TRABAJO DE FIN DE GRADO

Grado en Estadística

Título: Regularización en problemas lineales de regresión: Una visión Estadística y Económica aplicada a datos de desarrollo humano.

Autor: Álvaro Garnica Barco

Director: Josep Maria Ollé, Montserrat Termes Rifé i Esteban Vegas Lozano

Departamento: Econometría, Estadística y Economía aplicada

Convocatoria: Enero 2021

UNIVERSITAT DE BARCELONA

UNIVERSITAT POLITÈCNICA DE CATALUNYA
BARCELONATECH
Facultat de Matemàtiques i Estadística

Resumen

En este trabajo se estudian diferentes métodos de regularización aplicados a problemas predictivos de regresión. Para ello se utiliza una base de datos relativa a indicadores de desarrollo mundial y se enseñan los pasos típicos a seguir en este tipo de estudios, desde la recolecta y tratamiento de datos, hasta la obtención de predicciones y resultados. Este estudio abarca los campos del *Machine Learning* y de la Inteligencia Artificial, por lo que nos servirá también para hacer hincapié sobre la utilidad y las posibilidades de estos campos, y más concretamente de los *Open Data,* en la economía y en la sociedad en general. Destacar finalmente que, para complementar este trabajo se ha desarrollado una aplicación web, donde el usuario podrá visualizar lo explanado a lo largo de esta memoria de forma mucho más interactiva.

Palabras clave: Aprendizaje automatizado, Inteligencia Artificial, regresión, Reducción de la dimensionalidad, Clustering, Regularización, Predicción, Indicadores de desarrollo, Open Data

Abstract

This work studies different regularization methods applied to predictive regression problems. For this, a world development indicators database has been used. We will also see the classic steps to follow in such a task, from data collection and processing to obtaining predictions and results. This study falls into the fields of Machine Learning and Artificial Intelligence, so it will also serve to emphasize the usefulness and possibilities of these fields in economics studies. Finally, it should be noted that to complement this work a web application has been developed. In this the user can visualize what has been explained throughout this work in a much more interactive way.

Keywords: *Machine learning, Artificial Intelligence, regression, Dimensionality reduction, Clustering, Regularization, Prediction, Development indicators, Open Data*

Clasificación AMS: 62J05 *Linear regression*
62J07 *Ridge regression; shrinkage estimators*
62H25 *Factor analysis and principal components*
91C20 *Clustering*

Índice

1. Introducción

Los términos de *Machine Learning* y de Inteligencia Artificial están cada vez más a la orden del día. El *Machine Learning*, o aprendizaje automatizado, tiene por objetivo crear sistemas que aprendan automáticamente patrones complejos a partir de unos determinados datos. Es decir, se aplican algoritmos matemáticos y estadísticos sobre una serie de datos para posteriormente predecir comportamientos futuros. Se utilizan tales modelos en infinidad de campos y ámbitos, siendo uno de ellos por supuesto la economía. De aquí surge uno de los primeros objetivos de este trabajo. Se quiere combinar este tipo de técnicas estadísticas con datos de relevancia económica, ligando así los grados de estadística y economía. Otro objetivo será el de mostrar y acercar a un público de perfil menos técnico, los pasos que se suelen tomar antes y después de aplicar tales algoritmos predictivos.
Por otra lado, y enfocado más al grado de estadística, se han querido estudiar diferentes técnicas de regularización aplicadas a modelos lineales, que surgen como respuesta a uno de los fallos más comunes en modelos predictivos, el sobreajuste. Aunque esto se detalla en el trabajo, ya se adelanta aquí que el sobreajuste es un problema que conlleva que el modelo no prediga correctamente comportamientos futuros, por lo que es importante intentar corregirlo. Estas técnicas de regularización no se han estudiado a lo largo de la carrera, pero se combinan con técnicas sí vistas durante la carrera. Por lo tanto se espera que este trabajo de fin de grado pueda llegar a servir a otros estudiantes como complemento para asimilar mejor ciertos conceptos. Para ello además, se ha desarrollado una aplicación web interactiva, donde el usuario podrá interaccionar con los datos y modelos utilizados.

Una vez detallados los objetivos, podemos indagar algo más en los datos y técnicas utilizados. Como se ha dicho, se ha querido que los datos estuviesen relacionados con la economía, es por ello que la base datos está compuesta de varios indicadores para países de todo el mundo publicados por el Banco Mundial. La técnica base de *Machine Learning*, sobre la que se apoyan los métodos de regularización utilizados, es la regresión lineal. La regresión lineal, a menudo utilizada como método de inferencia, es aquí utilizada como método de predicción. Por lo que el análisis de los resultados obtenidos tendrá un enfoque puramente predictivo y no inferencial.

La estructura del trabajo es la siguiente:
En el primer apartado, se hace una descripción puramente teórica de los conceptos que se ven a lo largo de la memoria. Se ha intentado que las explicaciones lleguen al máximo de público posible, aún así el lector deberá tener un mínimo de conocimiento de conceptos matemáticos y estadísticos.
Seguidamente, ya con un punto de vista más practico, se hace una primera aproximación de los datos en forma de análisis exploratorio. La idea aquí es preprocesar y entender mejor los datos con los que se trabaja. Mediante técnicas como el análisis de componentes principales y el *clustering* se realiza un *profiling* de los países de la base de datos. Este apartado es especialmente interesante combinarlo con la aplicación web, ya que esta nos permite visualizar mejor las características de cada país.

Posteriormente, se modelizan y analizan los resultados obtenidos de los métodos de regularización y de predicción utilizados. Se comienza ajustando una regresión lineal, y se sigue ajustando los modelos de regularización en cuestión. La idea es por lo tanto ver y contrastar empíricamente como actúan, y como afectan a los resultados estos métodos. Los modelos implementados son: Ridge, Lasso, *group* Lasso, ElasticNet, regresión por componentes principales y regresión por mínimos cuadrados parciales.
En el siguiente apartado se hace hincapié sobre el potencial que tienen los datos abiertos sobre la sociedad y la economía. Sirve también para ver lo ligado que está el mundo del dato con sectores y servicios que quizás uno no se imagina.
Por último, se finaliza el trabajo con las conclusiones pertinentes.

2. Metodología

Se explica en este apartado la parte más teórica de la técnicas y modelos utilizados a lo largo del trabajo. Se verá también cual es la procedencia de los datos empleados así como los recursos informáticos que se han utilizado.

2.1. Datos y recursos informáticos

Los datos brutos, antes de ser procesados, han sido extraídos de la base de datos "*World Development Indicators"* del *DataBank* del Banco Mundial[1]. Esta base de datos contiene más de 1400 variables para todos los países del mundo y para fechas que se remontan en algunos casos hasta 1960. En este caso se han seleccionado todos los países del mundo y alrededor de 30 variables numéricas para datos relativos al año 2015[2]. Una vez extraídos estos datos en formato CSV se han realizado pequeños cambios, como por ejemplo cambiar la codificación de las variables. Posteriormente se han cargado estos datos en el entorno de desarrollo *RStudio*, donde se ha programado la totalidad del trabajo en lenguaje R. Por otra parte, se ha utilizado la plataforma *shinyapps.io*[3] para desplegar la aplicación *shiny* y la plataforma *github* para compartir el código desarrollado a lo largo del trabajo.

2.2. Análisis de componentes principales y clustering

a. Análisis de componentes principales

El análisis de componentes principales o PCA es una técnica de reducción de la dimensionalidad. Se busca describir un conjunto de datos en términos de nuevas variables (o componentes) que retengan el máximo posible de información. Este método convierte un conjunto de variables posiblemente correlacionadas en un nuevo conjunto de variables artificiales sin correlación entre ellas. A estas nuevas variables se les llama componentes principales, y cada una de estas es una combinación lineal del juego de variables original.
Esta técnica define las coordenadas de la primera componente de modo que la mayor varianza se recoja en este eje. La segunda mayor varianza será recogida por la segunda componente, y así sucesivamente. En total habrán tantas componentes como variables originales hayan, pero como se ha dicho, las primeras componentes principales serán las más interesantes ya que serán las que mas información recojan. La idea es que se pueden ignorar las últimas componentes, bastará con seleccionar[4] un número inferior de componentes para obtener casi la misma información que se obtendría con todas las variables, de ahí que se hable de reducción de la dimensionalidad.

[1] *World Developement Indicarors* - databank.worldbank.org/source/world-development-indicators

[2] A los datos más recientes les faltaba mucha información

[3] Shiny Apps - shinyapps.io

[4] en el apartado 4.2.1 se especifica como seleccionar el número de componentes apropiado

b. Clustering

El *clustering* (o algoritmo de agrupación) tiene por objetivo encontrar *'clusters'* (o grupos) de individuos de modo que los elementos de un *cluster* sean lo más parecidos posibles entre sí, y de forma que los *clusters* sean lo más diferente posible entre ellos. La formación de estos grupos permite sintetizar la descripción del conjunto de datos. Es decir, describiendo cada *cluster* se debería obtener una descripción bastante precisa de la mayoría de miembros de ese *cluster*.

Las dos principales técnicas de agrupamiento son:

- agrupamiento jerárquico
- agrupamiento no jerárquico[5]

En este apartado se explica tan solo el *clustering* no jerárquico conocido como *k-means*.
Este método genera *k* grupos en el que cada observación es asignada al grupo cuya media es más cercana. Es decir, se agrupan las observaciones en función de su similaridad, la cual se mide en términos de distancia (usualmente se utiliza la distancia Euclidea).

Los pasos del algoritmo son los siguientes:

a. Elegir *k*, el número de *clusters* que se desean obtener[6]
b. Se generan *k* centroides iniciales aleatoriamente
c. Se asigna cada observación al centroide más cercano
d. Se recalculan *k* centroides
e. Se repite c. y d. hasta lograr convergencia (cuando los centroides dejan de moverse en el punto d.)

Cabe destacar, que a menudo se suele realizar el *clustering* a partir de las componentes principales obtenidas del análisis de componentes principales. De este modo el análisis y el *profiling* de los grupos obtenidos puede ser más intuitivo, ya que basta con analizar las primeras componentes para obtener una idea precisa de en que se diferencian los *clusters* obtenidos.

2.3. ¿Qué es la regularización?

Antes de entender el concepto de regularización y su relación con el *Machine Learning* hay que preguntarse por qué hace falta la regularización.
Como bien sabemos el *Machine Learning* consiste en entrenar un modelo con datos relevantes y usar el modelo para predecir datos desconocidos. A menudo surge un problema conocido como *overfittin*g o sobreajuste. El *overfitting* aparece cuando se sobreentrena el modelo con unos datos particulares y a causa de esto el modelo no es capaz de predecir correctamente datos no vistos previamente. Es decir, el modelo se ajusta demasiado a las características específicas de los datos de entrenamiento y no logra generalizar

[5] en estos métodos el número de grupos *k* se determina de antemano

[6] en el apartado 3.4 se enseña como elegir el número óptimo

adecuadamente, por lo que no logra predecir situaciones distintas a las devenidas durante este entrenamiento.
Por otra parte, hay que entender el concepto de *bias-variance tradeoff* (compensación entre sesgo y varianza). Todo error de predicción para cualquier modelo de *Machine Learning* puede dividirse en 3 partes:

- Error de sesgo (*bias*)
- Error por varianza
- Error irreductible

El error irreducible, como su nombre indica, no puede reducirse. Este resulta del ruido a la hora de plantear el problema mismo.
El sesgo surge de las suposiciones erróneas o simplificadoras hechas por el algoritmo de aprendizaje, esto hace que la función objetivo sea más fácil de aprender.
La varianza puede ser vista como la cantidad que cambiará la estimación de la variable respuesta si se utilizasen diferentes datos de entrenamiento. Es decir, la varianza es un error de sensibilidad a pequeños cambios en el conjunto de entrenamiento.
La complejidad del modelo será normalmente la que dicte en que tipo de error se está incurriendo. En un modelo poco complejo no se conseguirá atrapar correctamente la relación entre las variables explicativas y la variable respuesta. Este generalizará demasiado, se tendrá baja varianza pero alto sesgo (desajuste). Un modelo muy complejo se adaptará muy bien a sus datos de entrenamiento pero no conseguirá generalizar adecuadamente. Se tendrá bajo sesgo pero alta varianza (sobreajuste).
Como se ve, puede resultar muy difícil conseguir tanto un bajo sesgo como una baja varianza. Por ello hay que encontrar un nivel de complejidad intermedio que consiga el mínimo posible de ambos errores.

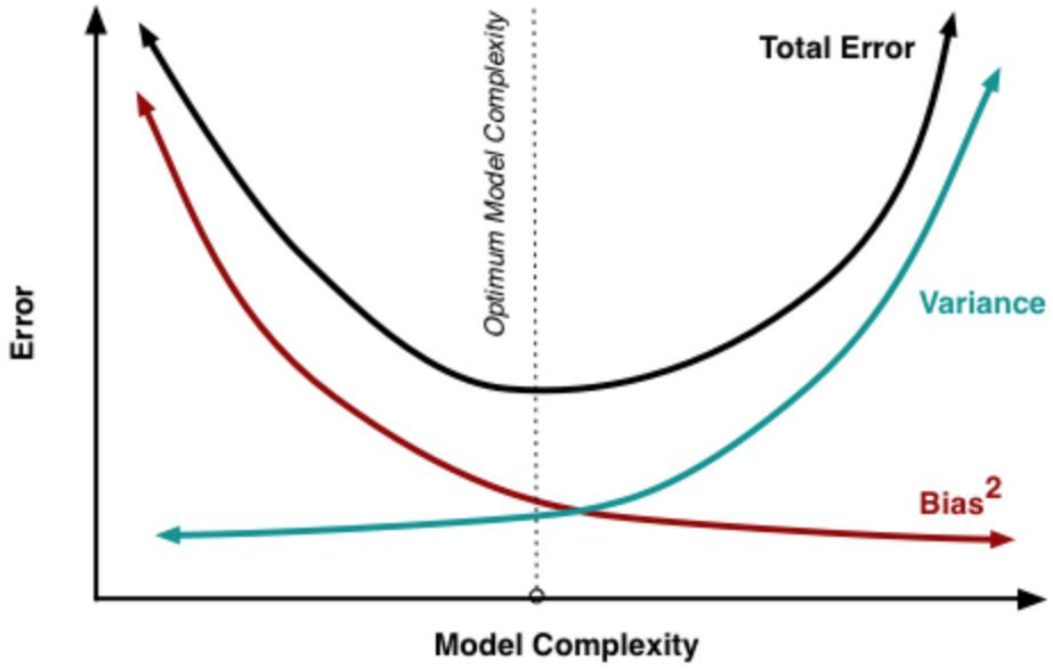


Fig. 2.3.1: *Bias-Variance Tradeoff*[7]

[7] Fuente: analyticsvidhya.com/blog/2020/08/bias-and-variance-tradeoff-machine-learning/

En el análisis de regresión el problema de sobreajuste es también común. En este caso el modelo será demasiado complejo cuando el número de variables comparado con el número de observaciones sea demasiado grande. Efectivamente, si en el modelo se añaden muchas variables sin relación con la variable respuesta estas pueden aparecer como estadísticamente significativas, estas variables se mantendrían por lo tanto en el modelo incurriendo a problemas de sobreajuste[8].
La regularización es un método que se aplica como solución a los problemas de sobreajuste a la hora de modelizar una regresión lineal. Esta técnica desalienta el aprendizaje de un modelo más complejo añadiendo una penalización a la función de coste[9], contrayendo los coeficientes hacia cero. Se intenta así evitar el *overfitting* minimizando la influencia en el modelo de los predictores menos relevantes.
En términos de descomposición sesgo-varianza, la regularización introduce sesgos en la solución de regresión para reducir la varianza.

2.4. Modelos de regresión que se implementan

2.4.1. Regresión Lineal Múltiple

La regresión lineal es un modelo matemático que usa una o varias variables explicativas[10] para predecir el resultado de una variable respuesta. El objetivo es modelizar la relación lineal entre las variables explicativas (independientes) y la variable respuesta (dependiente).

La fórmula que define este modelo es la siguiente:

$$Y = \beta_0 + \beta_1 X_1 + \cdots + \beta_p X_p + \varepsilon$$

Donde:
$Y =$ Variable dependiente
$X_1, X_2, \ldots, X_p =$ variables independientes
$\beta_0, \beta_1, \beta_2 \ldots, \beta_p =$ parámetros respectivos a cada variable independiente
p = número de variables independientes
ε = término de error

Los parámetros β será lo que se busque estimar, los cuales se calculan mediante mínimos cuadrados ordinarios (MCO). Este método halla las estimaciones minimizando la diferencia de sumas de cuadrados entre las respuestas observadas y las respuestas predichas por el modelo (suma de residuos cuadrados).

[8] a este hecho se le conoce como *paradoja de Freedman, Freedman, D. A. (1983) "A note on screening regression equations." The American Statistician, 37, 152–155.*

[9] este concepto se explica en los apartados siguientes

[10] cuando son varias hablamos de regresión lineal múltiple

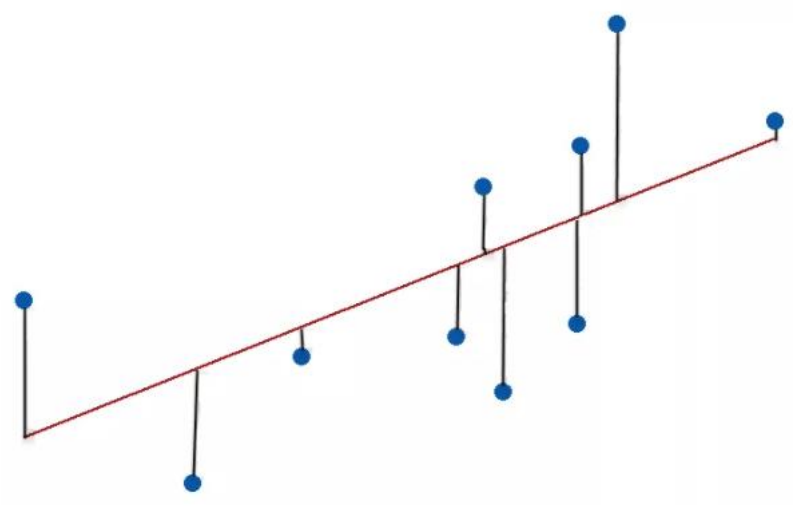


Fig. 2.4.1: representación gráfica MCO

En esta imagen se puede visualizar una representación de esta diferencia entre las respuestas observadas (puntos azules) y las respuestas predichas (proyecciones sobre la línea roja), en un caso con tan solo una variable independiente.

2.4.2. Regresión de Ridge

La regresión de Ridge es una variación de la regresión lineal. Más concretamente un método de regularización el cual reduce el valor de los coeficientes del modelo pero sin que lleguen nunca a ser cero. Para ello aplica una penalización también conocida como 'regularización L2' y es igual a la suma de los coeficientes elevados al cuadrado: $\sum_1^p \beta_j^2$. El objetivo de Ridge será resolver el problema siguiente:

$$\hat{\beta}^{\text{ridge}} = \underset{\beta}{\operatorname{argmin}} \sum_{i=1}^{N} \left(y_i - \beta_0 - \sum_{j=1}^{p} x_{ij}\beta_j \right)^2,$$

$$\text{subject to } \sum_{j=1}^{p} \beta_j^2 \leq t,$$

Fig. 2.4.2: función objetivo Ridge[11]

Si no se tiene en cuenta la restricción que aquí vemos el problema es el mismo que el de la regresión lineal por mínimos cuadrados ordinarios. Es decir, se busca estimar los coeficientes

[11] Fuente: Trevor Hastie, Robert Tibshirani, Jerome Friedman, 2004 - *The Elements of Statistical Learning*

β que minimicen la suma de diferencias al cuadrado entre los valores reales de la variable dependiente y los valores ajustados por el modelo (suma de residuos cuadrados).
Añadiendo la restricción que podemos ver se está limitando la suma de los coeficientes al cuadrado (esta deberá ser inferior o igual a un valor cualquiera deseado *t*), obteniendo en general valores reducidos de los coeficientes.
Esta ecuación suele escribirse en su forma Lagrangiana:

$$\hat{\beta}^{\text{ridge}} = \underset{\beta}{\operatorname{argmin}} \left\{ \sum_{i=1}^{N} \left(y_i - \beta_0 - \sum_{j=1}^{p} x_{ij}\beta_j \right)^2 + \lambda \sum_{j=1}^{p} \beta_j^2 \right\}$$

Fig. 2.4.3: función objetivo Lagrangiana de Ridge

Como vemos ahora se busca minimizar la suma de residuos cuadrados más la penalización L2 ponderada por el parámetro λ[12]. $\lambda \geq 0$ es el parámetro de complejidad que controlará la cantidad en la que se contraerán los parámetros. A valores más elevados de λ más restrictivo será el modelo y más se contraerán los coeficientes. En cambio si λ es nulo el resultado será el mismo que el obtenido por mínimos cuadrados ordinarios. Esto va ligado con el *bias-variance tradeoff* : A medida que se incrementa λ se reduce la varianza pero aumenta el sesgo.

