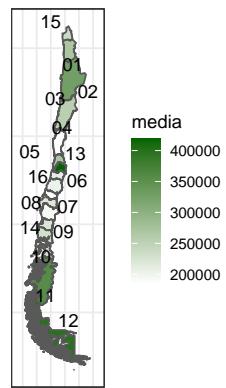


Figura 2: Índice de actividad por regiones

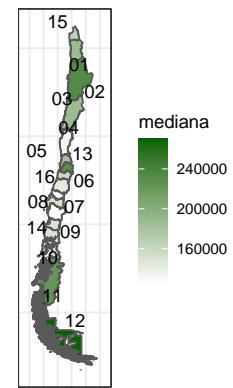
## Análisis de ingresos por Región:

Para este análisis usaremos 2 medidas de los ingresos, la media de estos y la mediana. Debido a que la media está sesgada por los valores extremos, en este caso las personas que tienen ingresos muy altos afectan al promedio, es por ello que también usaremos la mediana. De Figura 3a logramos divisar que la mayor media de los ingresos se encuentra principalmente en la región Metropolitana (XIII) con un ingreso promedio mayor de 400.000 pesos, seguida de la región de Magallanes (XII). Pero analizando la Figura 3b donde se presentan las medianas de los ingresos totales de las personas por región podemos ver que la región de Magallanes (XII) presenta mayor mediana de ingresos (más de 240.000 pesos), mientras que la región Metropolitana (XIII) tiene una mediana de ingresos entre 200.000 pesos y 240.000 pesos. Con

este caso de la región Metropolitana podemos analizar esta región para tener una visión de lo que pasa al interior de esta.



(a) Medios



(b) Medianos

Figura 3: Ingresos por regiones

Analizando lo que ocurre en la región metropolitana separando por comunas obtenemos las Figura 4a y Figura 4b. De Figura 4a observamos que la comuna de vitacura es la que presenta ingresos medios mayores a 1.250.000 pesos, por lo tanto esta comuna es la que incrementa la media de los ingresos en la región Metropolitana. Mientras que las demás comunas tienen un ingreso medio cercano a los 250.000 pesos.

Analizando la Figura 4b volvemos a observar que vitacura tiene la mayor mediana de los ingresos (más de 1.000.000 de pesos), después le siguen la comuna de Providencia y Las Condes con mediana de ingresos entre 750.000 y 1.000.000 pesos. Además, igual que en Figura 4a podemos apreciar que a excepción de estas comunas la mayoría tiene un ingreso (ya sea por media o mediana) cercano a 250.000 pesos. Por lo tanto es un aspecto a considerar que en santiago es un sector el cual presenta mayores ingresos, lo que afecta en la comparación entre los ingresos de las regiones.

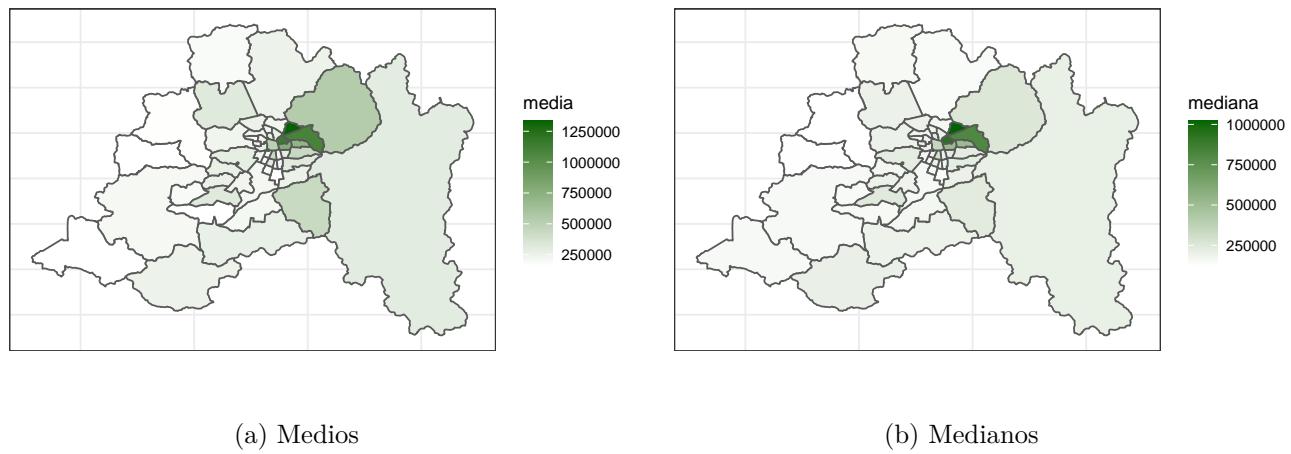


Figura 4: Ingresos por comunas de Santiago

## Análisis educacional por Región:

De Figura 5 podemos observar que las regiones de Aysén (XI) y Tarapacá (I) presentan mayor porcentajes de personas que no completaron ningún nivel de educación. Luego la región de Magallanes (XII) es la menor porcentaje de personas que no completaron ningún nivel educacional.