La regresión de Ridge esta ideada para mejorar la regresión lineal por mínimos cuadrados ordinarios, aún así presenta sus desventajas. Principalmente el problema es que Ridge incluye todos los predictores en el modelo final. Esto no supone un problema en cuanto a precisión del modelo, pero sí en su interpretabilidad.

- Interpretación geométrica

A continuación se hace una interpretación geométrica de cómo funciona la estimación de los parámetros en la regresión de Ridge (se utiliza un modelo con dos variables predictoras como ejemplo):

[12] la relación exacta entre λ y *t* depende de los datos y no se detalla en este trabajo, véase stats.stackexchange.com/the-proof-of-equivalent-formulas-of-ridge-regression para más información

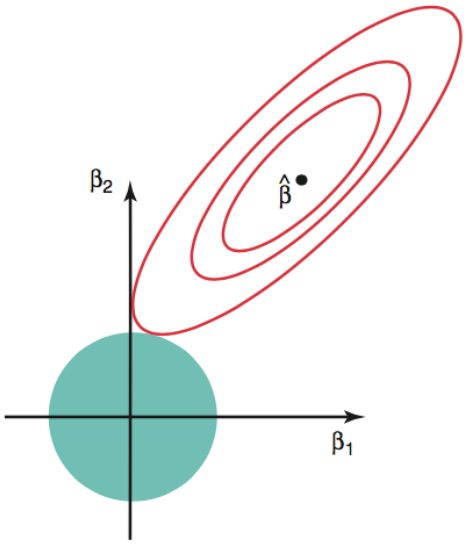


Fig.2.4.4: ilustración del error y de la función de restricción para la regresión de Ridge[13]

En esta ilustración la solución de la regresión lineal por MCO aparece como $\hat{\beta}$. Cada elipse roja representa una combinación de regresores donde la suma de residuos cuadrados es la misma (a medida que se aleja de $\hat{\beta}$ el error aumenta). El círculo verde representa la función de restricción de Ridge ($\beta_1^2 + \beta_2^2 \leq t$). Como vemos, si t es lo suficientemente grande las estimaciones serán las mismas que $\hat{\beta}$. Si no lo es, las estimaciones resultantes serán las correspondientes al primer punto de contacto entre una elipse y el círculo de restricción. Al ser esta restricción circular el punto de intersección no ocurrirá generalmente en ninguno de los ejes. Es decir, los coeficientes de Ridge no serán igual a cero.

2.4.3. Regresión de Lasso

La regularización de Lasso (*least absolute shrinkage and selection operator*) es también una variación de la regresión lineal. Al igual que Ridge este método reduce algunos coeficientes de los predictores. Se diferencia de la regresión de Ridge en que puede hacer que algunos coeficientes sean igual a cero, eliminando así algunos predictores del modelo. Para ello el método aplica una penalización también conocida como 'regularización L1' y es igual a la suma del valor absoluto de los coeficientes: $\sum_1^p |\beta_j|$. La regresión de Lasso resuelve el problema siguiente:

[13] Fuente de la imagen:Gareth James, Daniela Witten, Trevor Hastie, Robert Tibshirani, 2013 - *An introduction to Statistical Learning*

$$\hat{\beta}^{\text{lasso}} = \underset{\beta}{\operatorname{argmin}} \sum_{i=1}^{N} \left(y_i - \beta_0 - \sum_{j=1}^{p} x_{ij}\beta_j \right)^2$$

$$\text{subject to } \sum_{j=1}^{p} |\beta_j| \leq t.$$

Fig. 2.4.5: función objetivo de Lasso[14]

Como se ve tan solo ha cambiado la función de restricción respecto al método de Ridge visto anteriormente. De nuevo, se puede reescribir el modelo en forma Lagrangiana

$$\hat{\beta}^{\text{lasso}} = \underset{\beta}{\operatorname{argmin}} \left\{ \frac{1}{2} \sum_{i=1}^{N} (y_i - \beta_0 - \sum_{j=1}^{p} x_{ij}\beta_j)^2 + \lambda \sum_{j=1}^{p} |\beta_j| \right\}$$

Fig. 2.4.6: función objetivo Lagrangiana de Lasso

Se observa como de nuevo el grado de regularización está controlado por el parámetro λ. Cuanto más grande sea este más se contraerán los coeficientes y más variables se eliminarán.

En definitiva, la principal ventaja de este método radica es que no solo contrae los parámetros sino que también hace una selección de variables, haciendo que los modelos sean más interpretables.

- Interpretación geométrica

A continuación, se hace una representación gráfica de cómo se definen las coeficientes del modelo en la regresión de Lasso. Esto resulta muy útil para entender porque este método puede resultar en estimaciones de coeficientes exactamente iguales a cero:

[14] Fuente:Trevor Hastie, Robert Tibshirani, Jerome Friedman, 2004 - *The Elements of Statistical Learning*

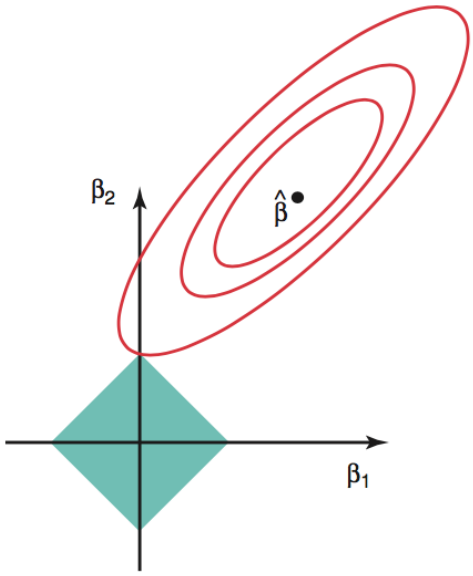


Fig.2.4.7: ilustración del error y de la función de restricción para la regresión de Ridge[15]

De nuevo la solución de la regresión lineal por MCO aparece como $\hat{\beta}$ y cada elipse roja representa una combinación de regresores donde la suma de residuos cuadrados es la misma. En cambio, esta vez la restricción tiene forma de cuadrado ($|\beta_1| + |\beta_2| \leq t$). Se comprueba otra vez que si *t* es lo suficientemente grande las estimaciones de Lasso serán las mismas que $\hat{\beta}$. En caso contrario las estimaciones serán las correspondientes al punto de contacto entre una de las elipses y una 'esquina' de la zona de restricción. Al encontrarse estas esquinas siempre sobre los ejes, la intersección será a menudo sobre uno (o varios[16]) de los ejes. Cuando esto ocurre uno (o varios) de los coeficientes será igual a cero.

- Comparación Ridge y Lasso

Resumiendo lo dicho en los apartados anteriores, se puede decir que ambos modelos han sido diseñados para evitar el sobreajuste a la hora de modelizar una regresión lineal por mínimos cuadrados ordinarios. Para ello, ambos reducen el valor de los coeficientes, disminuyendo así la varianza pero aumentando el sesgo. Se diferencian en que Lasso consigue que algunos coeficientes sean exactamente cero, por lo que este método también realiza una selección de variables. Esto resulta una clara ventaja en situaciones donde no todos los regresores son importantes y se desea eliminar los menos influyentes.
Por otra parte, Ridge puede resultar más adecuado en situaciones donde las variables predictoras estén muy correlacionadas entre ellas. Ridge reduce la influencia de todas estas variables a la vez, mientras que Lasso selecciona una de ellas, dándole toda la importancia y eliminando el resto. El problema radica en que bajo altas correlaciones, esta selección de

[15] Fuente de la imagen:Gareth James, Daniela Witten, Trevor Hastie, Robert Tibshirani, 2013 - *An introduction to Statistical Learning*

[16] este hecho también es válido en problemas de mayores dimensiones, dónde la intersección será sobre varios ejes simultáneamente

variables puede ser muy inestable bajo pequeños cambios en los datos de entrenamiento, arrojando así resultados poco consistentes.
Otro inconveniente de Lasso es que si se tienen más predictores que observaciones ($p > n$), Lasso solo seleccionará como máximo n variables.
Como vemos puede resultar difícil elegir entre un método y el otro. Por suerte, y como se ve en el siguiente apartado, existe un método llamado ElasticNet que combina ambos métodos.

2.4.4. ElasticNet

ElasticNet surge como solución a los problemas de Lasso, que como se ha dicho hace una selección de variables que puede ser demasiado dependiente de los datos, y por lo tanto inestable. Para ello combina las penalizaciones L1 y L2 de Lasso y Ridge. De nuevo se tienen dos fórmulas equivalentes:

$$\hat{\beta} = \underset{\beta}{\operatorname{argmin}} \left\{ \sum_{i=1}^{N} \left(y_i - \sum_{j=1}^{p} x_{ij}\beta_j \right)^2 \right\} \text{ s.t. } (1-\alpha) \sum_{j=1}^{p} |\beta_j| + \alpha \sum_{j=1}^{p} \beta_j^2 \leq t$$

Fig. 2.4.8: función objetivo ElasticNet[17]

y

$$\hat{\beta} = \underset{\beta}{\operatorname{argmin}} \left\{ \sum_{i=1}^{N} \left(y_i - \sum_{j=1}^{p} x_{ij}\beta_j \right)^2 + \lambda_1 \sum_{j=1}^{p} |\beta_j| + \lambda_2 \sum_{j=1}^{p} \beta_j^2 \right\}$$

Fig. 2.4.8: función objetivo Lagrangiana ElasticNet

Donde $\alpha = \frac{\lambda_2}{\lambda_1 + \lambda_2}$

Como vemos la restricción es ahora una combinación lineal de las restricciones de Lasso y de Ridge. El parámetro $0 \leq \alpha \leq 1$ será el que asigne más importancia a L1 o a L2. Es decir, cuanto más cercano a 1 sea más peso se dará a la penalización de Ridge, y cuanto más cercano a 0 más peso se le dará a la penalización de Lasso (cuando $\alpha = 0$ se aplica Lasso y cuando $\alpha = 1$ Ridge).
Combinar ambos métodos puede resultar en una combinación de lo mejor de cada uno:

- selección de variables
- reducción de coeficientes
- alienta el *grouping effect* (cuando hayan variables muy correlacionas ya no se conservará tan solo una de ellas)
- elimina la limitación en el número de variables seleccionadas cuando $p > n$

[17] Fuente: Hui Zou, Trevor Hastie, 2003 - *Regularization and Variable Selection via the Elastic Net*

- Interpretación geométrica

A continuación se intenta dar una explicación algo más visual a este método. Para ello se visualizan las tres funciones de restricción de un modelo con dos variables predictoras:

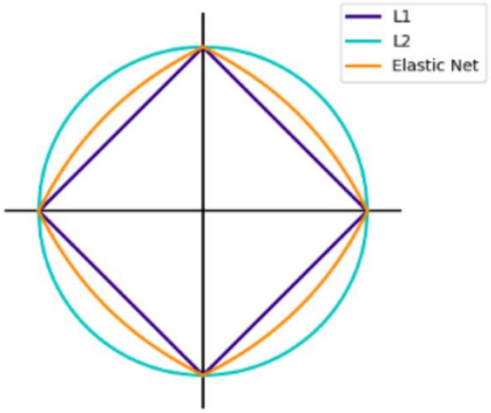


Fig. 2.4.9: ilustración de la función de restricción de Lasso, Ridge y ElasticNet

Como se observa, la penalización de ElasticNet se encuentra entre las de Lasso (L1) y Ridge (L2). Esto hace que su función de restricción muestre singularidad en los vértices (como Lasso), y que sea convexa entre cada vértice (como Ridge). Esta convexidad dependerá del valor de α. La singularidad en los vértices implica que se sigue haciendo una selección de variables, y la convexidad consigue que no se seleccione tan solo una variable cuando exista alta correlación entre variables predictoras (*grouping effect*).

- Tabla resumen de los 3 métodos vistos hasta ahora:

Modelo	Penalización	Características principales
Ridge	$\lambda \sum_{j=1}^{p} \beta_j^2$	• reducción de coeficientes
Lasso	$\lambda \sum_{j=1}^{p} \lvert\beta_j\rvert$	• reducción de coeficientes • selección de variables
ElasticNet	$\lambda_1 \sum_{j=1}^{p} \lvert\beta_j\rvert + \lambda_2 \sum_{j=1}^{p} \beta_j^2$	• combinación de Ridge y Lasso

Fig. 2.4.10: tabla resumen Lasso, Ridge y ElasticNet

2.4.5. *Group* Lasso

El método *Group* Lasso permite que un conjunto de grupos de variables predefinidos sean incluidos o no en el modelo. Existen situaciones en que ciertas variables predictoras tienen una estructura natural agrupada (común en bioestadística, análisis de bolsa, etc). Definiendo grupos se puede así conseguir que este método haga que grupos enteros de variables tenga o no coeficiente cero. Es decir, mientras que Lasso elimina variables individualmente, *Group* Lasso elimina grupos de variables[18] . Por supuesto este método también reduce el resto de coeficientes. Por lo tanto, *Group* Lasso puede ser visto como la aplicación de Lasso entre grupos y de Ridge intra-grupos.

La penalización de este modelo será igual a la suma ponderada de cuadrados de los coeficientes pertenecientes al mismo grupo (esta penalización será de nuevo más o menos importante dependiendo del parámetro λ):

$$\hat{\beta}^{Group\ lasso} = argmin\left\{\sum_{i=1}^{N}(y_i - \beta_0 - \sum_{j=1}^{p} x_{ij}\beta_j)^2 + \lambda \sum_{l=1}^{K} \sqrt{p_l}\left\|\beta^{(l)}\right\|_2\right\}$$

Donde:
K = nº de grupos
p_l = tamaño del grupo l

- Interpretación geométrica

[18] Nótese que si todos los grupos son de tamaño 1 será equivalente al modelo de Lasso.

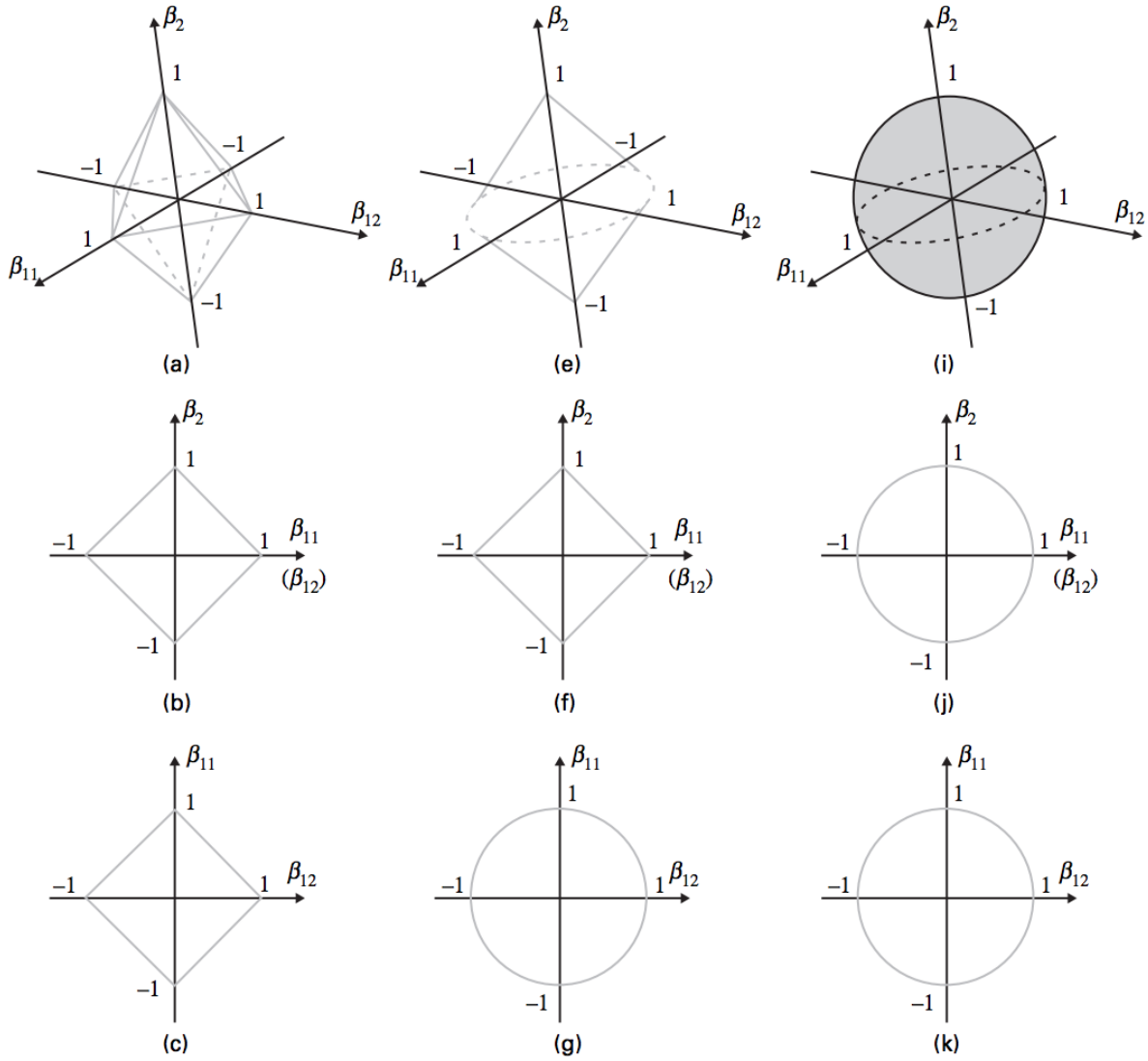


Fig. 2.4.12: penalización de Lasso (a)-(c), Group Lasso (e)-(g) y Ridge (i)-(k)[19]

En la imagen superior se representan las funciones de penalización de un modelo con 3 coeficientes, β_1, β_{11} y β_{12} para los modelos de Lasso, *Group* Lasso y Ridge. Los coeficientes β_{11} y β_{12} forman un grupo. En la figura (f) se ve como *Group* Lasso se comporta como Lasso entre grupos. En la figura (g) en cambio, se ve como el modelo se comporta como Ridge con las variables de un mismo grupo.

[19] Fuente: M. Yuan, Y. Lin, 2004 – Model selection and estimation in regression with grouped variables

2.4.6. Regresión por reducción de dimensionalidad

a. Regresión por componentes principales

La regresión por componentes principales (PCR) consiste en realizar un análisis de componentes principales sobre las variables predictoras, para a posteriori usar algunas[20] de las componentes generadas para ajustar un modelo de regresión lineal por mínimos cuadrados ordinarios.

PCR se usa principalmente por dos razones. La primera es reducir el número de variables predictoras. La segunda es deshacerse de la multicolinealidad entre variables. Es decir, al no estar las componentes correlacionadas entre sí, se ayuda a prevenir los posibles errores causados por usar regresores muy correlacionados.
Este método puede por lo tanto ser visto como un método de regularización, y es por consiguiente a menudo usado como técnica de prevención del sobreajuste a la hora de hacer predicciones.

Existen varios métodos a la hora de elegir el número de componentes principales que se quieren conservar para ajustar el modelo. Uno es elegir las primeras x componentes que conjuntamente expliquen una cantidad mínima de varianza deseada. Otro puede ser calcular alguna métrica de error mediante validación cruzada para cada cantidad de componentes, y escoger la cantidad de componentes que minimicen el error.

Una posible desventaja de PCR es que no hay garantía de que las componentes elegidas expliquen la variable respuesta (se pueden haber eliminado las componentes que sí lo hacían).

b. Regresión por mínimos cuadrados parciales

La regresión por mínimos cuadrados parciales, o *Partial Least Squares* (PLS), es una técnica muy similar a la de PCR. Ambos métodos emplean como regresores las componentes principales de las variables originales, y ambos pueden ser usados para resolver problemas de multicolienalidad entre predictores[21] . La diferencia es que PCR no tiene en ningún momento en cuenta la variable respuesta para construir las componentes principales, es un método de aprendizaje no supervisado. En cambio, a la hora de determinar las combinaciones lineales de las componentes, PLS busca aquellas combinaciones que no solo expliquen la varianza total observada, sino que predigan también la variable respuesta lo mejor posible. En otras palabras, PCR se preocupa por capturar la mayoría de la varianza de los predictores, sin tener en cuenta la varianza relacionada con la predicción de la variable respuesta (o acabando esta en alguna de la últimas componentes). PLS en cambio, crea componentes (a menudo llamadas variables latentes) que maximizan la covarianza entre los regresores y la respuesta. Es por lo tanto un método de aprendizaje supervisado.

[20] nótese que al ser las componentes combinaciones lineales de todos los predictores, si se usasen todas estas componentes se obtendría el mismo resultado que aplicando una regresión lineal

[21] otra ventaja es que estos métodos pueden usarse incluso cuando p>n, cosa que no se puede hacer en la regresión lineal por MCO

Aunque en este trabajo no se profundiza en estos, existen varios algoritmos para obtener las componentes de este modelo (e.g. kernel algorithm[22], wide kernel algorithm[23], SIMPLS[24]).