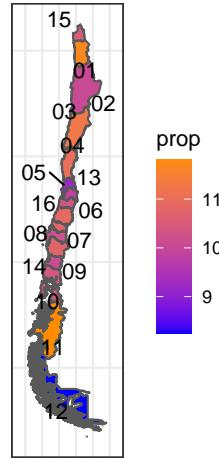
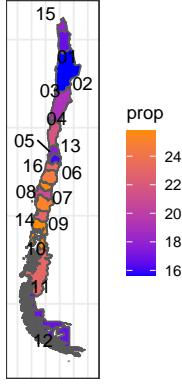


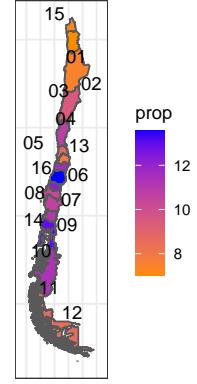
Figura 5: Porcentaje de personas sin educación por región

De Figura 6a se visualiza que las regiones de Maule (VIII), La Araucanía (IX) y de Los Lagos (X) presentan mayor porcentaje de personas que no han terminado la enseñanza básica. Es importante destacar que los porcentajes de personas que no terminaron la básica fue mayor al 24%, una cifra importante a considerar. También podemos observar que las regiones de Arica y Parinacota (XV), Tarapacá (I), Antofagasta (II), Metropolitana (XIII) y Magallanes (XII) con menor porcentaje de personas que no han terminado la básica.

De Figura 6b se visualiza que la región Libertador General Bernardo O'Higgins (VI) con el mayor porcentaje de personas que solo tienen terminada la enseñanza básica, mientras que las regiones de Arica y Parinacota (XV), Tarapacá (I), antofagasta(II), Metropolitana (XIII) y Magallanes (XII) como las regiones con menor porcentaje de personas que llegaron a completar la enseñanza básica.



(a) Incompleta

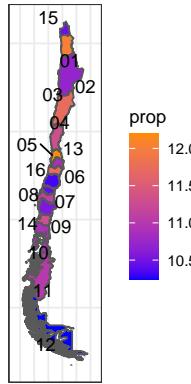


(b) Completa

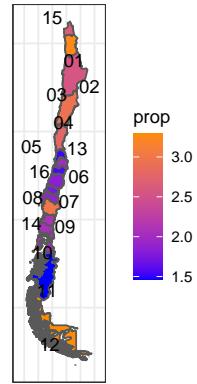
Figura 6: Porcentaje de personas con educación básica

De Figura 7a se observa que La región de Tarapacá (II) y Metropolitana (XIII) tienen mayor porcentaje de personas que no llegaron a completar la enseñanza media en un colegio/instituto humanista, mientras que las regiones de Magallanes (XII) y Libertador General Bernardo O'Higgins (VI) las regiones con menor porcentaje de personas que no completaron la enseñanza media en un colegio/instituto humanista.

De Figura 7b se logra apreciar que la región de Tarapacá (II), La Araucanía (IX) y Magallanes (XII) presentan mayor porcentaje de personas que no llegaron a completar la enseñanza media en un instituto Técnico Profesional. Mientras que las regiones de Aysén (XI), Ñuble(XVI) y Valparaíso (V) el menor porcentaje de personas que no llegaron a completar la enseñanza media en un instituto profesional.



(a) Humanista



(b) Técnico Profesional

Figura 7: Porcentaje de personas con educación media incompleta por región

De Figura 8a logramos observar que la región de Antofagasta (II) es la que presenta mayor porcentaje de personas que completaron la enseñanza media en colegios/institutos humanistas, mienytas que las regiones de la zona sur y un poco del centro (Magallanes, Aysén, Los Lagos, La araucanía, Los Ríos, Maule, Biobio, Ñuble) tienen un menor porcentaje de personas que completaron la enseñanza media en colegios humanistas.

Tenemos que Figura 8b se ve que las regiones de Tarapacá (II) y Magallanes (XII) presentan mayor porcentaje de personas que completaron la enseñanza media en institutos Técnicos Profesionales. Mientras que la región de Aysén (XI) es la que presenta menor porcentaje de personas que terminaron la enseñanza media en institutos Técnicos Profesionales.

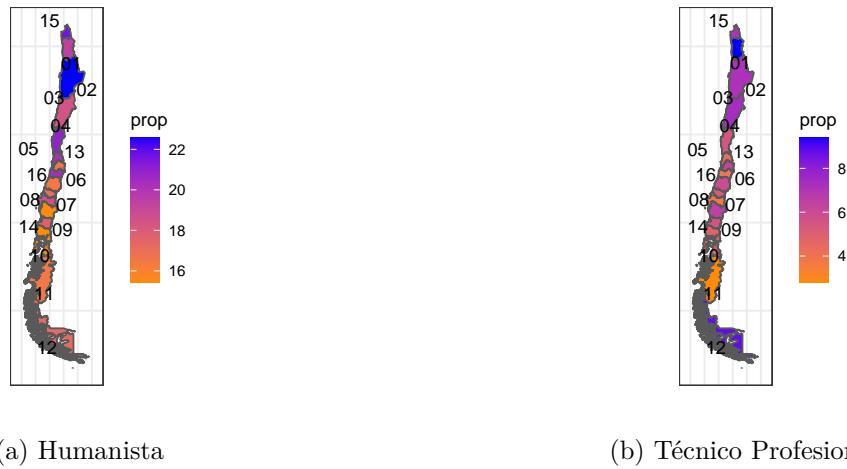


Figura 8: Porcentaje de personas con educación media completa por región

Observando Figura 9a obtenemos que las regiones de Antofagasta (II), Valparaíso (V), Metropolitana (XIII) y Magallanes (XII) con mayor porcentaje de personas que tienen educación Técnica incompleta. Cabe notar que este porcentaje es uno de los más bajos (no supera el 3,5%). Mientras que las regiones de Los Ríos (XIV) y La Araucanía (IX) presentan el menor porcentaje de personas con educación Técnica incompleta.

Analizando Figura 9b presenta mayor porcentaje de personas con educación Técnica completa la región de Magallanes (XII). mientras que múltiples regiones de la zona centro y sur (Tarapacá, Coquimbo, Maule, La Araucanía, Los Ríos y Los Lagos) tienen menor porcentaje de personas que completaron la educación Técnica.



Figura 9: Porcentaje de personas con educación técnica por región

Revisando Figura 10a se logra ver que las regiones de Antofagasta (II), Valparaíso (V), Metropolitana (XIII) y Ñuble (XVI) presentan mayor porcentaje de personas que tienen estudios profesionales incompletos. Por contraparte, las regiones de Coquimbo (IV) y Aysén (XI) presentan el menor porcentaje de personas con estudios profesionales incompletos.

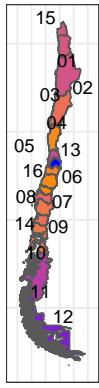
Observando Figura 10b podemos denotar que la región Metropolitana (XIII) es la única región con mayor porcentaje de personas con postgrados incompletos. Mientras que las demás regiones sus porcentajes son menores al 0.3%. Es interesante, igualmente, ver que este porcentaje no es mayor al 0.5% todos y que el más alto está en la región metropolitana.



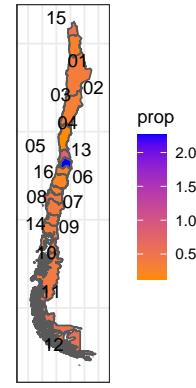
Figura 10: Porcentaje de personas con educación profesional por región

De Figura 11a podemos ver que el mayor porcentaje de personas que completaron estudios profesionales se encuentran concentrada en la región Metropolitana (XIII), seguida de la región de Magallanes (XII) con no más de 10%. Todas las demás regiones tienen un porcentaje menor a 8%.

Finalmente un caso muy interesante de analizar, en Figura 11b, observaremos el porcentaje de personas que terminaron un postgrado. De esta figura vemos una concentración extrema en la región Metropolitana (XIII), donde hay más de un 2% de personas que completaron un postgrado. Todas las demás regiones tienen menos de un 0.5% de personas con un postgrado completo. De esta información, al igual que los ingresos totales, analizaremos que sucede en la región Metropolitana según comunas.



(a) Incompleta



(b) Completa

Figura 11: Porcentaje de personas con profesional completo por región

Analizando el caso particular de Santiago sobre las personas que tienen un postgrado completo obtendremos la Figura 12

De Figura 12 podemos observar que existen comunas que no tienen personas que completaron un postgrado, que la mayoría tiene menos de un 2,5% de personas con un postgrado, y más importante que existen 3 comunas donde hay cerca de un 10% de personas con un postgrado, y exactamente esas comunas son Vitacura, Providencia y Las Condes, las cuales son las 3 comunas con el ingreso más alto según Figura 2. Entonces podemos suponer que esta comuna presenta ingresos altos debido a las personas con postgrado completo. Esta idea se fortalece debido al conocimiento popular que las personas que estudian un postgrado tiene sueldo mucho mayor al de las demás personas.