Cabe comentar que tanto PCR como PLS suelen ser utilizados en campos diferentes al que se estudia aquí. Suele ser utilizado en campos que tratan con muchas variables, y muy correlacionadas entre sí, como en las industrias química, de medicamentes o de alimentos.

2.5. Otros conceptos

2.5.1. Train/Test Split

El *train/test Split* o partición entrenamiento/test, común en todos los métodos de aprendizaje supervisado, consiste en dividir el conjunto de datos en dos partes: la parte entrenamiento, que servirá para entrenar el modelo y la parte *test*, que servirá para evaluar el modelo entrenado.

2.5.2. Validación Cruzada

La validación cruzada o *cross-validation* es una técnica para asegurar que los resultados de un análisis estadístico sean independientes de la partición entrenamiento/test. Es decir, a la hora de por ejemplo modelizar un algoritmo de predicción, los parámetros de este pueden verse demasiado influenciados por como se ha hecho la partición aleatoria de los datos, por lo que se busca reducir esta variabilidad.
Se suele aplicar el método conocido como *K-fold cross-validation*:

a. se dividen los datos en K subconjuntos aleatorios de mismo tamaño
b. uno de lo subconjuntos se usa como datos de prueba y el resto se usan para entrenar el modelo
c. se calcula el error (o métrica que se desee) a partir de esa partición
d. se repiten los pasos b. y c. hasta que todos los subconjuntos hayan sido una vez usados como datos de prueba
e. se calcula la media aritmética de los errores

Una extensión de este método es el *Leave-one-out cross-validation* (LOOCV). Es básicamente como el *K-fold cross-validation* pero creando tantos subconjuntos como observaciones hayan. Se empleará así solo una observación para calcular el error, por lo que este será muy variable. Pero repitiendo este proceso tantas veces como observaciones hayan se consigue eliminar la variabilidad que surge de dividir aleatoriamente los datos. Este método puede tener un coste computacional muy alto si se trabaja con muchos datos.

---

[22] Fredrik Lindgren, Paul Geladi, Svante Wold, 1993 – *The kernel algortihm for PLS*

[23] Rännar et al.

[24] De Jong, S., 1993 - *SIMPLS: an alternative approach to partial least squares regression*

2.5.3. Métricas de error

Las métricas de error sirven para evaluar la capacidad predictiva del modelo. Son estas las que nos dicen como de bien funciona nuestro modelo. En este trabajo se utilizan 3 diferentes para comparar y evaluar los modelos:

- RMSE

  Del inglés *root-mean-squared error*, el RMSE es una magnitud media del error. Es la raíz cuadrada de la media de las diferencias al cuadrado entre valores reales y predichos:

$$\sqrt{\frac{\sum_{i=1}^{N}(Predicción_i - Real_i)^2}{N}}$$

- MAE

  Del inglés *mean-absolute error*, el MAE es también una magnitud media del error. Es la media de las diferencias absolutas entre valores reales y predichos:

$$\sum_{i=1}^{N}\frac{|Predicción_i - Real_i|}{N}$$

- Coeficiente de determinación ajustado

Se llama R^2 al coeficiente de determinación. Este mide la proporción de varianza en la variable dependiente que es predecible por las variables independientes, o dicho en otras palabras: el coeficiente de determinación es la proporción de la varianza total de la variable respuesta explicada por la regresión. Su formula es la siguiente:

$$R^2 = \frac{\sum_{i=1}^{N}(Predicción_i - \overline{Real})^2}{\sum_{i=1}^{N}(Real_i - \overline{Real})^2}$$

Como se puede ver este valor oscila entre 0 y 1. Cuanto más cerca de 1 esté mejor será el ajuste del modelo a la variable dependiente. Pero este coeficiente conlleva un gran fallo, y es que a medida que se añaden variables su valor aumentará, aunque estas no sean significativas.

Es por eso que se debe utilizar el coeficiente de determinación ajustado R^2_{adj}. En efecto, este coeficiente penaliza la inclusión de variables, por lo que resulta mucho más apropiado si se va a utilizar más de una variable explicativa.

$$\mathrm{R}^2_{adj} = 1 - \frac{N-1}{N-p-1}[1-R^2]$$

3. Análisis exploratorio de datos

Se procede en este apartado a realizar aquellos pasos que, aunque no obligatorios, se recomiendan altamente llevar a cabo en este tipo de estudio antes de comenzar a implementar métodos y técnicas de predicción. Como indica el título se exploran y analizan los datos disponibles, esto nos permitirá entender mejor los datos. Es decir, entenderemos mejor en que se diferencian, en que se parecen, y que características tienen los diferentes países de la base de datos.

3.1. Preprocesamiento de los datos

- Tratamiento de *missings*

Aunque existen métodos de predicción que aceptan datos faltantes o *'missings',* es aconsejable tratar con ellos, ya sea eliminándolos o imputándolos. Antes de todo veamos gráficamente en que situación se encuentra nuestro *dataset:*

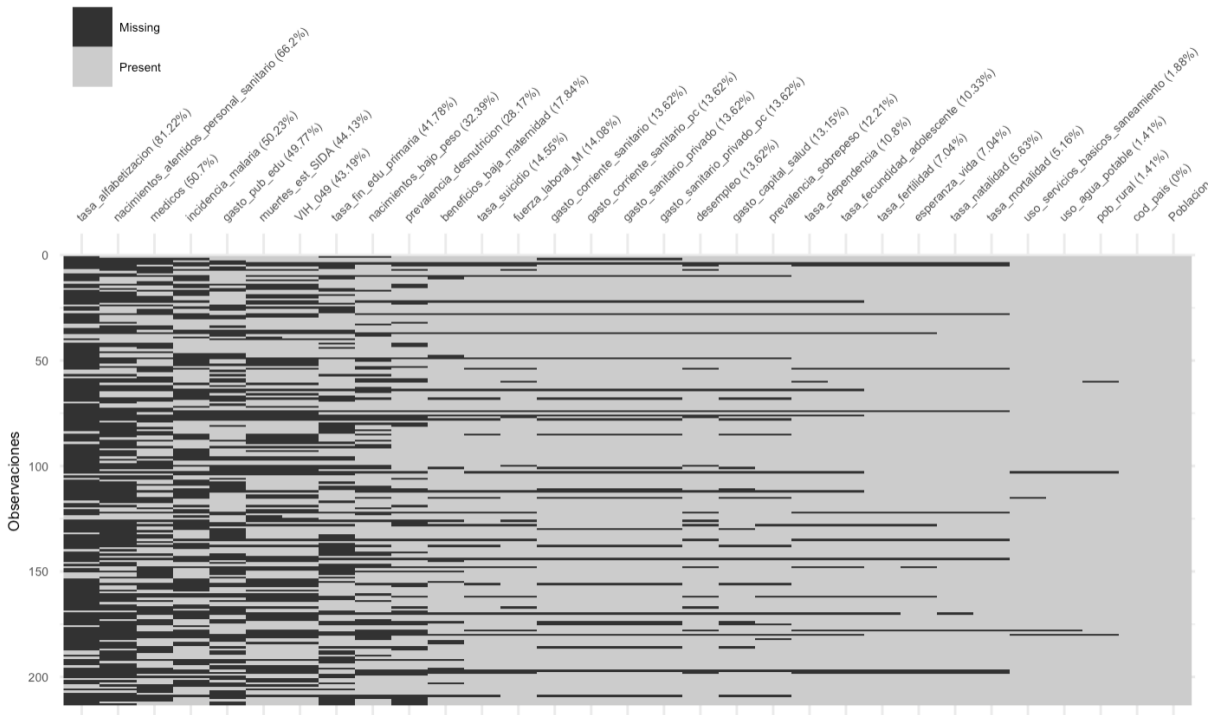


Fig. 3.1.1: Mapa de *Missings*

Los datos faltantes son considerables en algunas variables y en algunos países. Antes de estudiar métodos alternativos de imputación se procede a eliminar directamente ciertas variables y países a los que les falta demasiada información.
Primeramente se eliminan aquellos países que no disponen de información en la variable respuesta utilizada a lo largo del trabajo ('*esperanza_vida*'[25]). Esto es necesario ya que no se

[25] más adelante se explica la nomenclatura y el significado de cada variable

podrá entrenar ni validar ningún modelo cuya variable respuesta no esté completa. Los primeros países descartados son: Samoa Americana, Andorra, Islas Vírgenes Británicas, Islas Caimán, Dominica, Gibraltar, Islas Marshall, Mónaco, Nauru, Islas Mariana del Norte, Palaos, San Marino, San Cristóbal y Nieves, Islas Turcas y Caicos y Tuvalu.
Como comentario decir que sin duda resulta curioso ver que el Banco Mundial no disponga de los datos de esperanza de vida de países del primer mundo como Andorra, Mónaco o San Marino. Aún así todos los países son pequeñas islas o pequeños territorios por lo que no extraña tanto verlos aquí.
Seguidamente, se eliminan aquellos países cuya falta de información supere el 50% (que falten datos en más de la mitad de las variables). Estos países son: Aruba, Bermudas, Curazao, Islas Feroe, Polinesia Francesa, Groenlandia, Guam, Kosovo, Liechtenstein, Macao, Nueva Caledonia, Puerto Rico, San Martín, Islas Vírgenes.
Finalmente se decide eliminar aquellas variables que sigan conteniendo menos del 50% de los datos. Estas son '*nacimientos_atentidos_personal_sanitario'*, la cual mide el porcentaje de nacimientos que son atendidos por personal sanitario cualificado, y 'tasa_alfabetización'.

Se decide imputar ahora el resto de datos faltantes, esto quiere decir que se reemplaza el dato faltante por una estimación/adivinanza de este. Existen muchas técnicas de imputación, se puede por ejemplo reemplazar este valor por la media, mediana o moda de la variable en cuestión. Este método tan solo tiene en cuenta la distribución de la variable con datos faltantes, por lo que se suele aconsejar métodos alternativos, como la imputación por regresión o por 'k vecinos más próximos' que tengan en cuenta la relación de las variables y de los individuos entre sí. Este último método de los 'k vecinos más próximos', o K-NN (del inglés *K-nearest neighbors*) será el utilizado en este caso.

Es un método habitualmente usado como algoritmo de predicción que consiste en:

i. Calcular la distancia[26] entre el ítem a calcular y el resto de ítems ya conocidos (i.e. la distancia entre el país cuyo valor *missing* se quiere imputar y el resto de países)
ii. Seleccionar los k individuos[27], o países, más cercanos al ítem a calcular
iii. Asignar el valor medio o la mediana al ítem que se quería predecir (se asignaría el valor más frecuente si fuese un problema de clasificación)

Es decir, en este caso, se calculan los k países más 'cercanos' o parecidos al país cuya información faltante se quiere completar y se sustituye el valor faltante por la media de esos países en esa variable. En este trabajo se ha implementado esta imputación mediante la función *knnImputation()* del paquete de R 'DMwr'.

- Selección y transformación de variables

[26] Existen varias métricas de distancia entre individuos como la Euclidea, la de Mahapolnis o la de Manhattan. Se ha utilizado aquí la distancia Euclidea estandarizada

[27] se ha cogido k = 6

A menudo resulta conveniente hacer una selección o transformación de las variables (también conocido como *Feature Engineering*), ya sea mediante algún tipo de análisis de datos o por juicio y conocimiento propio.
Relativo a este *dataset,* se decide primeramente ponderar las variables '*VIH_049*' y '*muertes_est_SIDA'* en función de la población. Seguidamente se crea la variable '*tasa_crecimiento_anual*' (la cual indica la tasa a la que ha crecido un país en un año dado), a partir de las variables 'tasa_natalidad' y 'tasa_mortalidad'[28]. Y se eliminan estas dos últimas variables. De esta manera se obtiene una variable que reúne información de otras dos y que puede resultar más interesante.
Por otra lado, estudiando la correlación entre variables[29], se decide eliminar la variable '*VIH_049*' al tener una correlación por encima de 0.95 con la variable '*muertes_est_SIDA*'. Aunque existen otras variables con alta correlación entre sí (por encima de 0.9 incluso) se ha decidido mantenerlas, ya que esto nos permitirá estudiar el comportamiento de los modelos de regresión estimados más adelante en presencia de (multi)colinealidad.
Finalmente, tenemos dos juegos de variables muy parecidos, 'gasto_corriente_sanitario' y 'gasto_corriente_sanitario_pc', y 'gasto_sanitario_privado' y 'gasto_sanitario_privado_pc'. Se diferencian únicamente en que unas se miden en relación al PIB y las otras en términos per cápita. Se eliminan las variables en términos per cápita.

3.2. Análisis descriptivo

Antes de seguir con el análisis se presenta una tabla con la descripción y unidad de medida de cada variable de la base de datos final:

Variable	**Descripción**	**Unidad de medida**
cod_pais	Código del País	
tasa_fertilidad	Tasa de fertilidad	hijos por mujer
tasa_fecundidad_adolescente	Tasa de fecundidad adolescente	nacimientos por cada 1000 mujeres de 15 a 19 años
nacimientos_bajo_peso	Nacimientos bajo peso	% del total
muertes_est_SIDA	Muertes estimadas por SIDA	por 1000 habitantes
medicos	Médicos	por 1000 habitantes

[28] tasa de crecimiento anual = tasa de natalidad - tasa de mortalidad

[29] véase matriz de correlación en el apartado 3.2.

prevalencia_desnutricion	Prevalencia de la desnutrición	% de la población
prevalencia_sobrepeso	Prevalencia del sobrepeso	% de la población adulta
uso_agua_potable	Uso de agua potable	% de la población
uso_servicios_basicos_saneamiento	Uso servicios básicos de saneamiento	% de la población
incidencia_malaria	Incidencia malaria	casos por año por 100 habitantes en riesgo
gasto_capital_salud	Gasto de capital sanitario	% del PIB
gasto_corriente_sanitario	Gasto corriente sanitario	% del PIB
gasto_sanitario_privado	Gasto privado sanitario	% del gasto corriente sanitario
tasa_suicidio	Tasa de suicidio	por 100000 habitantes
tasa_dependencia	Tasa de dependencia	% de la población en edad laboral
desempleo	Tasa de desempleo	% de la fuerza laboral total
fuerza_laboral_M	Fuerza laboral, mujeres	% de la fuerza laboral total
beneficios_baja_maternidad	Beneficios por baja de maternidad	% del sueldo remunerado
tasa_fin_edu_primaria	Tasa de finalización de educación primaria	% del grupo de edad relevante
gasto_pub_edu	Gasto público en educación	% del PIB
pob_rural	Población rural	% de la población total
tasa_crecimiento_anual	Tasa de crecimiento anual	%

Fig. 3.2.1: Tabla descriptiva de variables

Toda la información que se muestra a continuación está disponible en la página 'EDA' de la *webapp* Shiny anteriormente comentada. Lo que aquí se muestra son simplemente capturas

de pantalla que sirven como ejemplo de lo que se puede encontrar en dicha aplicación, por lo tanto, se recomienda fuertemente al lector que acuda a la aplicación para un análisis mucho más elaborado y completo.

- Análisis univariante

Un análisis univariante nos permite ver características de cada variable independientemente del comportamiento del resto de variables. Se muestra a continuación un análisis para una sola variable a modo de ejemplo (*tasa_fertiliddad*). En la aplicación referida anteriormente se puede obtener esta misma información para el resto de variables.

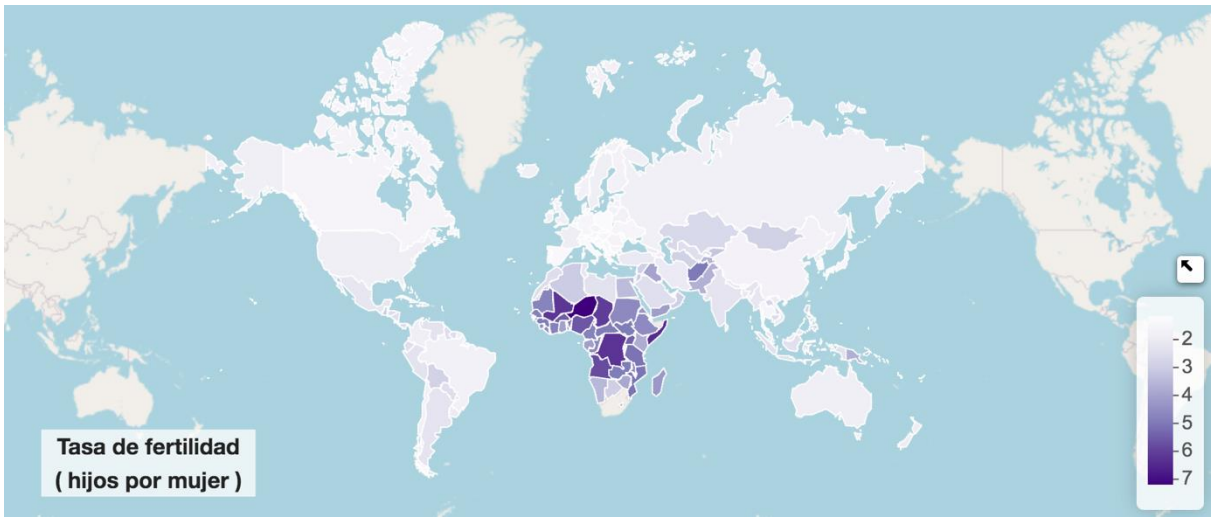
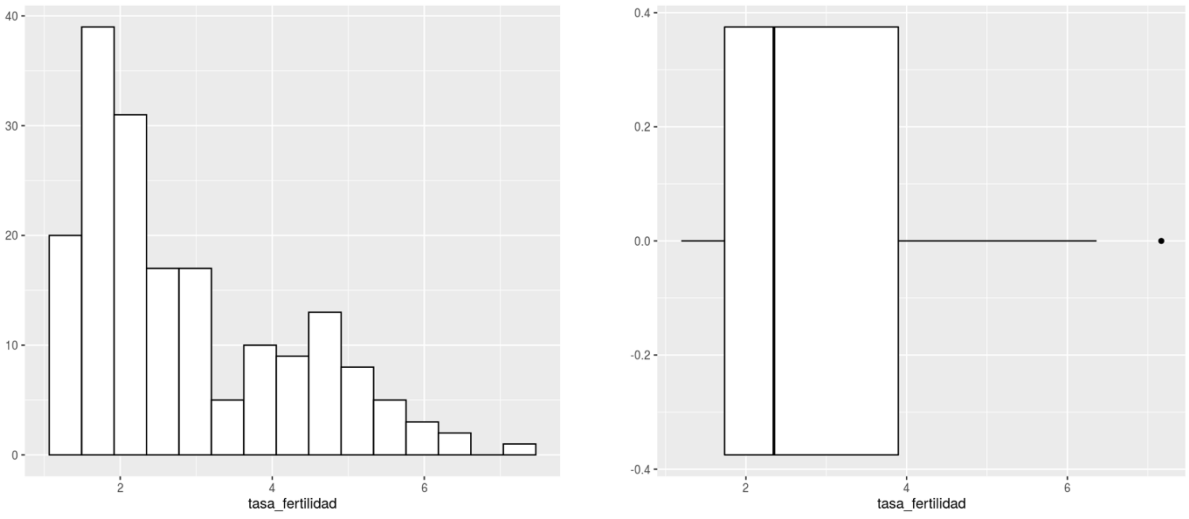


Fig. 3.2.2: Mapa descriptivo de la tasa de fertilidad




```
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
  1.196   1.734   2.348   2.833   3.897   7.169
```

Fig. 3.2.3: Histograma, Boxplot y estadísticos de la tasa de fertilidad

En la memoria no se hace un análisis individual para cada variable, pero una vez observada cada una de ellas, se puede decir que las variables parecen seguir una distribución normal centrada o sesgada hacia algún lado. Tampoco se detecta presencia de *outliers*[30] .

- Análisis multivariante

El análisis multivariante permite ver la relación entre varias variables. Se dispone en la aplicación de un diagrama de dispersión por cada pareja de variables, así como del mismo diagrama por niveles de una tercera variable.