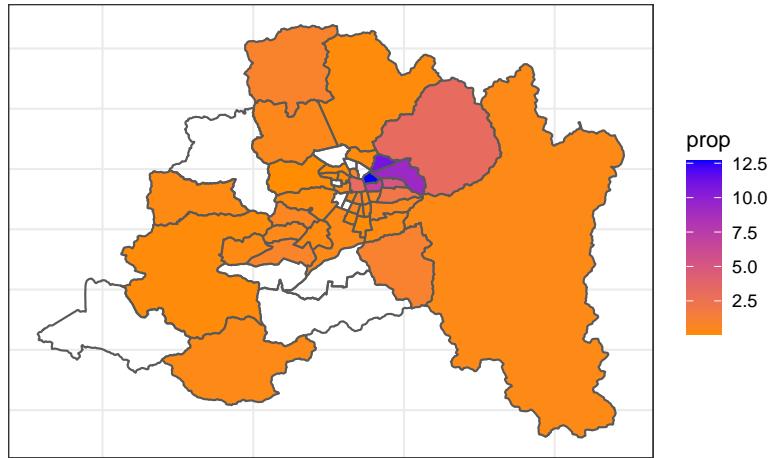


Figura 12: Porcentaje de personas con Postgrado completo por comunas de Santiago

Con todos los datos anteriores podemos ver semejanzas, de que existen una relación entre los ingresos, la pobreza, la demografía, los estudios, entre otros. Tal como muestra Figura 13 con las correlaciones de las variables ocupadas.

Podemos observar en Figura 13 la correlación que existe entre la variable de Allegamiento Interno y la variable de número de personas en el hogar, la variable de escolaridad y condición de actividad, hacinamiento y Allegamiento Externo, entre otros.

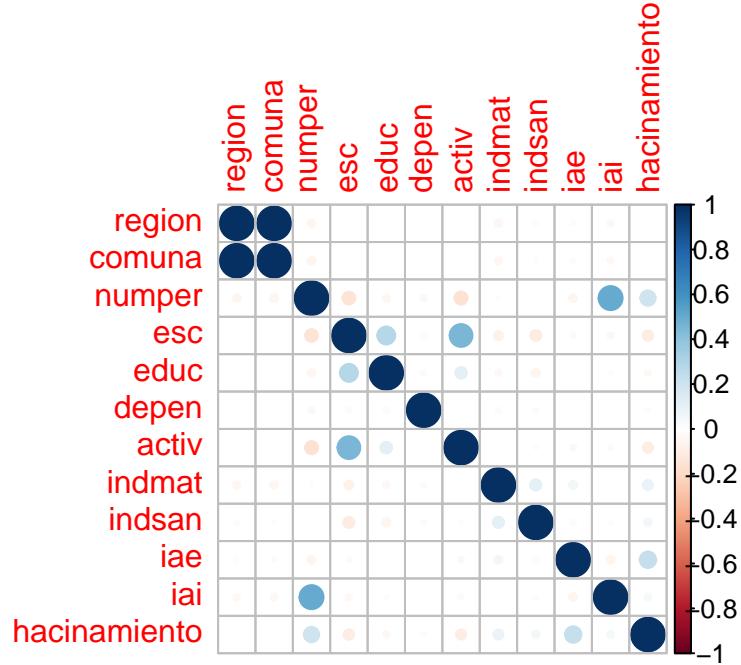


Figura 13: Correlograma de las variables

Con estos datos se puede plantear un modelo factorial (Debido a que tratamos con variables categóricas), con ellos poder ver cuanto son las cantidades de factores necesarios para describir el modelo con la mayor variabilidad posible.

Tal como describimos en **Metodología**, utilizaremos el método de varianza explicada por factores. Tal como muestra Figura 14 se muestra el porcentaje de varianza explicada por cada factor.

Con Figura 14 mostramos la varianza explicada por cada factor, a simple vista no logramos observar la cantidad de factores necesarios según el criterio de varianza explicada. Es por esta razón que creamos Figura 15 donde se muestran las varianzas explicadas por los factores acumulados, es decir la varianza explicado por el factor más todos los que le anteceden.

Como podemos observar en Figura 15 dado que buscamos una varianza explicada entre un 70% y 90%. Con ese intervalo vemos que con 8 factores tenemos la mayor varianza explicada que se encuentra dentro del rango acordado. Con un total de 8 factores logramos explicar el 86%.

Por ello nos quedaremos solamente con 8 factores. Usando este nuevo modelo, obtendremos la Figura 16 donde se muestran las varianzas y varianzas acumuladas de cada factor.

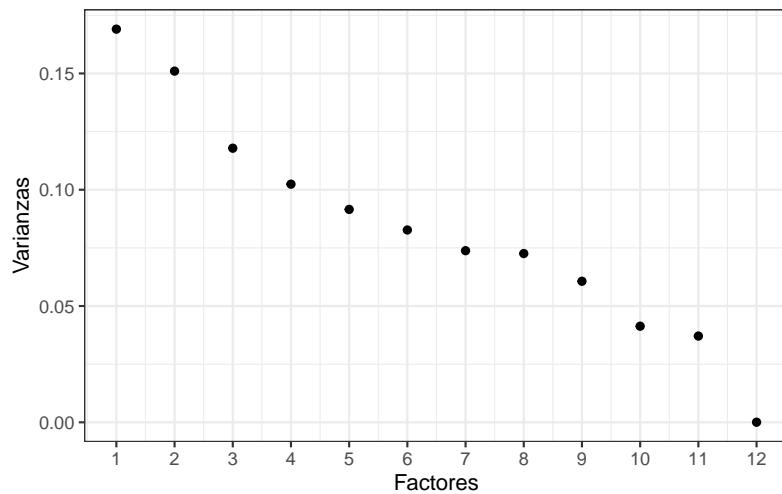


Figura 14: Porcentaje de varianza explicada por cada factor

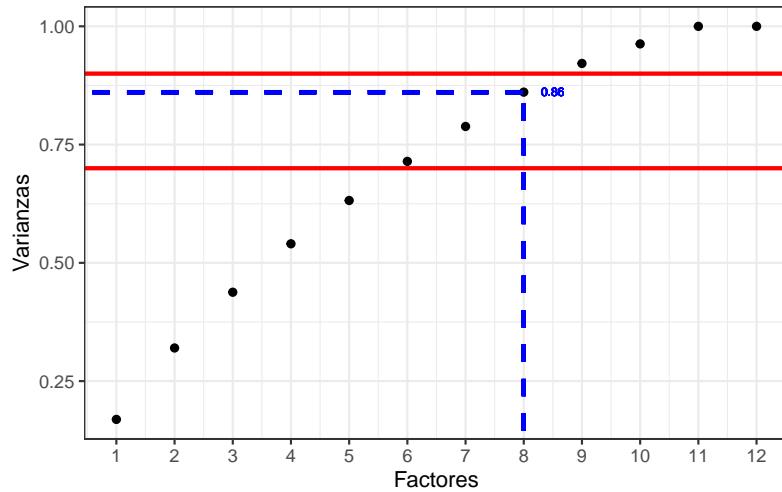


Figura 15: Porcentaje de varianza explicada por cada factor acumulado

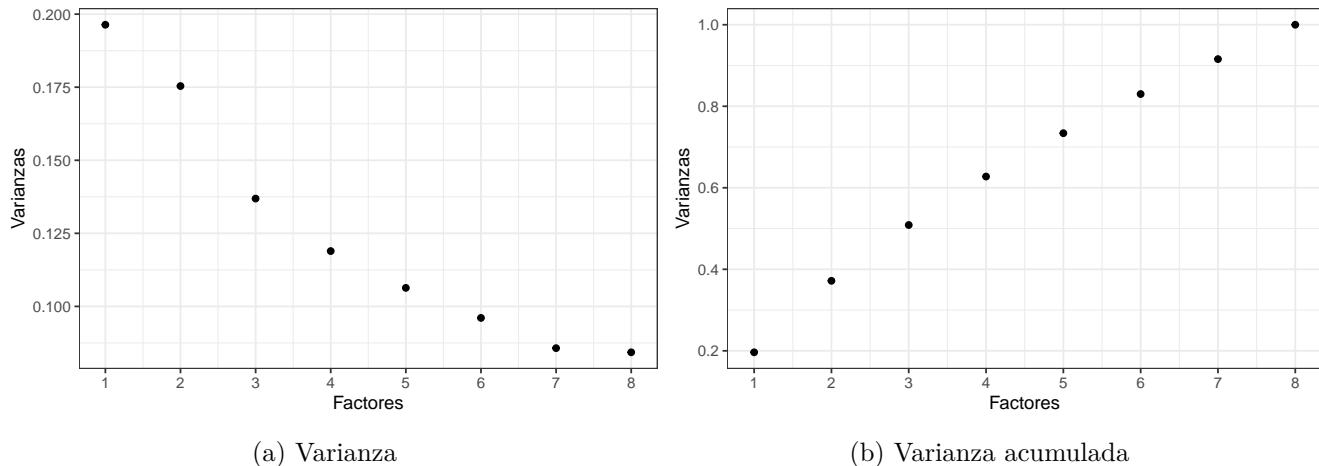


Figura 16: Varianza explicada por cada factor

Ya con la cantidad de factores elegidos y la variabilidad explicada de cada uno, ahora podemos analizar las correlaciones de cada factor (o componente) con las variables como muestra la Figura 17