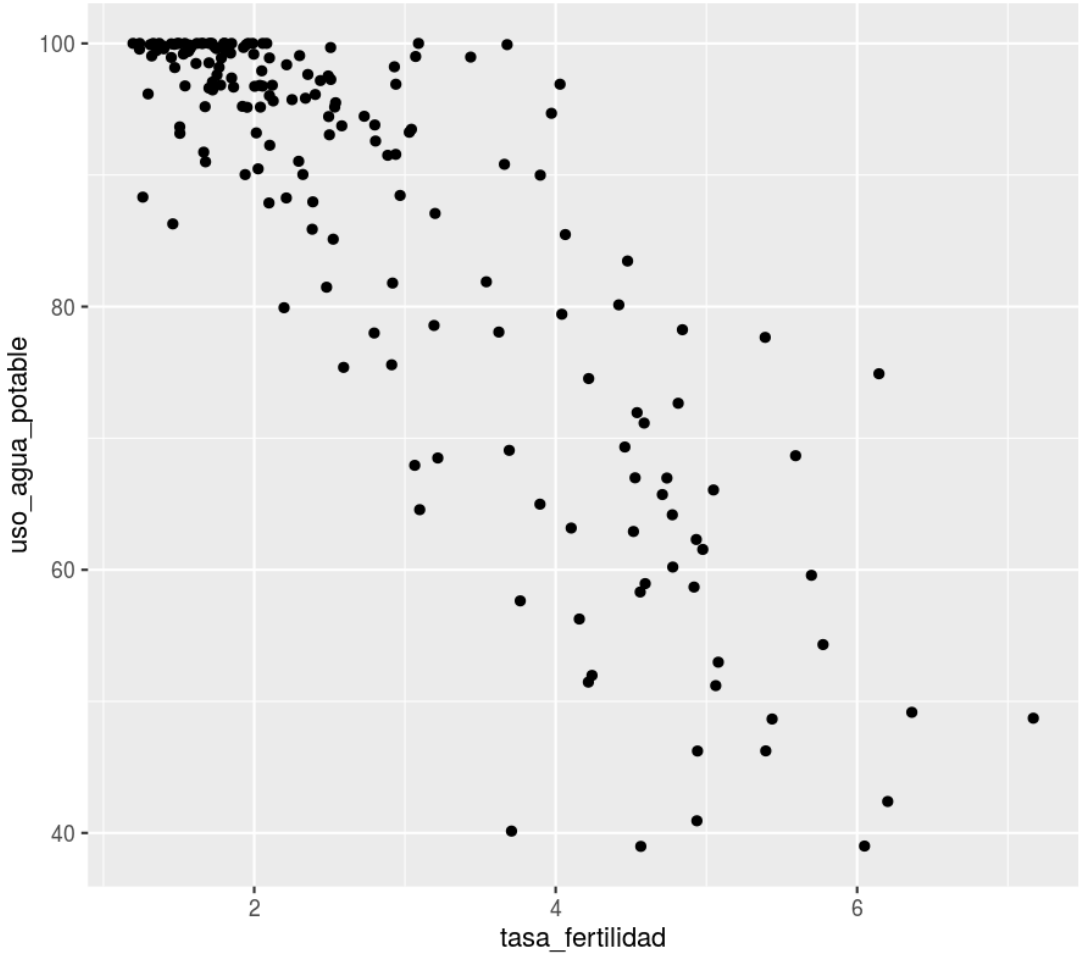


Fig. 3.2.4: Diagrama de dispersión uso_agua_potable vs tasa_fertilidad

[30] Se define como *outlier* a aquel valor que se desvía extremadamente del resto de observaciones.

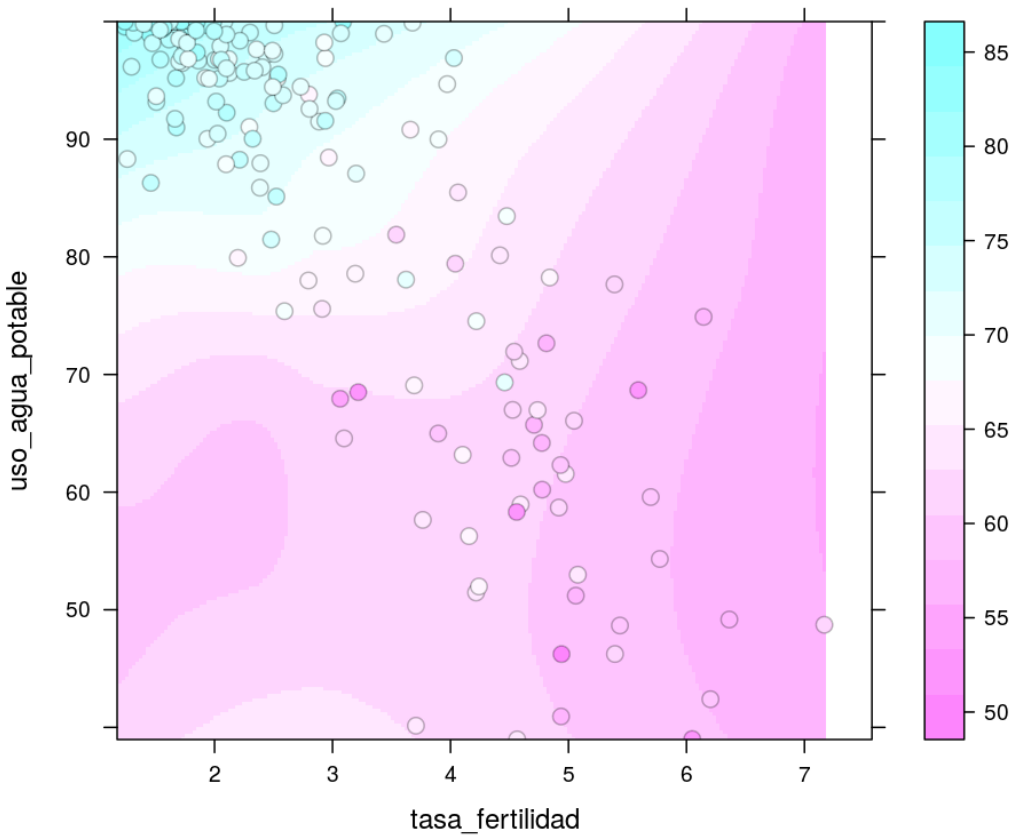


Fig. 3.2.5: Diagrama de dispersión uso_agua_potable vs tasa_fertilidad por niveles de esperanza_vida

Aunque resulta interesante ver las relaciones entre todas las variables, suele ser especialmente interesante observar como es la relación entre la variable respuesta y el resto de variables. Así por ejemplo, de los gráficos anteriores aprendemos que los países con altas tasas de fertilidad y bajo uso de agua potable tienen una esperanza de vida más baja. Conclusiones similares como esta se pueden extraer formando distintas combinaciones de variables.

Esta relación o correlación entre variables puede también visualizarse en forma de *heatmap*:

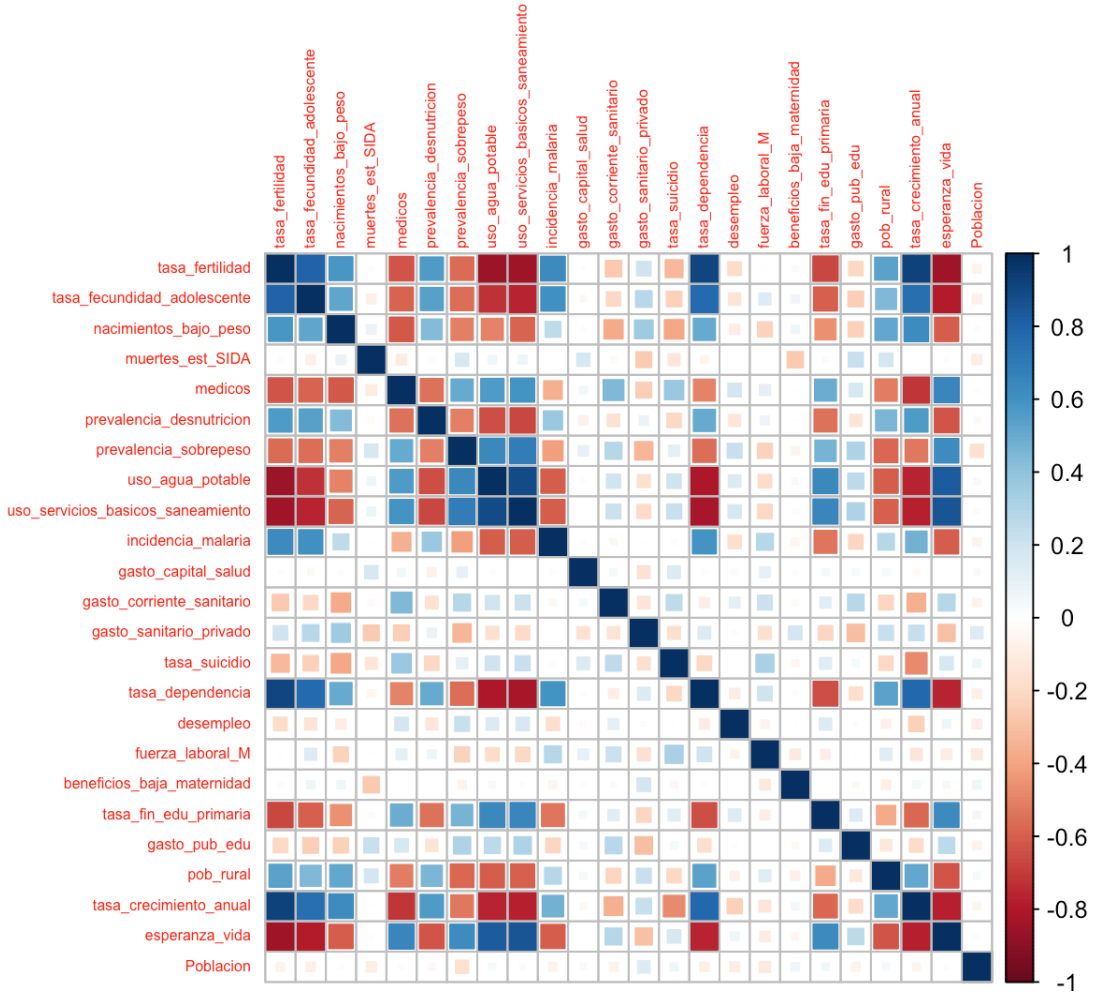


Fig. 3.2.6: *Heatmap* de correlación de variables

De aquí podríamos sacar tanto conclusiones esperadas como no tan esperadas. Se lista a continuación algún ejemplo:

- correlación fuerte negativa entre la esperanza vida y la tasa de fertilidad
- correlación fuerte positiva entre la esperanza de vida y el uso de agua potable
- no correlación entre la incidencia de la malaria y las variables relativas a gastos en sanidad
- no correlación entre la esperanza de vida y los beneficios por baja de maternidad

3.3. Análisis de componentes principales

Se realiza ahora un análisis de componentes principales (véase apartado 2.2.a.). Antes de nada, comentar que de nuevo la mayor parte de información que se encuentra aquí está presente en la aplicación *Shiny.*

La función *prcomp()* nos permite ajustar el análisis directamente (la propia función escala y centra los datos). Tal y como se ha dicho, el objetivo es centrar la mayor parte de la varianza en las primeras componentes. Veamos cuanta varianza explica cada componente:

	PC1	PC2	PC3	PC4	PC5	PC6	PC7	PC8	PC9	PC10	PC11	PC12	PC13
Standard deviation	3.082	1.4665	1.3162	1.1225	1.0422	1.0118	0.9804	0.9014	0.863	0.8093	0.7929	0.7659	0.6943
Proportion of Variance	0.396	0.0896	0.0722	0.0525	0.0453	0.0427	0.0401	0.0339	0.031	0.0273	0.0262	0.0244	0.0201
Cumulative Proportion	0.396	0.4855	0.5576	0.6101	0.6554	0.6981	0.7381	0.7720	0.803	0.8303	0.8565	0.8809	0.9010

	PC14	PC15	PC16	PC17	PC18	PC19	PC20	PC21	PC22	PC23	PC24
Standard deviation	0.639	0.6163	0.5868	0.5548	0.5003	0.46943	0.41934	0.33774	0.28688	0.27529	0.1298
Proportion of Variance	0.017	0.0158	0.0143	0.0128	0.0104	0.00918	0.00733	0.00475	0.00343	0.00316	0.0007
Cumulative Proportion	0.918	0.9338	0.9482	0.9610	0.9715	0.98063	0.98796	0.99271	0.99614	0.99930	1.0000

Fig. 3.3.1: Varianza, % de la varianza, y varianza acumulada de la componentes principales

Se muestra aquí la desviación estándar de cada componente (el cuadrado de este valor se le llama *eigenvalue*), la proporción de varianza que explica[31], así como la proporción acumulada de varianza explicada hasta esa componente.

De esta información se extrae que, como era de esperar, la primera componente recoge la mayor parte de la información, con casi el 40% de la varianza. En cambio, la segunda y tercera no son tan explicativas como se podía esperar (tan solo 9% y 7% de la varianza respectivamente). Habrá que centrarse principalmente en esta primera componente para extraer conclusiones. Aún así se aprecia la utilidad de este método, ya que con tan solo las primeras 9 componentes se obtiene un 80% de varianza acumulada.

Por otro lado, resulta interesante estudiar la correlación entre las componentes principales y las variables. Es decir, como ya se ha mencionado, las componentes son combinaciones lineales de las variables originales, así que de este modo podremos ver de que variables depende esencialmente cada componente principal. Se muestra a continuación esta información en forma de *heatmap* y de gráfico (para las cuatro primeras componentes[32]):

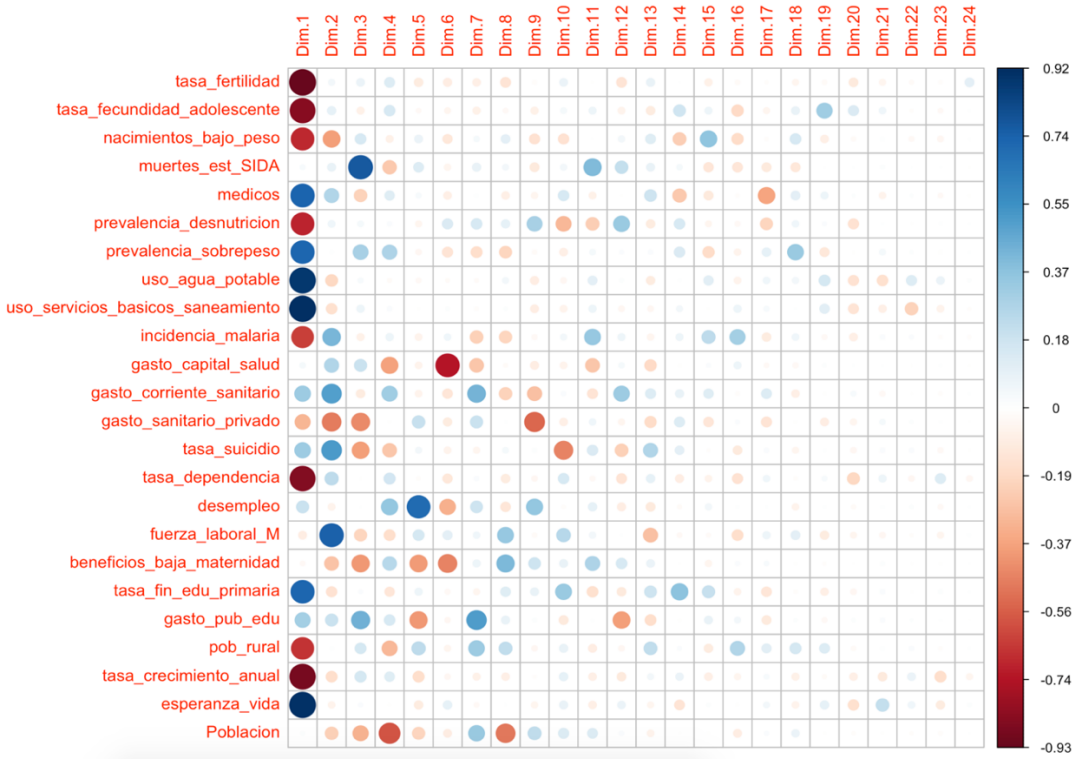


Fig. 3.3.2: *Heatmap* de la correlación entre variables y componentes

[31] La varianza total es igual al número de variables (24 en este caso), dividiendo el *eigenvalue* por este número se obtiene la proporción de varianza

[32] en la aplicación *Shiny* se pueden visualizar todas las componentes

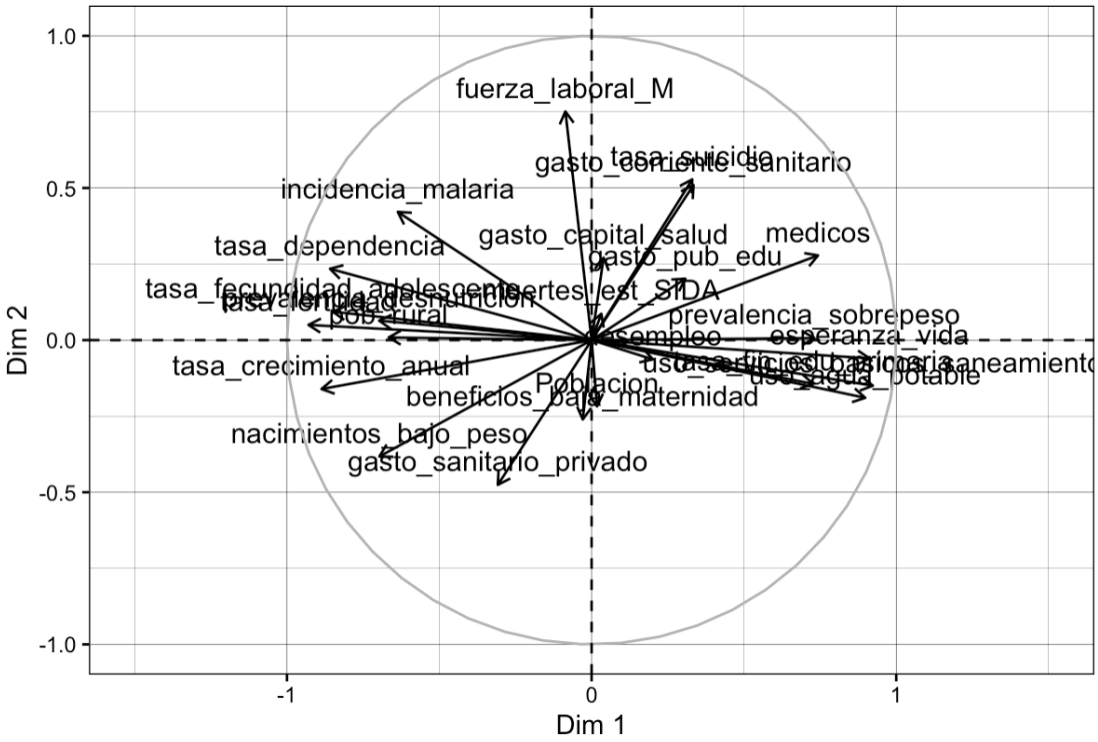


Fig. 3.3.3: proyección de las variables sobre las componentes 1 y 2

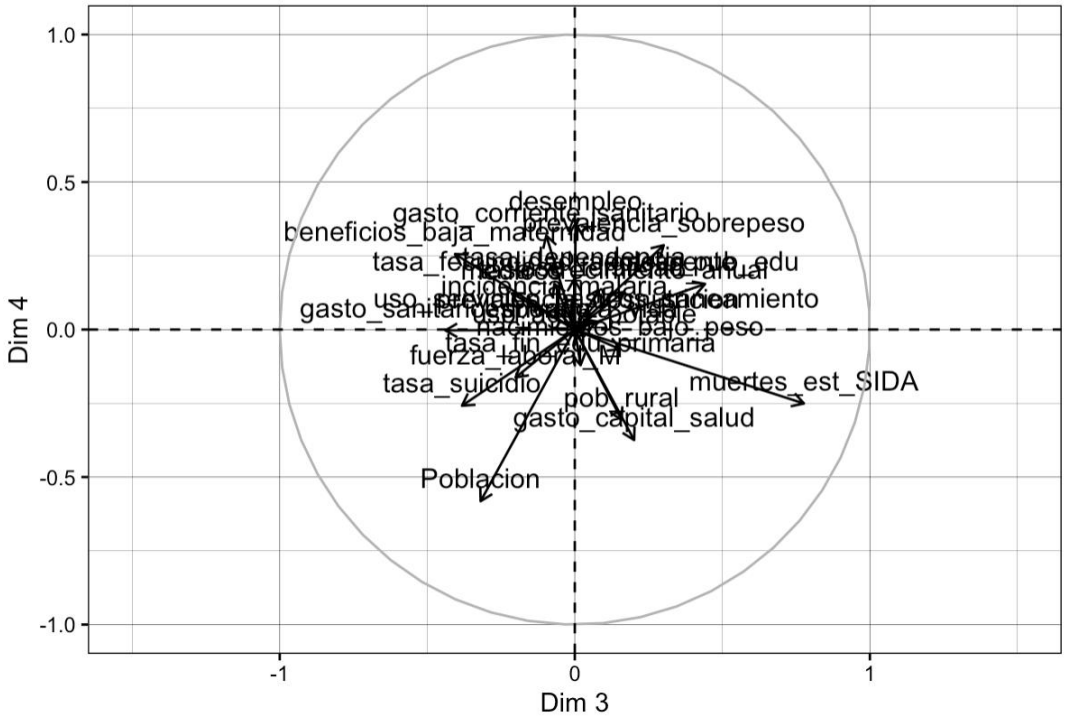


Fig. 3.3.4: proyección de las variables sobre las componentes 3 y 4

Vemos así que la primera componente está muy correlacionada positivamente con las variables médicos, prevalencia_sobrepeso, uso_agua_potable, uso_servicios_basicos_saneamiento, tasa_fin_edu_primaria y esperanza_vida, y correlacionada negativamente con las variables tasa_fertilidad, tasa_fertilidad_adolescente,

nacimientos_bajo_peso, prevalencia_desnutricion, tasa_dependencia, pob_rural y tasa_crecimiento_anual. La segunda componente está correlacionada positivamente con fuerza_laboral_M, tasa_suicidio y gasto_corriente_sanitario, y negativamente con nacimientos_bajo_peso y gasto_sanitario_privado. De esta manera vemos claramente como la primera componente toma valores dependiendo de variables relativas al bienestar, a la sanidad y a la educación. Y la segunda componente se comporta de manera parecida, pero centrándose en otras variables algo más socioeconómicas como la tasa de suicidio o la fuerza laboral femenina. Finalmente en la tercera y cuarta componente únicamente se destaca respectivamente una correlación negativa con la variable *población* y una correlación positiva con *muertes_est_SIDA*.