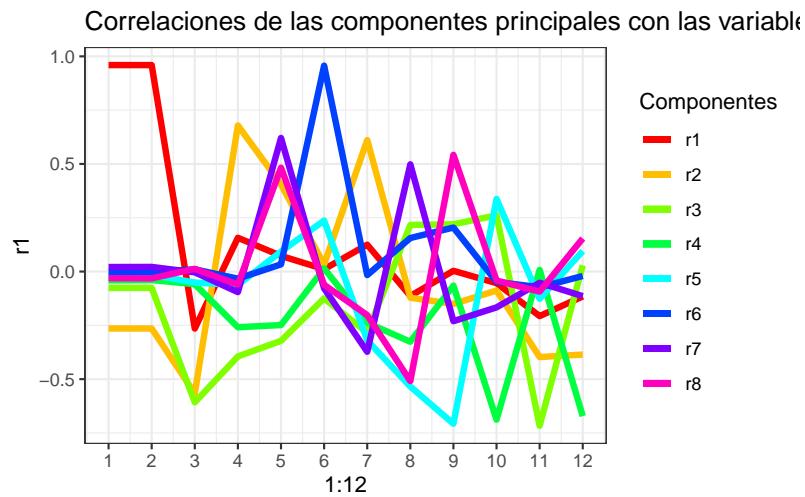


Figura 17: Correlaciones entre componentes y variables

Podemos apreciar de Figura 17 que tenemos muchas variabilidad de casos, hay algunos componentes que estan muy correlacionados con las variables, como es el caso del primer componente con las primeras 2 variables, el sexto componente con la sexta variable, tomaremos como muy correlacionados aquellos componentes con correlación mayor a 0.5 y menor a -0.5 de la Figura 17. También existen componentes que no estan correlacionados con algunas variables, como es el caso de todos los que se encuentran entre -0.5 y 0.5 de la Figura 17.

Con este modelo podemos observar que tanto explica el modelo a cada variable por separado. Tal como muestra la Figura 18

Ya con todos los datos evaluados y vemos que podemos trabajar con las 8 componentes principales que obtuvimos. Ahora ha llegado el momento de generar una regresión logística con los 8 componentes

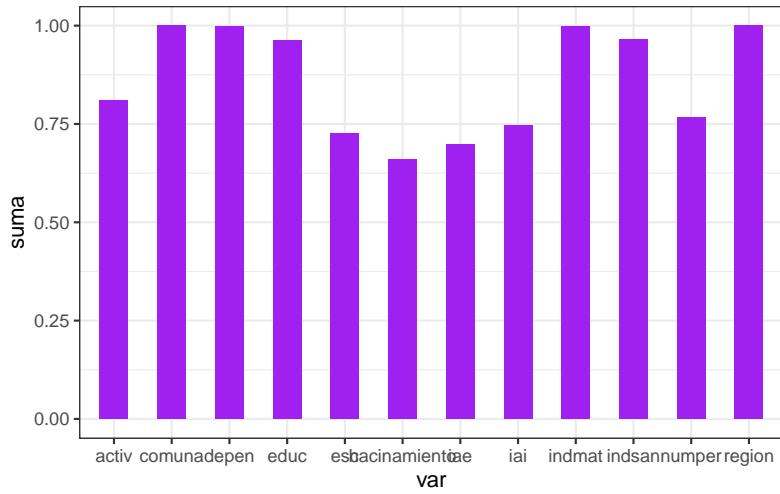


Figura 18: variabilidad explicada del modelo por cada variable

principales como planteamos en el enunciado.

## Regresión logística

Utilizando los 8 componentes para poder predecir el índice de pobreza, agrupando a las personas pobres no extremas y pobres extremas en el mismo grupo, pobre.

Al generar un modelo logístico obtenemos los valores de significancia que muestra Tabla 1.

	Estimate	Std. Error	z value	Pr(> z )
(Intercept)	-1.950	0.068	-28.833	0
X1	-0.069	0.003	-22.181	0
X2	-0.261	0.007	-39.864	0
X3	0.207	0.010	20.050	0
X4	-0.437	0.022	-19.942	0
X5	-0.804	0.025	-32.343	0
X6	0.241	0.007	35.187	0
X7	-0.193	0.018	-11.038	0
X8	0.415	0.024	17.496	0

Tabla 1: Resultados obtenidos

Por Tabla 1 observamos que nuestros 8 componentes principales son significativos para predecir la pobreza. Ahora para cada componentes junto con el intercepto, se calculará el intervalo de confianza para estimarlos. Tal como muestra Tabla 2 de esta manera quedan estimados los coeficientes con una confianza del 95%.