Por otro lado, hay mirar que valores cogen los países sobre estas componentes[33]:

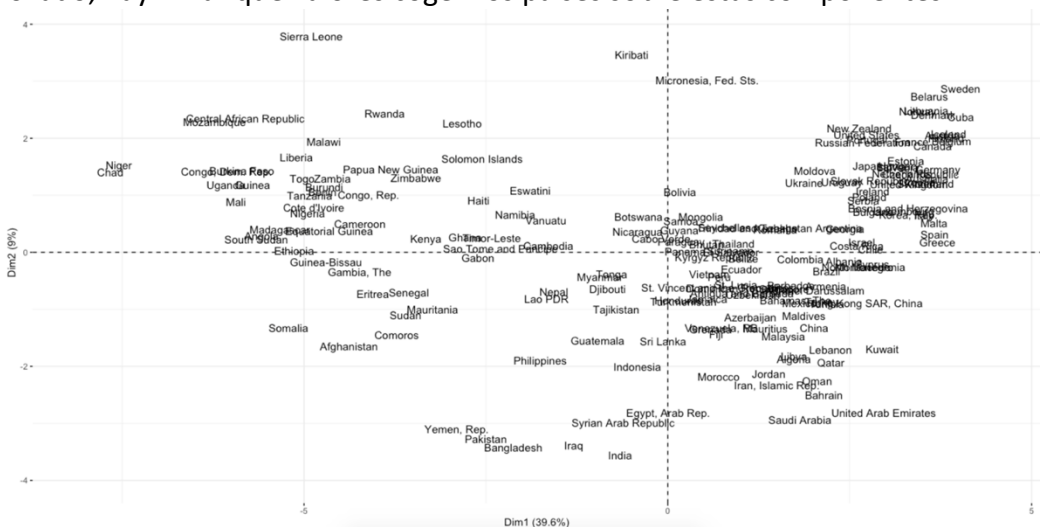


Fig. 3.3.5: Gráfico de individuos sobre las componentes 1 y 2

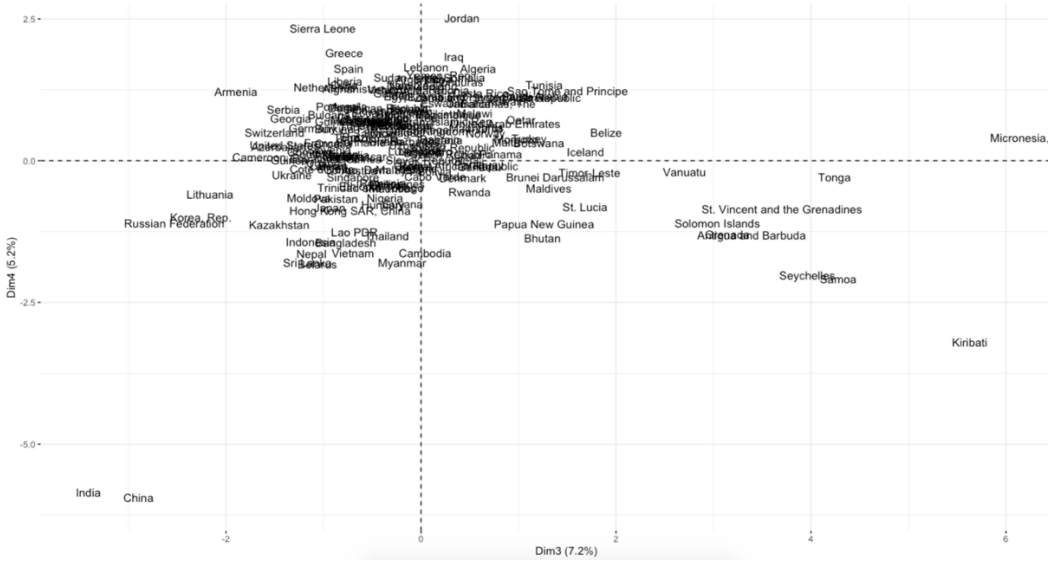


Fig. 3.3.6: Gráfico de individuos sobre las componentes 3 y 4

[33] se recomienda visualizar esta información en la aplicación *Shiny* ya que se puede filtrar por país además de que permite crear grupos en función de cualquier variable:

Esto nos permite ya tener una idea aproximada de que países tienen parecidos y diferencias entre sí. Así por ejemplo podemos decir que en términos de la primera componente el Chad y Suecia parecen ser los países más alejados, o que España, Malta y Grecia son países con características muy similares.
Por otro lado, puede resultar interesante visualizar que niveles cogen los individuos en función de alguna variable. En la aplicación web se puede elegir la variable así como el número de 'cortes'[34] a realizar. Se muestra aquí un ejemplo cogiendo 3 particiones de una variable muy correlaciona con la primera componente como es *tasa_fertilidad* y otra poco correlacionada como es *desempleo*:

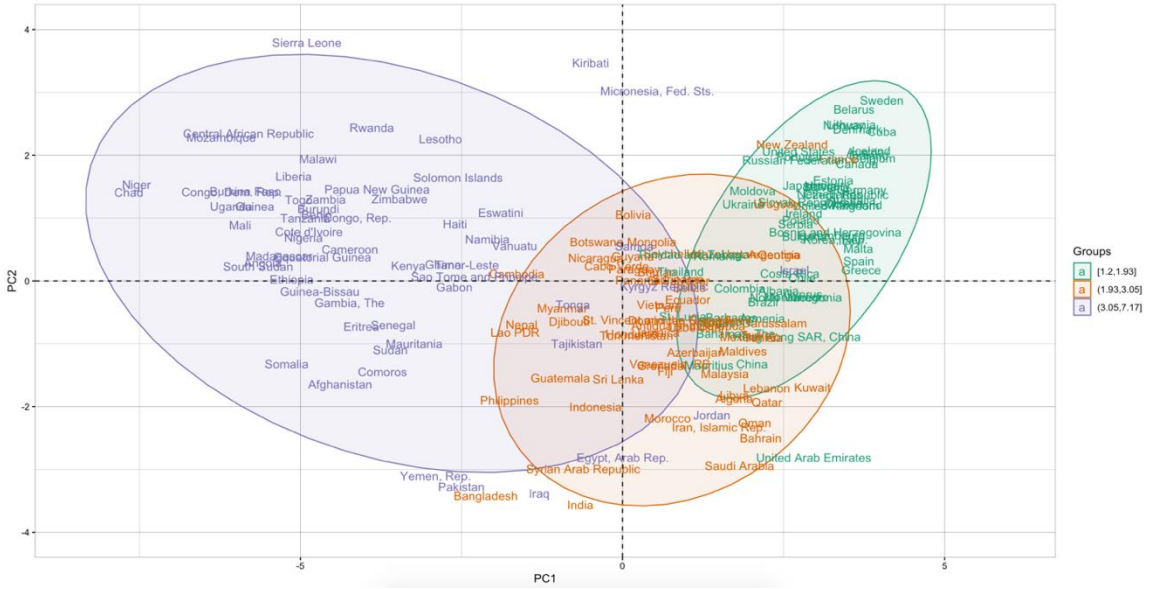


Fig. 3.3.7: Gráfico de países sobre las componentes 1 y 2 segmentado en 3 grupos de la variable tasa_fertilidad

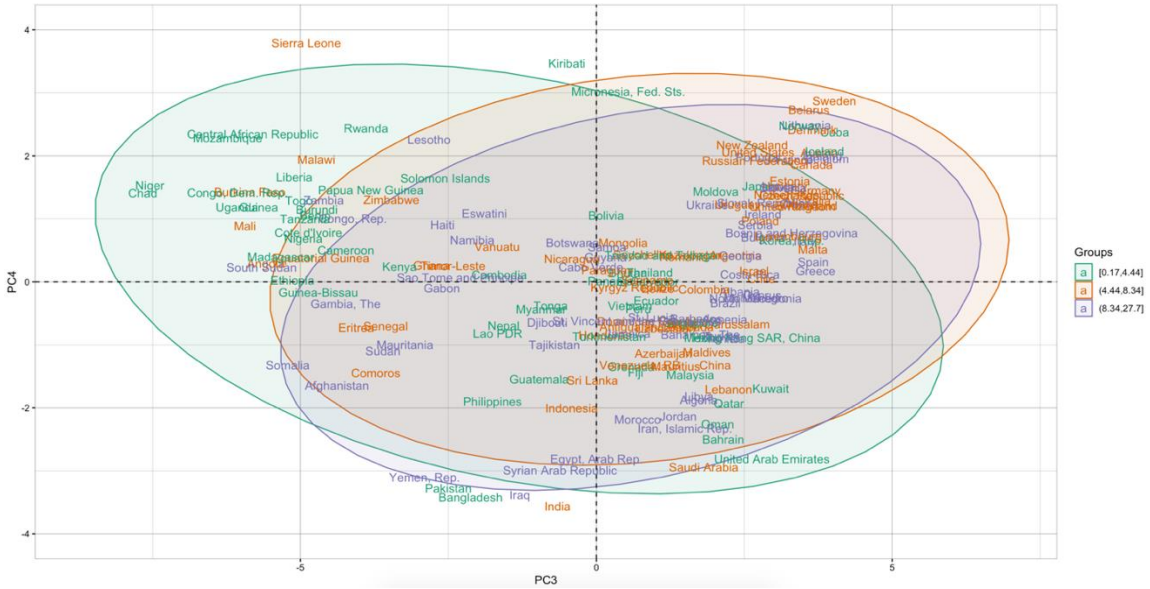


Fig. 3.3.8: Gráfico de países sobre las componentes 3 y 4 segmentado en 3 grupos de la variable desempleo

Se ve claramente en el primer gráfico el efecto de la variable tasa_fertilidad sobre la primera componente, ya que vemos que se han formado 3 grupos muy distintos. En el segundo gráfico

[34] se utilizan los percentiles para realizar las particiones

se ve todo lo contrario, no se puede medir el nivel de desempleo mirando las dos primeras componentes. Se pueden sacar muchas conclusiones como esta, así que se anima al lector a jugar con la variable por la que categorizar así como el número de grupos a hacer.

Por otra parte, esta información debería combinarse con lo dicho anteriormente sobre la correlación entre las variables y las componentes. Es decir, ahora sabemos de que variables depende principalmente cada componente y que valor coge cada individuo sobre cada componente, así que proyectando esta información simultáneamente podremos saber cuales son las características de cada país:

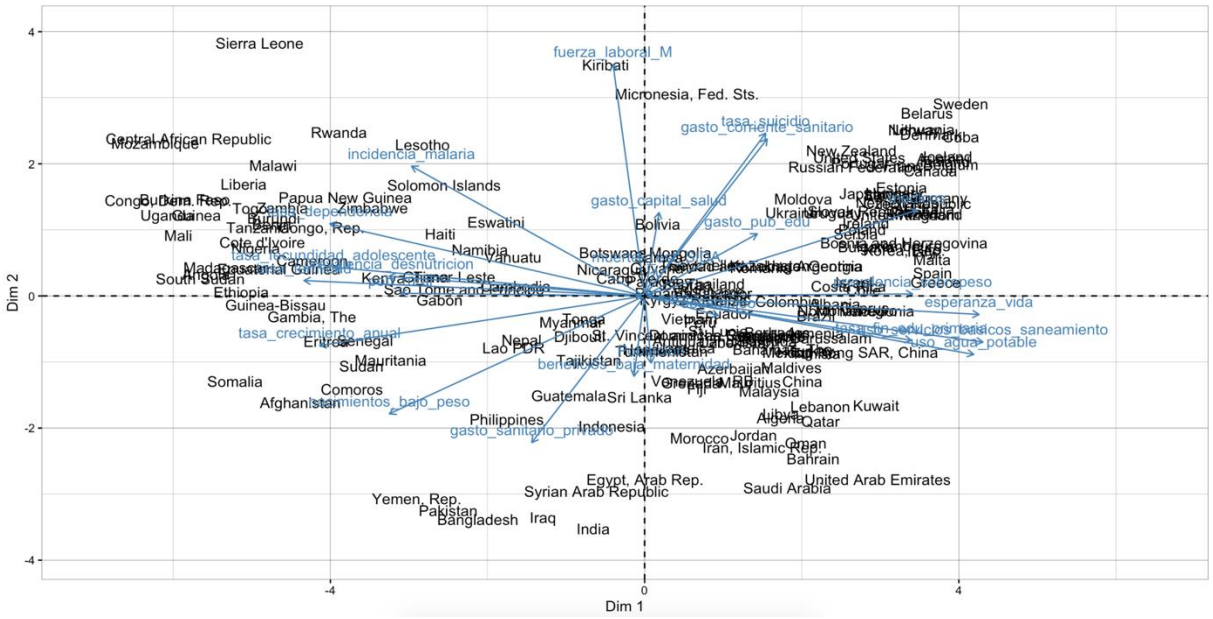


Fig. 3.3.9: *Biplot* de países y variables sobre las componentes 1 y 2

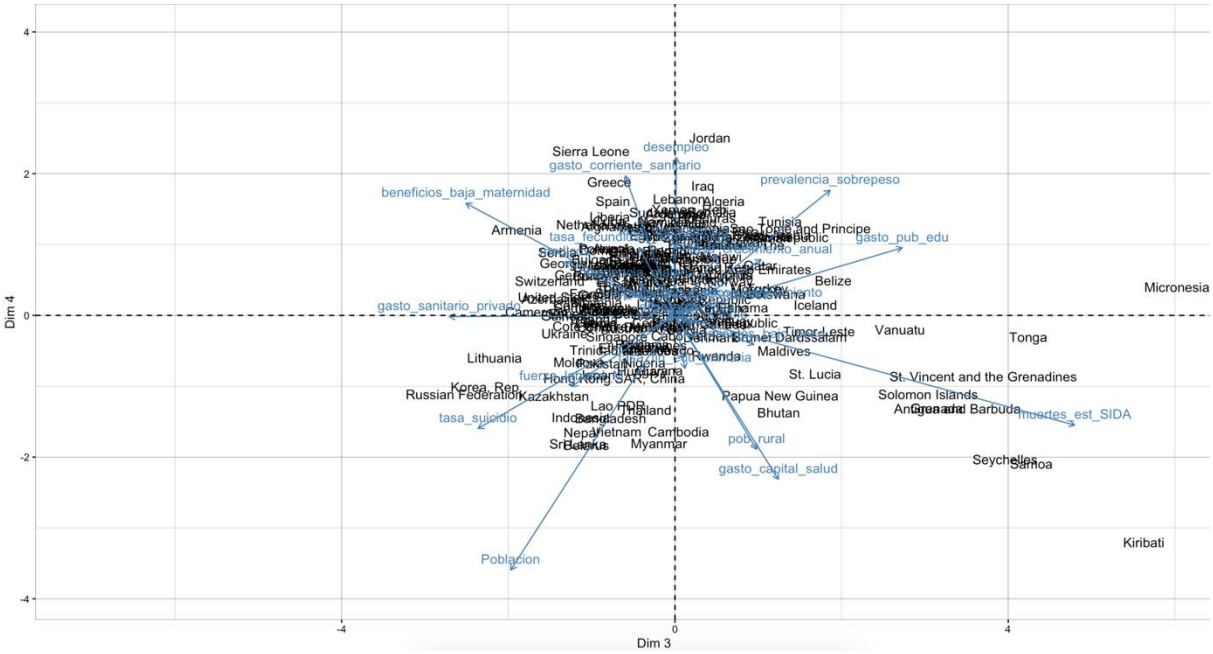


Fig. 3.3.10: *Biplot* de países y variables sobre las componentes 3 y 4

La idea detrás de este *biplot* es que se puedan sacar conclusiones como la siguiente: Suecia, Bielorrusia y Dinamarca tienen un valor alto de la primera y segunda componente, por lo que son países de los que se puede esperar una esperanza de vida alta[35], una baja tasa de fertilidad o un elevado número de fuerza laboral de la mujer.

Aunque a continuación se seguirá el análisis de estas componentes principales mediante un *clustering* y un *profiling,* la idea de este apartado del trabajo era sentar las bases del análisis de componentes principales para que el lector pueda navegar por la aplicación web *Shiny* y sacar las conclusiones que le interesen. Así por ejemplo podría filtrar por los países o las variables que le interesen, o profundizar más y estudiar tantas componentes como quiera.

3.4. Clustering

A continuación se procede a realizar un *clustering* de los países utilizando las componentes principales obtenidas anteriormente. Se aplica aquí el algoritmo no jerárquico *k-means* (véase apartado 2.2.b). Como ya se ha adelantado este método requiere elegir previamente el número de *clusters* que se quieren realizar. Para ello se utilizará la técnica conocida como *Elbow Rule*, la cual consiste en:

- computar el algoritmo *k-means* para diferentes valores de *k* (de 1 a 10 en este caso)
- para cada *k* calcular la suma de cuadrados total *intra-cluster*[36]
- gráficar la curva de suma de cuadrados en función de *k*
- se coge la *k* que corresponda al 'pliegue' (al codo) de la curva

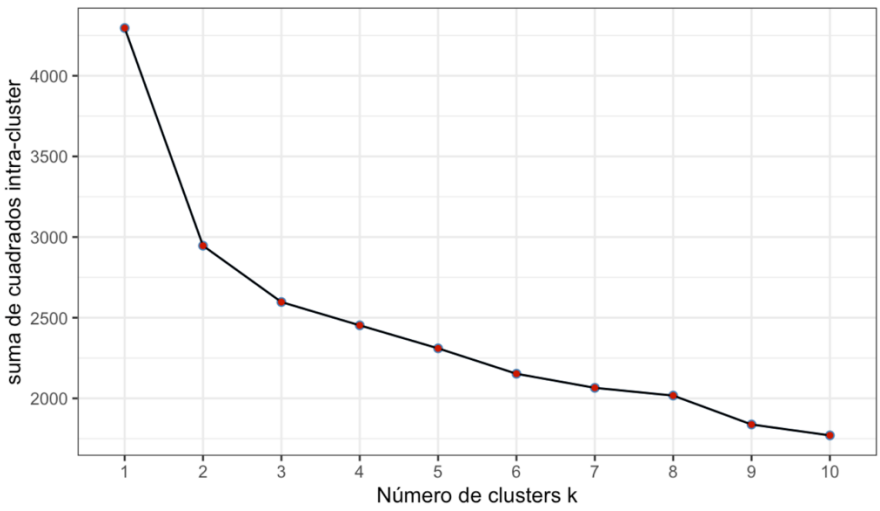


Fig. 3.4.1: suma de cuadrados *intra-cluster* en función del número de clusters *k*

[35] decimos que 'se puede esperar' porque hay que recordar que las componentes son combinaciones lineales de todas las variables, un país podría tener un valor bajo en alguna de las variables muy correlacionadas con una componente dada pero valores altos en el resto de variables correlacionas, por lo que obtendría aún así un valor elevado en la componente (véase justamente el caso de Bielorrusia, país con esperanza de vida media pero con valor elevado en la primera componente)

[36] la suma de cuadrados *intra-cluster* es una medida de variabilidad de las observaciones dentro de casa *cluster*, se quieren *cluster* compactos y parecidos entre sí, de ahí que se desee un valor relativamente bajo

Observamos, como era de esperar, que la suma de cuadrados disminuye a medida que creamos grupos (los grupos son cada vez más pequeños y compactos). Aplicando la *elbow rule* se escogerían 3 *clusters*. Es decir, la suma de cuadrados sigue disminuyendo después de $k = 3$, pero la disminución a partir de este punto no es lo suficientemente grande como para esperar una gran mejora en la variabilidad *intra-cluster.*
Una vez elegido el número de grupos a formar se aplica el algoritmo mediante la función *kmeans().* En esta función se ha especificado que el número de iteraciones máximas (i.e. el número de veces que se recalculan los centroides) serán 10 y que se generarán 25 juegos aleatorios de centroides[37]. Es decir, se recalcula el algoritmo 25 veces con 3 centroides iniciales aleatorios cada vez y se guarda el resultado más común. De esta manera se asegura la convergencia.

Obviamente, cuando se trabaja con un número elevado de dimensiones, resulta imposible representar los grupos formados. De ahí que se aplique el algoritmo sobre las componentes principales. Y así, como ya se ha explicado, bastará con analizar las primeras componentes para obtener una idea precisa de cómo se distribuyen estos *clusters*.