	2.5 %	97.5 %
(Intercept)	-2.083	-1.818
X1	-0.075	-0.062
X2	-0.274	-0.248
X3	0.187	0.227
X4	-0.481	-0.395
X5	-0.853	-0.755
X6	0.228	0.255
X7	-0.228	-0.159
X8	0.368	0.461

Tabla 2: Intervalos de confianza obtenidos

Finalmente se obtendrá la cantidad

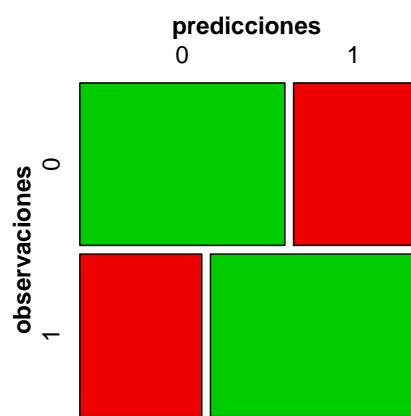


Figura 19: Mosaico de predicciones y valores reales del modelo logístico

## Regresión lineal

Ahora usaremos los 8 componentes principales para generar una regresión lineal para poder estimar los ingresos totales. Luego de crear este modelo con los 8 componentes principales, compararemos con modelos sin alguno de estos componentes y compararemos los AIC de los modelos, tal como muestra Figura 20.

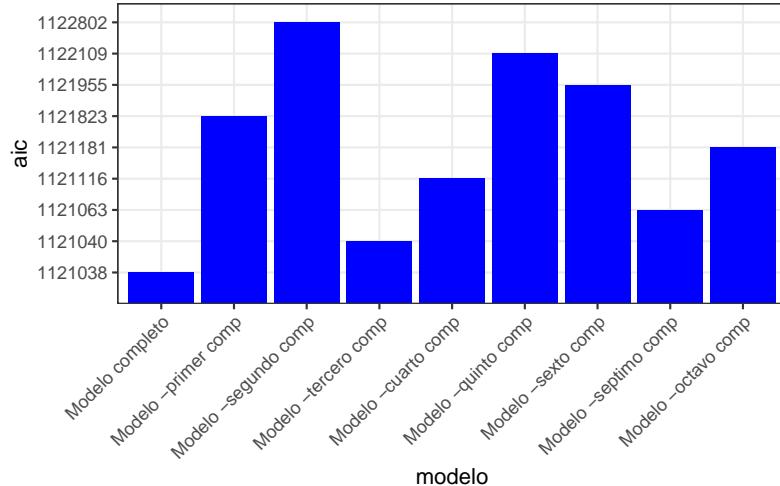


Figura 20: AIC de los modelos

Según Figura 20, el modelo completo es aquel que presenta menor criterio de información de Akaike (AIC), por lo tanto es correcto utilizar estos 8 componentes.

Analizando los coeficientes de la regresión lineal obtenemos lo visto en Tabla 3.

	Estimate	Std. Error	t value	Pr(> t )
(Intercept)	350915.587	8548.667	41.049	0.000
V1	10725.297	380.490	28.188	0.000
V2	33534.123	789.253	42.488	0.000
V3	-2590.542	1315.965	-1.969	0.049
V4	22878.651	2551.309	8.967	0.000
V5	95589.037	2899.023	32.973	0.000
V6	-24685.751	809.725	-30.487	0.000
V7	10473.847	2017.477	5.192	0.000
V8	-32551.171	2704.169	-12.037	0.000

Tabla 3: Resultados obtenidos

Viendo Tabla 3 podemos observar que el tercer componente resultó no ser tan significativo como el resto, pero como ya comprobamos, si lo eliminamos aumenta el AIC. Pese a no ser tan significante como el resto. Ahora, ya analizado los coeficientes, vamos a crear un intervalo de confianza para cada coeficiente. A continuación, en @tble-4 se mostrará los intervalos de confianza de cada uno.

	2.5 %	97.5 %
(Intercept)	334160.001	367671.17
V1	9979.529	11471.07
V2	31987.168	35081.08
V3	-5169.865	-11.22
V4	17878.026	27879.28
V5	89906.884	101271.19
V6	-26272.830	-23098.67
V7	6519.546	14428.15
V8	-37851.406	-27250.94

Tabla 4: Intervalos de confianza obtenidos

Sin embargo, pese a que la cantidad de componentes que genera el modelo es el más eficiente, el modelo no es buen predictor debido a la existencia de valores atípicos. Estos valores atípicos no permiten ajustar el modelo de la mejor manera posible, esto se logra apreciar a travez de la Figura 21.

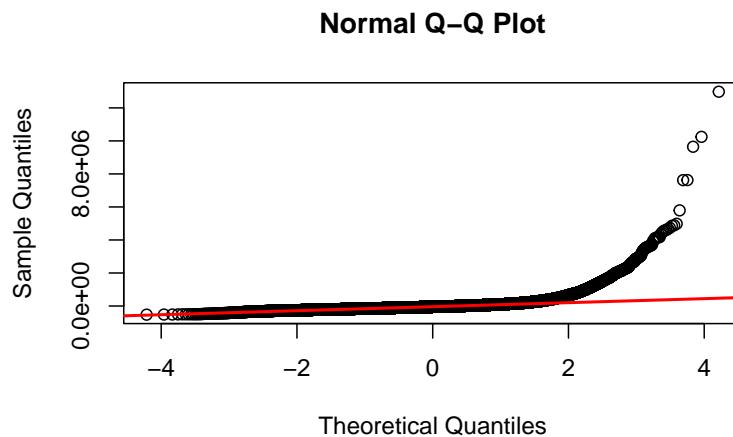


Figura 21: Q-Q Plot del modelo

Con Figura 21, se muestra claramente que existen una cantidad considerable de datos atípicos, los cuales aumentan los residuos del modelo. Tal como analizamos anteriormente, cuando hablamos de ingresos en chile, existe un grupo (concentrado en la comuna de vitacura) el cual tiene ingresos elevados.

## **Conclusión**

Tras realizar un análisis factorial logramos crear 12 factores. De los 12 factores, por criterio de varianza explicada, calculamos que con los 8 primeros factores son capaces de explicar suficiente varianza de los datos. Con esos 8 componentes principales se pudo reducir la dimensionalidad de los datos, y poder crear un modelo logístico que logra caracterizar y, por ende, logra predecir bien qué sectores de la población serán considerados o no dentro de la pobreza. Aquí, los 8 componentes principales resultaron significativos y el modelo fue capaz de predecir de manera muy acertada.

Además, se comprobó que no es posible realizar un análisis de regresión lineal el cual intente explicar los ingresos totales. Al analizar los ingresos se observa que existe una importante cantidad de

## Bibliografía (EDITAR DESPUÉS)

1. Chung-hong Chan, Geoffrey CH Chan, Thomas J. Leeper, and Jason Becker (2021). *rio*: A Swiss-army knife for data file I/O. R package version 0.5.29.
2. Hijmans R (2022). *raster: Geographic Data Analysis and Modeling*. R package version 3.6-11, <https://CRAN.R-project.org/package=raster>.
3. Libro de códigos Base de Datos Casen 2017
4. Meyer D, Zeileis A, Hornik K (2006). “The Strucplot Framework: Visualizing Multi-Way Contingency Tables with vcd.” *Journal of Statistical Software*, 17(3), 1-48. doi:10.18637/jss.v017.i03 <https://doi.org/10.18637/jss.v017.i03>
5. Pebesma, E., 2018. Simple Features for R: Standardized Support for Spatial Vector Data. *The R Journal* 10 (1), 439-446, <https://doi.org/10.32614/RJ-2018-009>
6. Pedersen T (2022). *ggforce: Accelerating ‘ggplot2’*. R package version 0.4.1, <https://CRAN.R-project.org/package=ggforce>.
7. Slowikowski K (2022). *ggrepel: Automatically Position Non-Overlapping Text Labels with ‘ggplot2’*. R package version 0.9.2, <https://CRAN.R-project.org/package=ggrepel>.
8. South A (2012). *rworldxtra: Country boundaries at high resolution..* R package version 1.01, <https://CRAN.R-project.org/package=rworldxtra>
9. Taiyun Wei and Viliam Simko (2021). R package ‘corrplot’: Visualization of a Correlation Matrix (Version 0.92). Available from <https://github.com/taiyun/corrplot>
10. Wickham H, Averick M, Bryan J, Chang W, McGowan LD, François R, Grolemund G, Hayes A, Henry L, Hester J, Kuhn M, Pedersen TL, Miller E, Bache SM, Müller K, Ooms J, Robinson D, Seidel DP, Spinu V, Takahashi K, Vaughan D, Wilke C, Woo K, Yutani H (2019). “Welcome to the tidyverse.” *Journal of Open Source Software*, 4(43), 1686. doi:10.21105/joss.01686 <https://doi.org/10.21105/joss.01686>.
11. Wickham H, François R, Henry L, Müller K (2022). *dplyr: A Grammar of Data Manipulation*. R package version 1.0.9, <https://CRAN.R-project.org/package=dplyr>.
12. Wickham H, Girlich M (2022). *tidyr: Tidy Messy Data*. R package version 1.2.0, <https://CRAN.R-project.org/package=tidyr>.
13. Wickham H, Seidel D (2022). *scales: Scale Functions for Visualization*. R package version 1.2.1, <https://CRAN.R-project.org/package=scales>.