Antes de nada veamos a que *cluster* se ha asignado cada país:

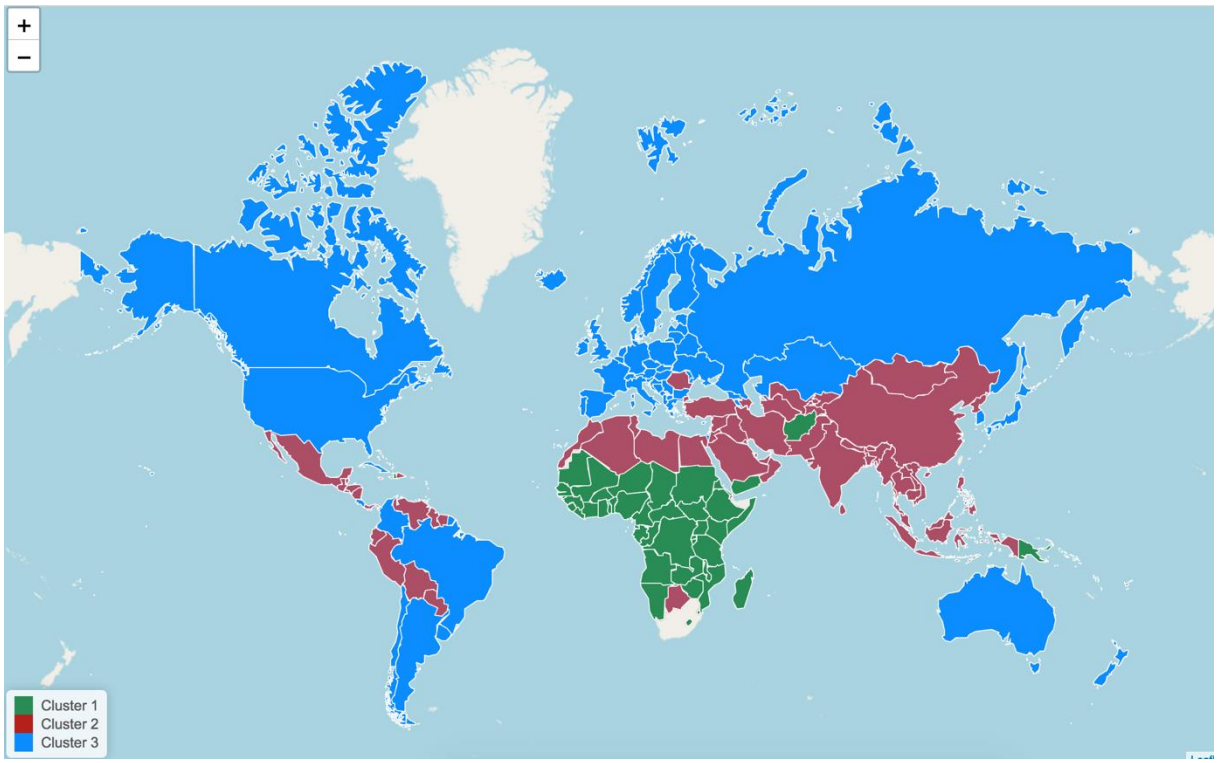


Fig. 3.4.2: mapa mundial de los clusters formados

[37] la función *kmeans* también da la opción de insertar las coordenadas iniciales de los centroides manualmente

Proyectemos ahora estos países sobre las cuatro primeras componentes:

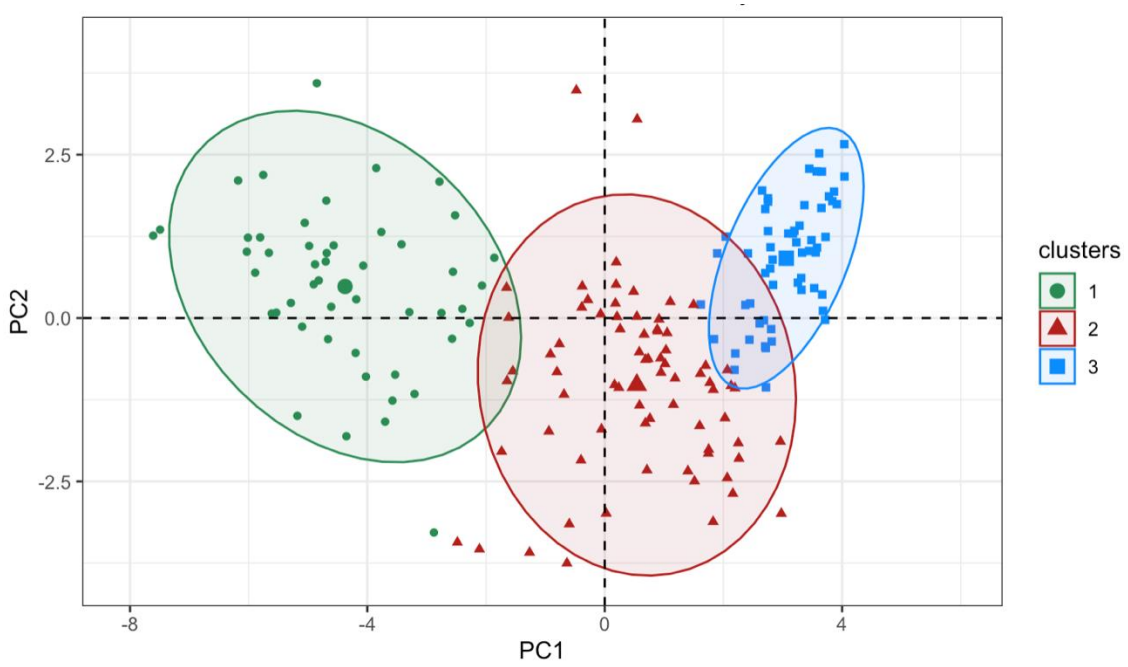


Fig. 3.4.3: clusters sobre las componentes principales 1 y 2

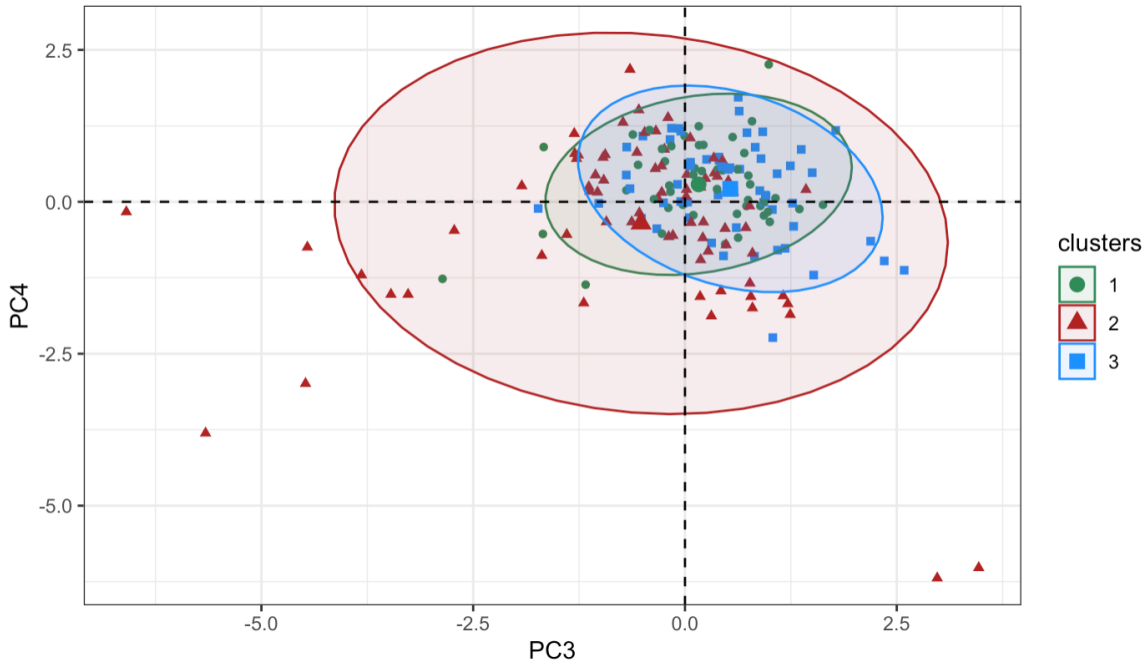


Fig. 3.4.4: clusters sobre las componentes principales 3 y 4

Se observa como claramente los *clusters* han sido principalmente formados a partir de la primera y segunda componente. Lo cual era de esperar ya que reúnen entre ellas casi el 50% de la varianza total. Así que, observando el primer gráfico, vemos que los países del primer *cluster* se caracterizan por valores bajos en la primera componente y medios-altos en la segunda. El segundo *cluster* se caracteriza por valores medios en la primera componente y

bajos-medios en la segundo. Finalmente el tercer *cluster* se caracteriza por valores altos en la primera componente y medios-altos en la segundo.
Ahora sabiendo que con variable se correlaciona cada componente sabremos que características tiene cada *cluster.* Aunque esta información ya se ha desarrollado en el apartado 3.3, se proyectan a continuación los individuos categorizados por *cluster*, así como las 12 variables que más contribuyen en términos de varianza sobre las dos primeras componentes:

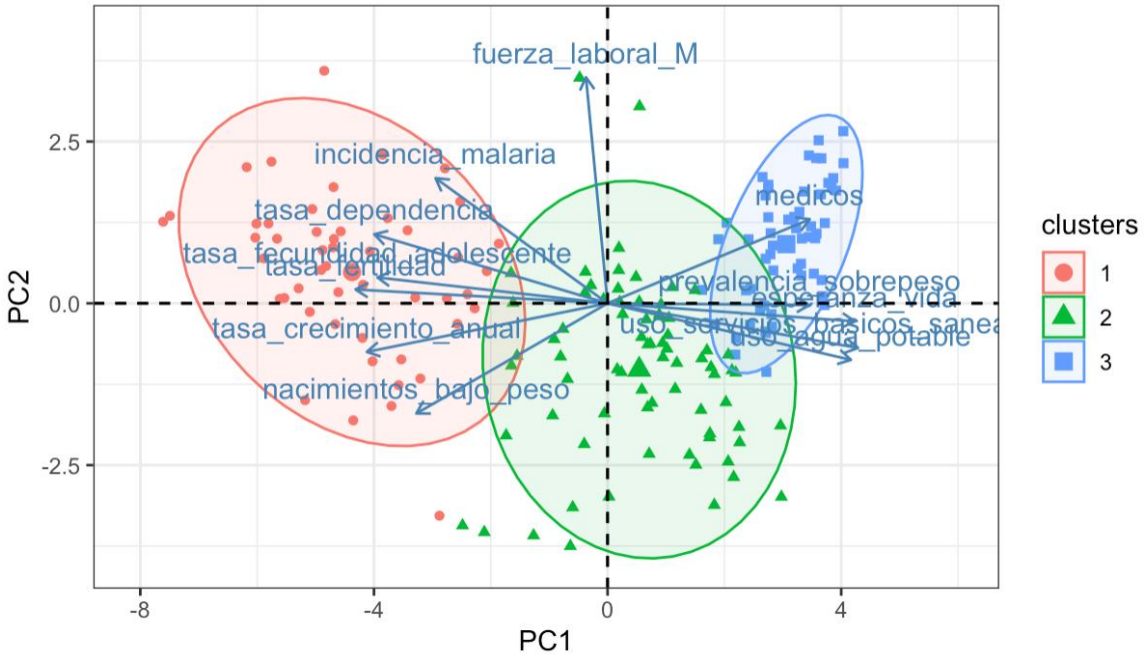


Fig. 3.4.5: clusters y 12 variables más contribuyentes sobre las componentes 1 y 2

Combinando la información de este gráfico con el mapa de la figura 3.4.2 queda claro que se han formado 3 c*lusters* por nivel de pobreza, salud y bienestar en general de la población.
Aunque no se puede hablar exactamente de primer, segundo y tercer mundo si que es verdad que el tercer *cluster* está principalmente formado por países de altos ingresos de Europa, América y Oceanía. El primer *cluster* lo forman en cambio países de muy bajos ingresos, principalmente de África. Y finalmente el segundo *cluster* lo componen países algo más avanzados de Asía y América del Sur y Central.

A continuación se detalla la media de cada variable en cada *cluster*:

	Cluster 1	Cluster 2	Cluster 3
Tasa de fertilidad	4.74	2.47	1.66
Tasa de fecundidad adolescente	101.0	41.3	18.7
Nacimientos bajo peso	15.02	11.98	6.64
Muertes estimadas por SIDA	0.865	2.484	0.634
Médicos	0.204	1.023	3.091
Prevalencia de la desnutrición	18.91	9.36	3.01
Prevalencia del sobrepeso	29.9	50.8	57.2
Uso de agua potable	61.8	92.8	98.3
Uso servicios básicos de saneamiento	30.8	83.6	96.6
Incidencia malaria	187.8	20.6	34.3
Gasto de capital sanitario	0.203	0.228	0.247
Gasto corriente sanitario	5.80	5.39	8.43
Gasto privado sanitario	45.6	43.1	34.2
Tasa de suicidio	7.45	7.01	14.17
Tasa de dependencia	84.2	51.4	49.5
Tasa de desempleo	7.12	6.53	9.39
Fuerza laboral, mujeres	44.1	36.2	45.5
Beneficios por baja de maternidad	92.9	91.3	90.2
Tasa de finalización de educación primaria	65.9	90.1	96.1
Gasto público en educación	4.24	4.82	5.02
Población rural	61.3	43.6	25.2
Tasa de crecimiento anual	26.84	14.05	2.07
Esperanza de vida	61.0	72.8	78.7
Población	20797735	61777845	29066268

Fig. 3.4.6: Media de las variables por *cluster*

En el siguiente apartado se resume más ampliamente las características de cada grupo.

3.5. Profiling

Se muestra en este apartado a modo de resumen un *profiling,* o elaboración de perfiles, de los 3 *clusters* obtenidos.

Cluster 1	Cluster 2	Cluster 3
• Países de África del oeste, del este y del centro. Además de países de otros continentes como Papúa Nueva Guinea, Haití o Afganistán	• Países de África del norte, gran parte de Centroamérica, varios países de Sudamérica como Perú, Bolivia o Venezuela. Países de Oriente Medio y prácticamente el resto de Asia al completo.	• Ciertos países de Sudamérica como Brasil, Argentina o Chile. América del Norte y prácticamente toda Europa y Oceanía.
• 14.1% de la población mundial	• 63.1% de la población mundial	• 22.9% de la población mundial
• Altas tasas de fertilidad, de fecundidad adolescente, de nacimientos bajo peso, de crecimiento anual y gran incidencia de la malaria.	• Tasas medias de fertilidad, de fecundidad adolescente, de nacimientos bajo peso, de crecimiento anual y gran incidencia de la malaria.	• Bajas tasas de fertilidad, de fecundidad adolescente, de nacimientos bajo peso, de crecimiento anual y gran incidencia de la malaria.
• Bajos índices en sobrepeso, en uso de servicios básicos de saneamiento y en uso de agua potable. Muy baja esperanza de vida (61 años aprox.) y muy bajo número de médicos.	• Tasas medias-altas en sobrepeso, en uso de servicios básicos de saneamiento y en uso de agua potable. Esperanza de vida media (72.8 años) y bajo número de médicos.	• Altos índices en sobrepeso, en uso de servicios básicos de saneamiento y en uso de agua potable. Gran esperanza de vida (79 años aprox.) y alto número de médicos.
• Alto índice de fuerza laboral femenina y población mayoritariamente rural.	• Bajo índice de fuerza laboral femenina y población tanto rural como urbana	• Alto índice de fuerza laboral femenina y población mayoritariamente urbana.

Fig. 3.5.1: *Profiling* de los clusters

4. Modelización y regularización

Una vez explorado y entendido mejor el comportamiento de los datos se proceden a aplicar ya los modelos de predicción. Como se ha comentado en la introducción, el objetivo principal de este trabajo es quizás el de estudiar distintos métodos de regularización. Para ello se comienza aplicando una regresión lineal, se seguirá con los métodos de regresión por reducción de dimensionalidad, y se acabará con los métodos de regularización de Ridge, Lasso, ElasticNet y *group* Lasso.
Para ello se utilizarán dos tercios de los datos para entrenar los modelos y el resto para 'testear' el modelo. La variable a predecir será *esperanza_vida*. Aunque ya se ha comentado que el objetivo principal es estudiar los métodos de regularización, y no tanto interesarse realmente en la variable respuesta elegida. Aún así, se ha creído oportuno elegir esta variable ya que puede resultar interesante comprobar en que medida se puede predecir la esperanza de vida de un país basándose en indicadores de salud, bienestar y economía.

Antes de seguir, comentar que también para este apartado se recomienda acudir a la aplicación web para complementar la lectura.

4.1. Regresión Lineal Múltiple

A continuación se ajusta una regresión lineal múltiple por mínimos cuadrados ordinarios. Para ello se ha aplicado la función *lm()* de R sobre los datos de entrenamiento. Se puede ahora comprobar que coeficientes son significativos:

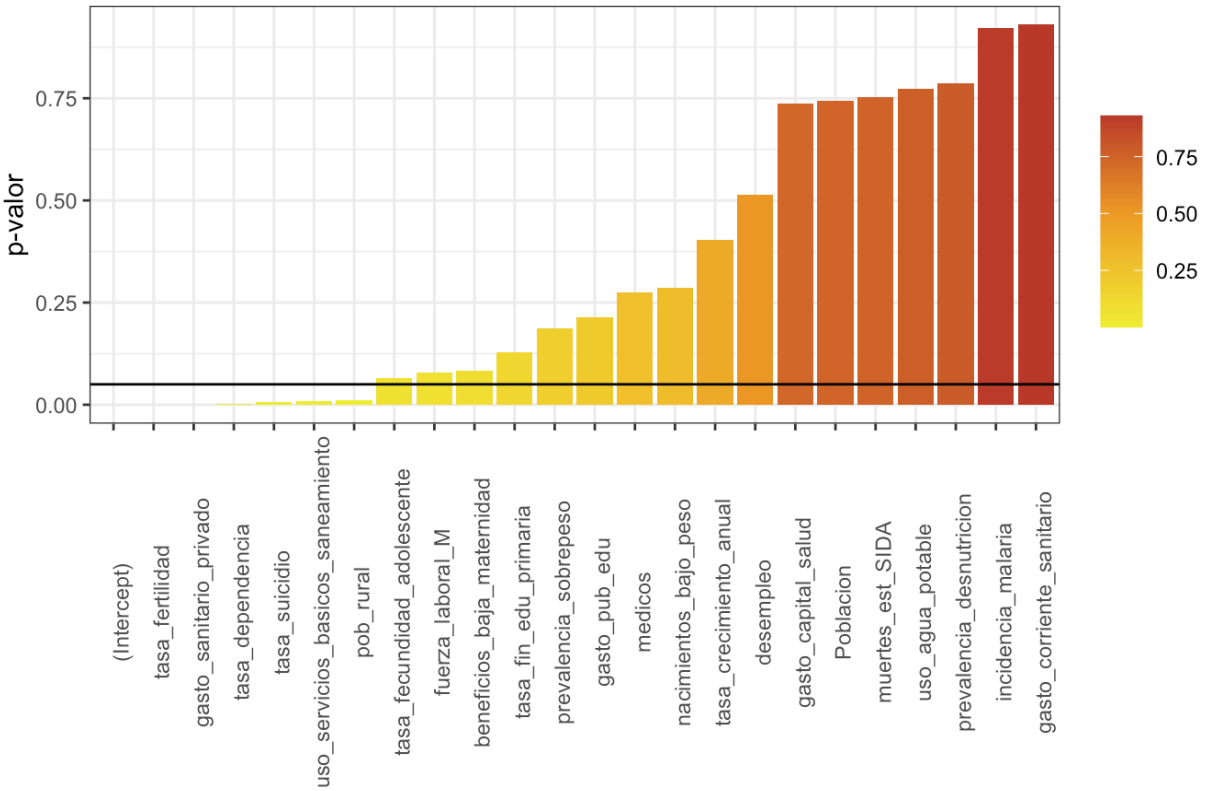


Fig. 4.1.1: p-valor del estadístico t-Student de los regresores del modelo

En el gráfico anterior se muestran las variables ordenadas por el p-valor del test de hipótesis de t de Student sobre los coeficientes. Este test comprueba si un coeficiente de regresión (parámetro β) es diferente de 0. Así cuanto más pequeño sea este valor más significativa es la variable. Las variables que mejor explican la variable respuesta (p-valor bien inferior a 0.05) son: *tasa_fertilidad, gasto_sanitario_privado*, *tasa_dependencia*, *tasa_suicidio*, *uso_servicios_basicos_saneamiento* y *pob_rural*.

Una vez ajustado el modelo se calculan las métricas de error explicadas en el apartado 2.5.3. Estas métricas se aplican tanto al conjunto de entrenamiento como al de *test*:

Regresión Lineal	**RMSE**	**MAE**	R^2_{adj}
Train	2.230	1.750	0.887
Test	4.091	3.040	0.654

Fig. 4.1.2: Tabla de medidas de error para los datos de *train* y *test* de la regresión lineal por MCO

Primeramente hay que comentar que el modelo se ajusta muy bien sobre los datos de entrenamiento. Pero como vemos las resultados empeoran cuando se trata de los datos de *test*. Aunque suele ser lo habitual, podría tratarse de un problema sobreajuste. El modelo podría estar ajustándose demasiado a las características específicas de los datos de entrenamiento, haciendo que no consiga generalizar correctamente con datos no vistos previamente.

4.2. Regresión de Ridge

Para modelizar las regresiones de Ridge, Lasso y ElasticNet se hará uso de las funciones *cv.glmnet()* y *glmnet()* del paquete *glmnet*. Estas funciones nos permiten introducir manualmente los parámetros α y λ deseados. Recordemos que α controla que modelo se aplica, en este caso se fija $\alpha = 0$ para obtener la regresión de Ridge[38].
En cuanto al parámetro de complejidad λ, no hay forma de saber de antemano qué valor será el más adecuado. Por ello mediante *cv.glmnet()* se puede calcular el error por validación cruzada de una serie de valores λ que se deseen. En otras palabras, la función calcula el error por validación cruzada (véase apartado 2.5.2.) para todos los valores de λ que se proporcionen. En este caso los valores de λ que se prueban son aquellos que van de 0 a 5 de 0.01 en 0.01 [0, 0.01, 0.02, …, 4.99, 5.00]. La validación cruzada se ha decidido hacerla por *Leave-one-out cross-validation* (no se disponen de muchos datos por lo que el coste computacional no es un problema[39]).

[38] se había dicho anteriormente en la metodología que $\alpha = 0$ era Lasso, esto simplemente es así por que la función glmnet() asigna α al revés de lo visto

[39] aún así se ajustará el modelo unas 58116 veces para hallar λ óptima (nº individuos * nº de λs)

Una vez aplicada la función se obtiene que el parámetro de complejidad λ que minimiza el error es λ = 0.04. A continuación se muestra cual sería el error con el resto de valores introducidos de λ:

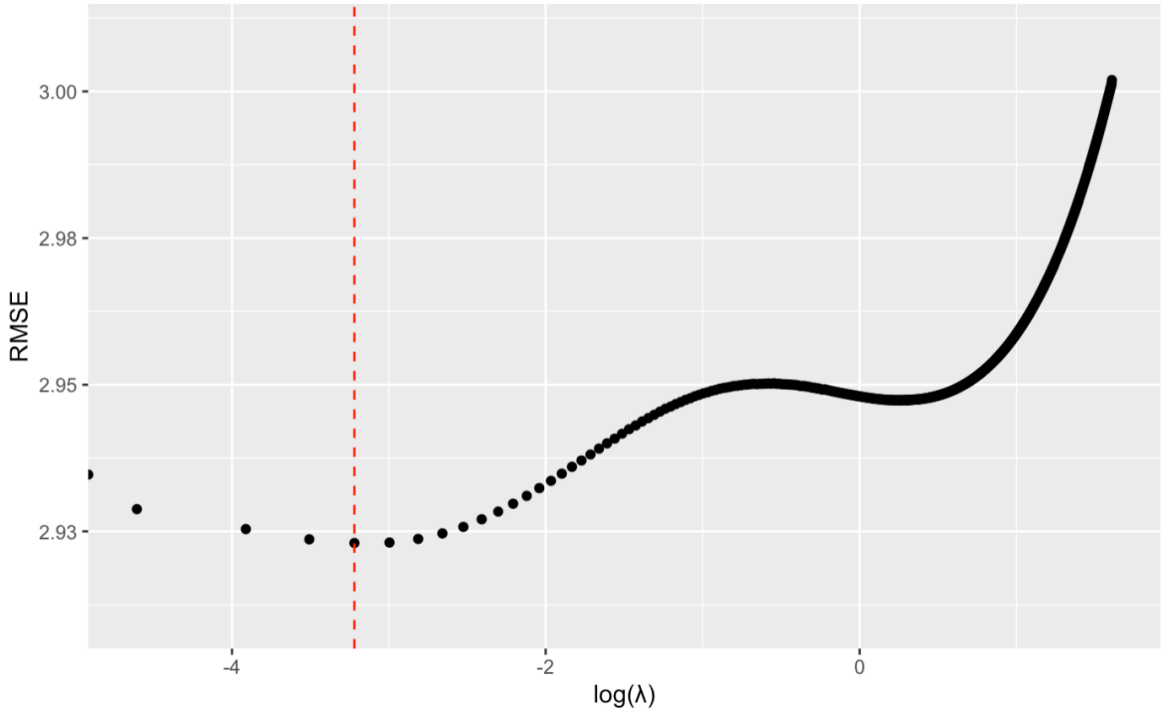


Fig. 4.2.1: RMSE por validación cruzada de la regresión de Ridge en función del logaritmo[40] de λ

El valor obtenido de λ es bastante bajo, lo cual quiere decir que la penalización que se está aplicando no es muy dura, y por consecuencia la reducción de los coeficientes no será excesiva.

Se muestra ahora el cambio porcentual de los coeficientes en base a la regresión lineal del apartado anterior:

[40] se usa el logaritmo por cuestiones visuales

predictor	coeficiente
<chr>	<dbl>
(Intercept)	-0.0269
tasa_fertilidad	-0.202
tasa_fecundidad_adolescente	-0.0285
nacimientos_bajo_peso	-0.0502
muertes_est_SIDA	0.525
medicos	-0.0222
prevalencia_desnutricion	-0.109
prevalencia_sobrepeso	-0.0635
uso_agua_potable	1.22
uso_servicios_basicos_saneamiento	-0.0630
incidencia_malaria	3.99
gasto_capital_salud	-0.401
gasto_corriente_sanitario	-2.76
gasto_sanitario_privado	0.0165
tasa_suicidio	0.0321
tasa_dependencia	-0.227
desempleo	-0.279
fuerza_laboral_M	-0.200
beneficios_baja_maternidad	-0.0325
tasa_fin_edu_primaria	0.0411
gasto_pub_edu	-0.0214
pob_rural	-0.0752
tasa_crecimiento_anual	-0.667
Poblacion	-0.153

Fig. 4.2.2: cambio porcentual de los coeficientes entre la regresión lineal y Ridge

En rojo aparecen los coeficientes que han disminuido, y como se ve estos son la gran mayoría. Hay algunas anomalías como *incidencia_malaria* (aumento de un 400%) o *muertes_est_SIDA* (aumento en un 50%), pero estas son unas variables muy poco significativas con coeficientes muy bajos, por lo que este aumento es insignificante en sí mismo.

Puede resultar interesante por otra parte ver como evolucionan los coeficientes del modelo en función de λ:

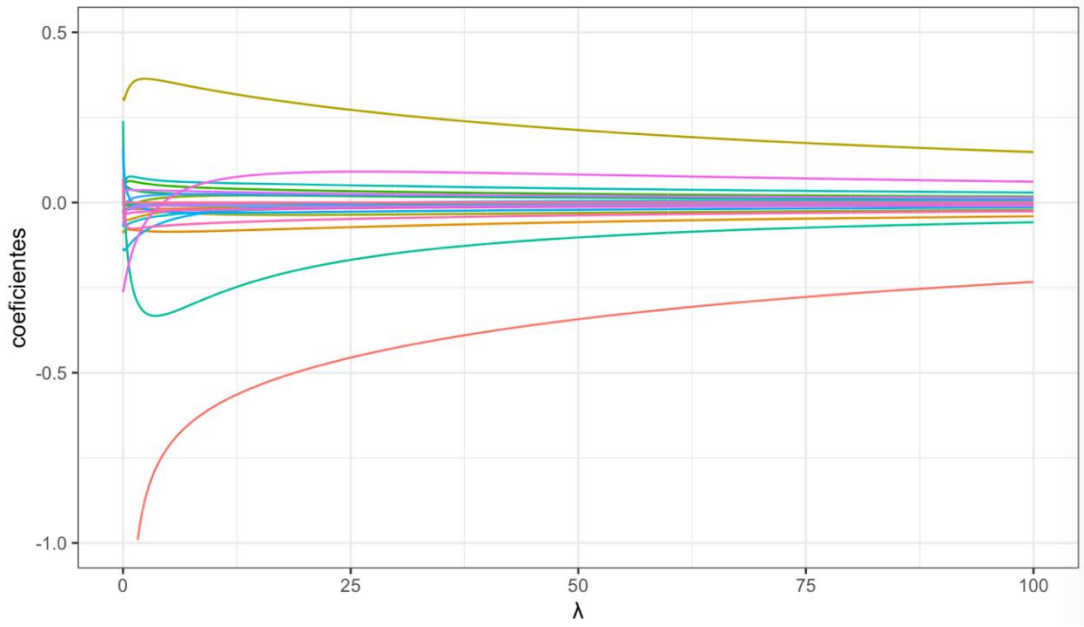


Fig. 4.2.3: coeficientes del modelo de Ridge en función de λ

Podemos observar efectivamente como la regresión de Ridge hace que los coeficientes tiendan hacia cero, sin llegar nunca a ser exactamente cero por muy grande que sea el parámetro de complejidad.

Ya tan solo queda por ver que tan bien ajusta y predice este modelo sobre datos nunca vistos anteriormente. Para ello se vuelven a calcular, tanto para los datos de entrenamiento como de test, las tres métricas de error anteriores:

Ridge	**RMSE**	**MAE**	$\boldsymbol{R^2_{adj}}$
Train	2.242	1.749	0.885
Test	4.103	3.087	0.652

Fig. 4.2.4: Tabla de medidas de error para los datos de *train* y *test* del modelo de Ridge

Estas medidas son prácticamente idénticas, aunque algo peores, a las obtenidas en la regresión lineal por mínimos cuadrados ordinarios. En los datos de entrenamiento los errores por Ridge siempre serán peores que por mínimos cuadrados, ya que esa es la definición misma de MCO, es el método lineal que mejor describe un set de datos ya que minimiza la suma de residuos cuadrados. Por otro lado, el objetivo de una regresión regularizada es mejorar la precisión de predicción en los datos de test, lo cual no se ha conseguido.
Aún así, estos resultados podían esperarse ya que el coeficiente de complejidad es muy bajo, por lo que el modelo es poco restrictivo y parecido al de la regresión lineal. Esto no significa que se haya hecho algo mal, simplemente que el sesgo introducido en la estimaciones no se ha visto seguido por una reducción de la varianza (*bias-variance tradeoff*).

4.3. Regresión de Lasso

El proceso de modelización de la regresión de Lasso es muy parecido al realizado anteriormente con Ridge. Se utilizan las mismas funciones y el valor de λ sigue eligiéndose por *cross-validation*. El valor de α en cambio deberá ser ahora igual a 1 para aplicar Lasso.

Se obtiene aquí que el parámetro de complejidad λ con el cual se obtiene un error mínimo es $\lambda = 0.23$. Se muestra a continuación el error del modelo en función de este parámetro:

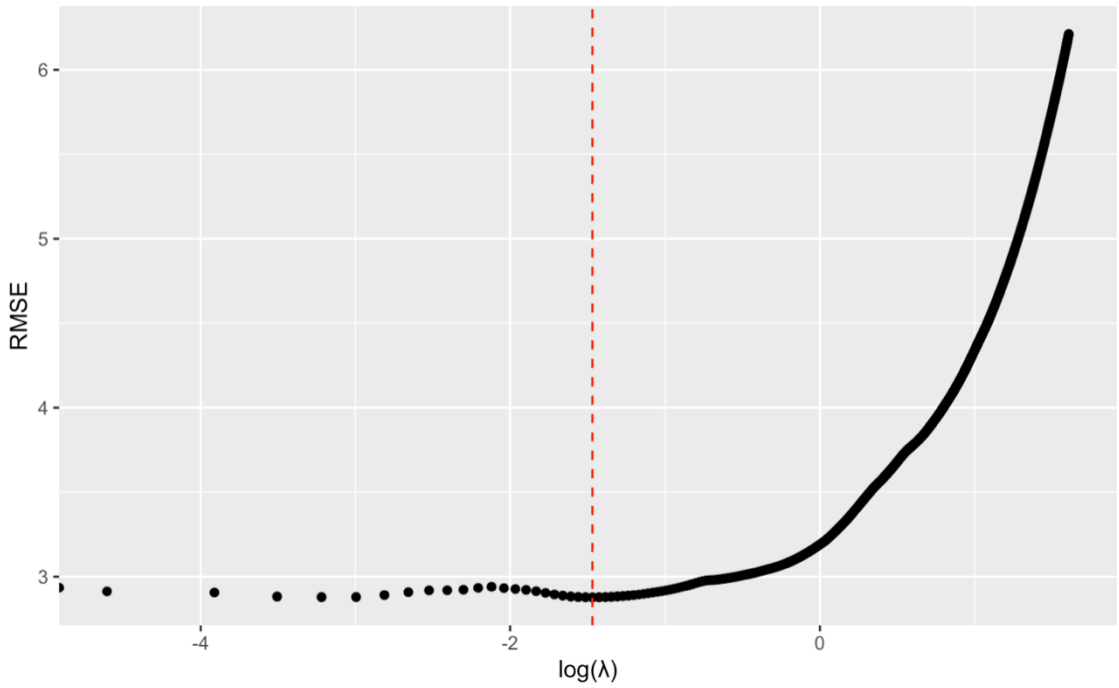


Fig. 4.3.1: RMSE por validación cruzada de la regresión de Lasso en función del logaritmo de λ

El valor de λ es esta vez algo superior que en el caso de Ridge. Esto implicará una penalización más estricta.

Veamos como han variado esta vez los coeficientes entre este modelo y la regresión lineal por mínimos cuadrado ordinarios:

predictor	coeficiente
<chr>	<dbl>
(Intercept)	-0.100
tasa_fertilidad	-0.545
tasa_fecundidad_adolescente	-0.159
nacimientos_bajo_peso	-0.752
muertes_est_SIDA	-1
medicos	0.290
prevalencia_desnutricion	-1
prevalencia_sobrepeso	-1
uso_agua_potable	1.25
uso_servicios_basicos_saneamiento	-0.298
incidencia_malaria	6.19
gasto_capital_salud	-1
gasto_corriente_sanitario	-1
gasto_sanitario_privado	-0.253
tasa_suicidio	-0.480
tasa_dependencia	-1
desempleo	-1
fuerza_laboral_M	-1
beneficios_baja_maternidad	-1
tasa_fin_edu_primaria	-0.119
gasto_pub_edu	-1
pob_rural	-0.289
tasa_crecimiento_anual	-1
Poblacion	-1

Fig. 4.3.2: cambio porcentual de los coeficientes entre la regresión lineal y Lasso

Se observa que han disminuido los coeficientes en todos menos en tres coeficientes. Además, como era de esperar hasta 12 variables han sido completamente eliminadas del modelo. De esta eliminación de variables se destaca que la mayoría eran variables poco o muy poco significativas, únicamente la variable *tasa_dependencia* era muy significativa. Esto puede ser debido a que como se ha dicho Lasso tiende a eliminar la mayoría de predictores muy correlacionados entre sí. En este caso, esta variable esta muy correlacionada con variables como *tasa_fertilidad* o *uso_servicios_basicos_saneamiento*, las cuales efectivamente no han sido eliminadas.

Se puede por otro lado apreciar como se eliminan variables a medida que aumenta el nivel de regularización:

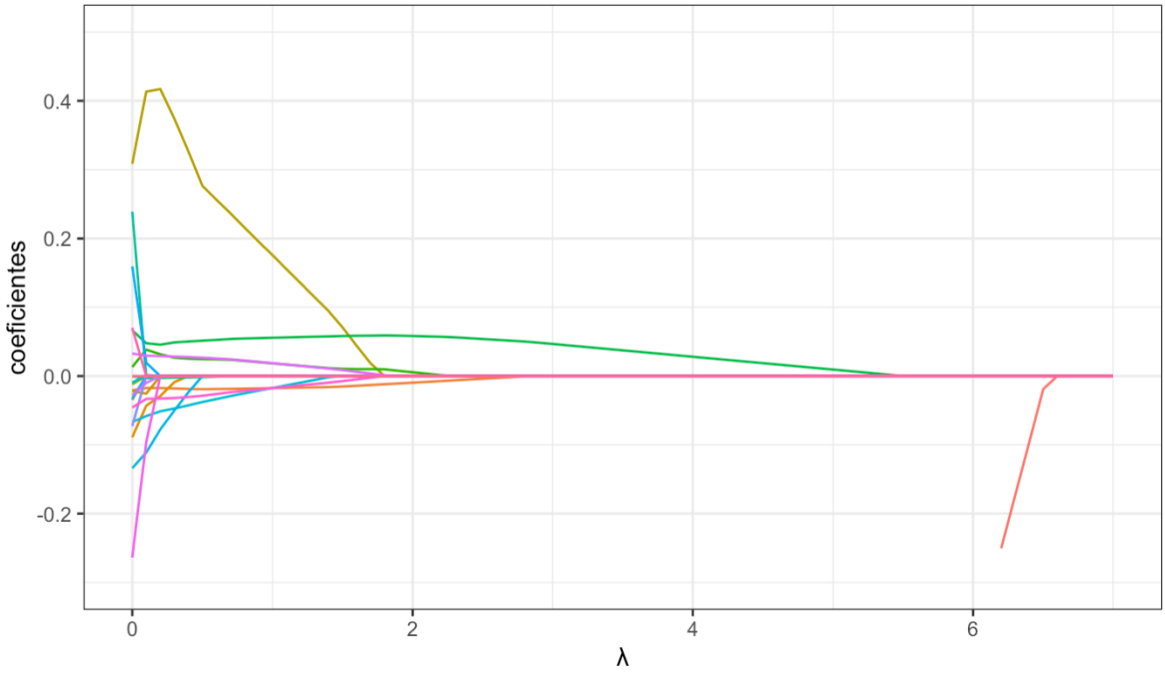


Fig. 4.3.3: coeficientes del modelo de Lasso en función de λ

En efecto, se aprecia en este gráfico como los coeficientes disminuyen a medida que aumenta λ, y como se convierten eventualmente en cero.

Una vez implementado el modelo toca evaluar su capacidad de predicción. Se muestran aquí las medidas de error obtenidas tanto en los datos de entrenamiento como de test:

Lasso	**RMSE**	**MAE**	R^2_{adj}
Train	2.490	1.94	0.861
Test	4.286	3.342	0.703

Fig. 4.3.4: Tabla de medidas de error para los datos de *train* y *test* del modelo de Lasso

Primeramente, se observan en los datos de entrenamiento mayores errores de predicción que en el modelo de Ridge. Recordemos que ahora se está siendo más restrictivo (mayor λ), lo cual implica en el caso de Lasso la eliminación de varios predictores. Por lo tanto se está perdiendo información, haciendo que resulte complicado que el modelo pueda describir el conjunto de datos con la misma precisión.
En cuanto a la capacidad predictiva en los datos de test, también ha empeorado respecto al modelo de Ridge si tenemos en cuenta el RMSE o el MAE. Aún así el empeoramiento no es excesivo (alrededor de un 5%) y hay que tener en cuenta que lo que se pierde en capacidad predictiva se gana en interpretabilidad y eficiencia. De hecho si se tiene en cuenta el coeficiente de determinación ajustado, el cual penaliza el uso de muchos predictores, entonces el modelo es mejor que ambos modelos anteriores. Sería trabajo del investigar decidir que prefiere, minimizar lo máximo posible el error, o sacrificar algo de capacidad predictiva por un modelo más simple.

4.4. Regresión ElasticNet

Anteriormente se ha dicho que el modelo por regresión lineal y el modelo por regresión de Ridge tenían mejor capacidad predictiva que el de Lasso, pero que este último también presentaba sus ventajas en forma de selección de variables y por lo tanto de interpretabilidad. ElasticNet nos permite convenientemente combinar las regresiones de Ridge y de Lasso. Se ha visto en el apartado de *metodología* que este método requiere especificar dos parámetros: el parámetro α que controla si el modelo se parece más a Ridge o a Lasso, y λ, que sigue siendo el parámetro de complejidad. Se recurre de nuevo a la validación cruzada por *leave-one-out cross-validaition* para calcular la combinación de parámetros que minimicen el error en los datos de entrenamiento. Para ello se cogerán valores de λ de 0 a 1.25 (0, 0.01, 0.02, …, 1.25) y valores de α de 0 a 1 (0, 0.1, 0.2, …, 0.9, 1).
Es decir, para cada una de estas parejas de valores se calcula el RMSE por *leave-one-out cross-validation*. Fíjese que ahora hay que calcular el error para cada combinación de ambos parámetros, resultando en un coste computacional más elevado[41]. A continuación se muestran dos maneras diferentes de visualizar el RMSE en función de estos parámetros:

[41] se realizan en total 160776 iteraciones

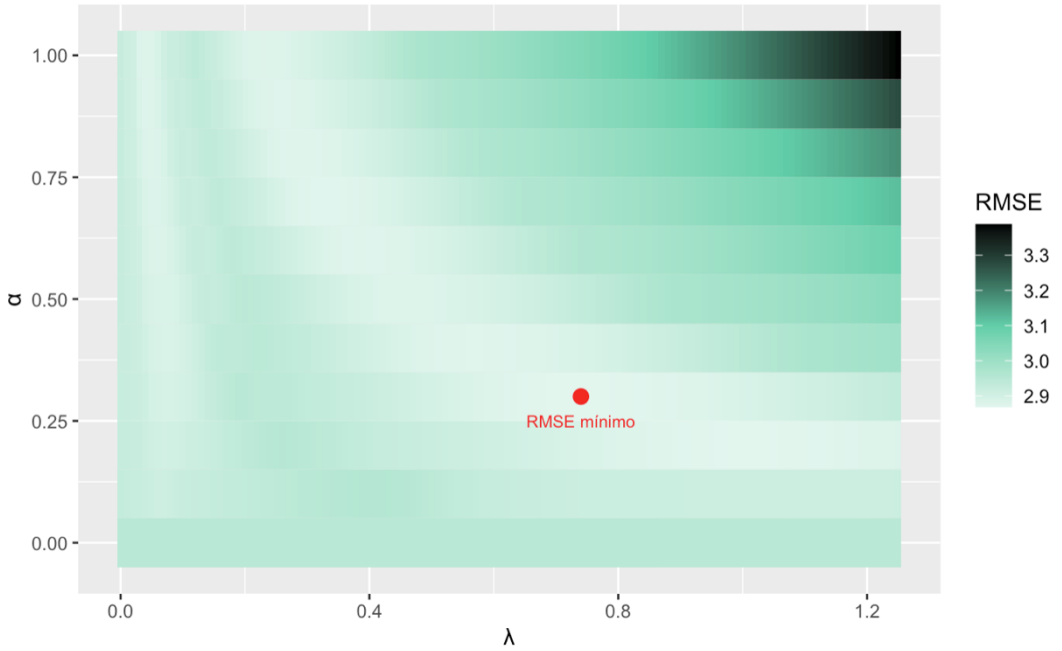


Fig. 4.4.1: RMSE por validación cruzada de la regresión ElasticNet en función de λ y de α

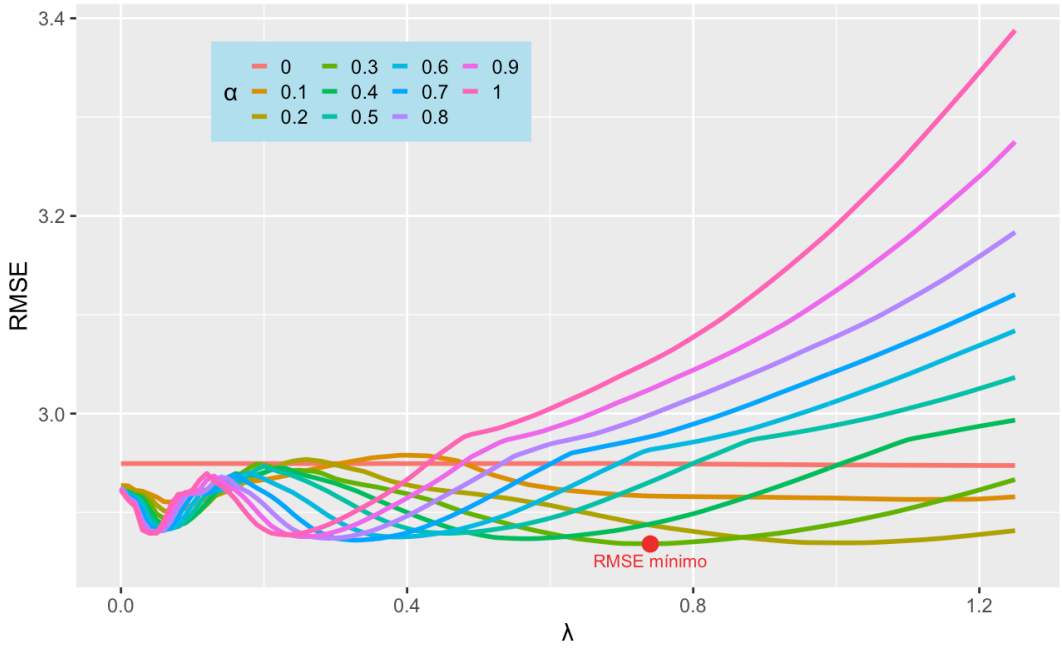


Fig. 4.4.2: RMSE por validación cruzada de la regresión ElasticNet en función de λ y de α (2)

Los valores de los parámetros correspondientes a este valor mínimo de error son $\alpha = 0.3$ y $\lambda = 0.74$. El primer valor indica que el modelo se acerca más al modelo de Ridge que al de Lasso, es decir que se aplicarán las dos penalizaciones pero dándole más peso a la de Ridge. El segundo valor es en esta ocasión algo más alto que en modelos anteriores, aunque al ser ElasticNet una combinación de estos otros dos modelos no se puede asegurar que esto implique una penalización más fuerte.

Veamos como se traducen estos parámetros en términos de coeficientes, para ello se vuelven a comparar los coeficientes con los obtenidos por mínimos cuadrados ordinarios:

predictor	coeficiente
<chr>	<dbl>
(Intercept)	-0.135
tasa_fertilidad	-0.696
tasa_fecundidad_adolescente	-0.0529
nacimientos_bajo_peso	-0.453
muertes_est_SIDA	-1
medicos	0.199
prevalencia_desnutricion	-1
prevalencia_sobrepeso	-1
uso_agua_potable	2.79
uso_servicios_basicos_saneamiento	-0.391
incidencia_malaria	10.5
gasto_capital_salud	-1
gasto_corriente_sanitario	-1
gasto_sanitario_privado	-0.308
tasa_suicidio	-0.503
tasa_dependencia	-1
desempleo	-1
fuerza_laboral_M	-1
beneficios_baja_maternidad	-1
tasa_fin_edu_primaria	0.0421
gasto_pub_edu	-1
pob_rural	-0.317
tasa_crecimiento_anual	-1.60
Poblacion	-1

Fig. 4.4.3: cambio porcentual de los coeficientes entre la regresión lineal y ElasticNet

De esta información se puede destacar que, al igual que en los modelos de Ridge y Lasso, la mayoría de coeficientes se han reducido. Por otro lado se han eliminado 11 predictores, uno menos que Lasso (no se elimina la variable *tasa_crecimiento_anual* y el resto de las que se eliminan coinciden). Se puede decir que ElasticNet 'quita' importancia a Lasso para dársela a Ridge, por lo que es de esperar que se seleccionen menos variables (aunque teniendo en cuenta que esto también depende del parámetro λ).

Veamos que tal predice este modelo:

ElasticNet	**RMSE**	**MAE**	R^2_{adj}
Train	2.513	1.950	0.859
Test	4.258	3.316	0.701

Fig. 4.4.4: Tabla de medidas de error para los datos de *train* y *test* del modelo de Lasso

Si se miran los errores sobre los datos de entrenamiento, este modelo resulta algo peor que el de Lasso. Pero lo que nos interesa realmente es la capacidad predictiva sobre los datos de test, donde ElasticNet obtiene errores menores, aunque muy cercanos, a los de Lasso.
Más adelante, se compara de forma más clara la capacidad predictiva, así como las ventajas y desventajas, de cada uno de estos y del resto de modelos aún no vistos.

4.5. *Group* Lasso

Se aplica ahora el método *Group* Lasso, el cual recordemos permite tratar las variables como grupos. Este modelo suele ser especialmente útil cuando se tiene una variable categórica codificada mediante varias variables binarias. Aunque este no es nuestro caso, también puede resultar útil cuando se tienen variables agrupadas por características en común. Se han creado así los siguientes grupos:

Grupo	Variables
1- Índices de fertilidad	*tasa_fertilidad, tasa_fecundidad_adolescente*
2- Índices de salud	*nacimientos_bajo_peso, muertes_est_SIDA, médicos, prevalencia_desnutricion, prevalencia_sobrepeso, uso_agua_potable, uso_servicios_basicos_saneamiento, incidencia_malaria*
3- Índices de gasto en sanidad	*gasto_capital_salud,gasto_corriente_sanitario, gasto_sanitario_privado*
4- Índices laborales	*tasa_dependencia, desempleo, fuerza_laboral_M, beneficios_baja_maternidad*
5- Índices de educación	*tasa_fin_edu_primaria, gasto_pub_edu*
6- Índices poblacionales	*pob_rural, tasa_crecimiento_anual, Poblacion*
7- Otro	*tasa_suicidio*

Fig. 4.5.1: Grupos de variables

Una vez se ha asignado cada variable a su grupo, hay que elegir el parámetro λ óptimo. De nuevo, este parámetro se determinará por *leave-one-out cross-validation*. Para obtener este parámetro y para ajustar el modelo se han utilizado las funciones *cv.gglasso()* y *gglasso()* del paquete *gglasso.*

Se obtiene que el valor óptimo de λ es 0.6. Y los coeficientes del modelo son:

(Intercept)	71.4623529127
tasa_fertilidad	-0.0093889248
tasa_fecundidad_adolescente	-0.0249354011
nacimientos_bajo_peso	-0.0309227353
muertes_est_SIDA	-0.0358628303
medicos	0.0187885812
prevalencia_desnutricion	-0.0148315030
prevalencia_sobrepeso	-0.0279251531
uso_agua_potable	0.0494947194
uso_servicios_basicos_saneamiento	0.0611324367
incidencia_malaria	-0.0081540108
gasto_capital_salud	-0.0035410973
gasto_corriente_sanitario	0.0381595767
gasto_sanitario_privado	-0.0656928972
tasa_suicidio	-0.1222849501
tasa_dependencia	0.0016180755
desempleo	0.0041943411
fuerza_laboral_M	0.0027275866
beneficios_baja_maternidad	-0.0112695157
tasa_fin_edu_primaria	0.0425815452
gasto_pub_edu	-0.0132801212
pob_rural	-0.0478727436
tasa_crecimiento_anual	-0.1834988801
Poblacion	-0.0001323889

Fig. 4.5.2: Coeficientes del modelo Group Lasso

Es decir, no se ha eliminado ningún grupo. Por lo tanto en este caso el modelo no está sirviendo como selección de variables, aunque si como reducción de coeficientes. Se puede comprobar como con una penalización más alta, sí que se eliminarían varios grupos de variables:

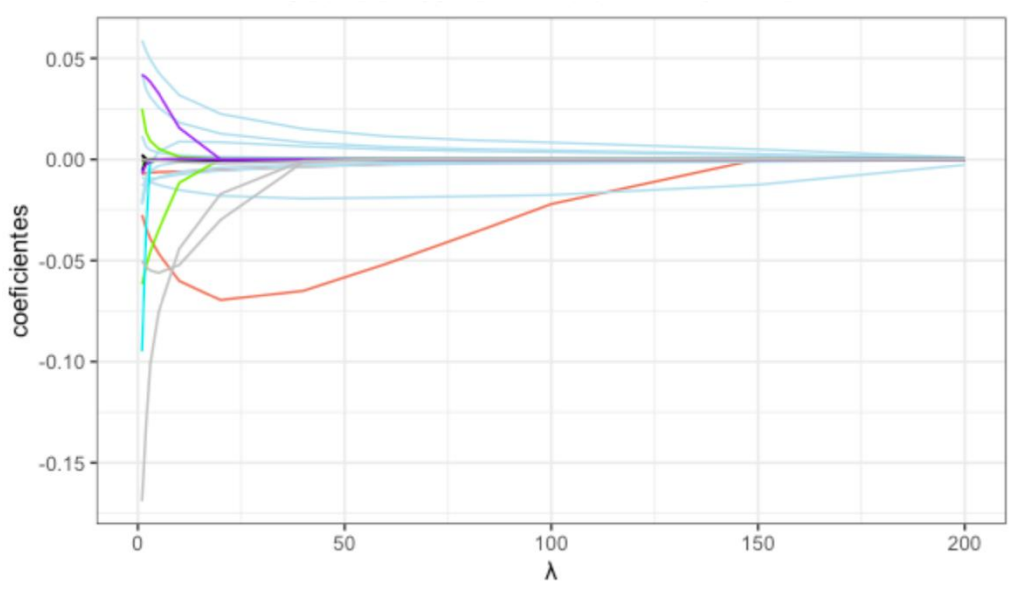
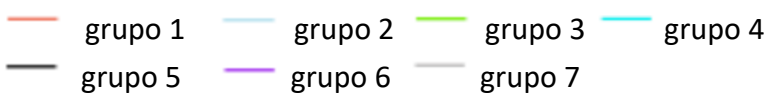


Fig. 4.5.3: Coeficientes del modelo Group Lasso en función del parámetro λ



En el gráfico anterior aparecen los coeficientes del modelo en función del valor de λ. Cada grupo aparece de un mismo color. Se observa como los coeficientes se reducen a medida que aumenta la penalización. Pero lo interesante aquí es ver como efectivamente los grupos de variables suelen alcanzar al mismo tiempo el valor 0. Estos grupos se eliminan en función de su significación conjunta. Así por ejemplo el grupo 2 parece ser el más significativo ya que es el último en eliminarse.

Comprobemos el poder predictivo de este modelo:

Group Lasso	**RMSE**	**MAE**	$\boldsymbol{R^2_{adj}}$
Train	2.510	1.961	0.860
Test	4.310	3.370	0.616

Fig. 4.5.4: Tabla de medidas de error para los datos de *train* y *test* del modelo PCR

El error obtenido es el peor hasta ahora. Como ya se ha dicho, esta técnica suele ser empleada en otro tipo de *datasets*, por lo que no sorprenden los resultados.

4.6. Regresión por reducción de dimensionalidad

En los métodos vistos anteriormente se emplean las variables originales como regresores, sin modificarlas (a parte de estandarizarlas). Seguidamente se verán los métodos explicados en el apartado 2.4.5, que se basan en utilizar como regresores las variables resultantes del análisis de componentes principales.[42]

4.6.1. Regresión por componentes principales

Se ajusta a continuación una regresión por componentes principales (PCR). Se utilizará para ello la función *pcr()* del paquete *pls*. Recordemos que este modo consiste en realizar una regresión lineal, pero utilizando como regresoras una subconjunto de las componentes principales de las variables explicativas.

Para elegir cuantas componentes usar se calcula, para cada número de componentes (i.e. 1era componentes, 1era y 2nda componentes, …, todas las componentes), la raíz del error cuadrático medio (RMSE) por validación cruzada. En este caso se utiliza el *leave-one-out cross-validation*:

[42] téngase en cuenta que estas componentes no serán exactamente iguales que las obtenidas en apartados anteriores, en este caso la variable respuesta *esperanza_vida* no se tiene en cuenta a la hora de calcular las componentes

componentes	RMSE	diferencia
<dbl>	<dbl>	<dbl>
0	7.49	0
1	3.07	-0.591
2	3.09	0.006
3	3.09	0.003
4	3.12	0.007
5	3.14	0.006
6	3.14	0.003
7	3.21	0.022
8	3.29	0.023
9	3.16	-0.037
10	3.13	-0.011
11	3.01	-0.039
12	2.99	-0.005
13	3.03	0.014
14	2.95	-0.027
15	2.96	0.003
16	3.07	0.036
17	3.15	0.026
18	3.17	0.008
19	3.22	0.014
20	3.10	-0.037
21	3.11	0.003
22	3.11	-0.001
23	2.93	-0.055

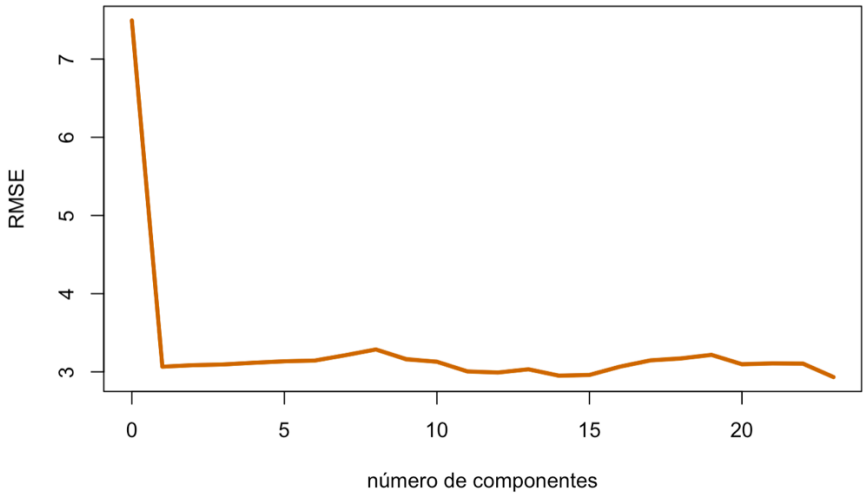


Fig. 4.6.1: evolución del RMSE de PCR en función del nº de componentes

Se observa como el mínimo se alcanza cuando se utilizan cada una de las 23 componentes. Pero sería poco interesante ajustar el modelo con todas las componentes (so obtendrían exactamente los mismos resultados que con la regresión lineal pero perdiendo en interpretabilidad). Por ello se deciden coger las 14 primeras componentes, ya que el error es prácticamente el mismo (2.93 por 2.95). Se puede además comprobar que estas primeras componentes explican gran parte de la varianza total:

TRAINING: % variance explained

	1 comps	2 comps	3 comps	4 comps	5 comps	6 comps	7 comps	8 comps	9 comps	10 comps	11 comps	12 comps
X	37.25	47.03	55.27	61.49	66.43	70.80	75.10	78.65	81.56	84.31	86.86	89.16
esperanza_vida	83.97	84.04	84.30	84.34	84.47	84.48	84.74	84.82	86.46	86.61	87.55	87.59

	13 comps	14 comps	15 comps	16 comps	17 comps	18 comps	19 comps	20 comps	21 comps	22 comps	23 comps
X	90.96	92.63	94.17	95.66	96.84	97.76	98.59	99.28	99.65	99.92	100.00
esperanza_vida	87.61	88.14	88.15	88.19	88.20	88.59	88.60	89.32	89.43	89.59	90.97

Fig. 4.6.2: % de varianza total y de la variable respuesta explicada acumulada[43]

Se comprueba como en efecto las primeras 14 componentes explican el 92.63% de la varianza total, y el 88.14% de la varianza de la variable respuesta (las 23 componentes explican el 90.97% de la variable respuesta, lo cual equivale al coeficiente de determinación R^2 de la regresión lineal). Veamos si este modelo PCR con 14 componentes nos ha servido para mejorar la capacidad de predicción:

[43] la fila 'X' muestra la varianza total acumulada y la fila 'esperanza_vida' muestra la varianza de la variable respuesta acumulada que explican las componentes

PCR	**RMSE**	**MAE**	R^2_{adj}
Train	2.558	2.041	0.850
Test	4.579	3.606	0.641

Fig. 4.6.3: Tabla de medidas de error para los datos de *train* y *test* del modelo PCR

Parece ser que de momento este método es el que peor se ajusta a los datos de test. Aún así este modelo parecer funcionar más que correctamente. Recordemos, que anteriormente se ha dicho que este método puede sufrir en presencia de componentes con poca varianza altamente correlacionadas con la variable respuesta. Se muestra aquí la correlación de cada componente con la variable respuesta *esperanza_vida*:

```
Comp 1  Comp 2  Comp 3 Comp 4 Comp 5  Comp 6 Comp 7  Comp 8 Comp 9 Comp 10 Comp 11 Comp 12
 0.916 -0.0263 -0.0506 0.0208 0.0364 0.00244 0.0517 -0.0286 -0.128  0.0386 -0.0966  0.0198
Comp 13 Comp 14  Comp 15 Comp 16  Comp 17 Comp 18  Comp 19 Comp 20 Comp 21 Comp 22 Comp 23
-0.0143  0.0734 -0.00406  -0.021 -0.00729  0.0631 -0.00893   0.085 -0.0321 -0.0399  -0.118
```

Fig. 4.6.4: Correlación entre las componentes principales y la variable respuesta

Efectivamente parece ser que no se está dejando fuera ninguna componente altamente correlaciona con la respuesta. Además se comprueba que la única componente altamente correlacionada es la primera. De ahí que el modelo arroje resultados aceptables.

4.6.2. Regresión por mínimos cuadrados parciales

Otro método de regresión por reducción de la dimensionalidad es el de la regresión por mínimos cuadrados parciales, o PLS. Como ya se ha comentado, este método se diferencia de PCR en que ahora sí se tiene en cuenta la variable respuesta a la hora de calcular las componentes. Para ajustar este modelo se utilizará la función *plsr()* del paquete *pls* (esta función utiliza el algortimo *kernel*).

Al igual que para PCR, se pueden elegir el número de componentes a utilizar mediante validación cruzada, de nuevo por *leave-one-out cross-validation*:

componentes	RMSE	diferencia
<dbl>	<dbl>	<dbl>
0	7.49	0
1	3.04	-0.594
2	3.11	0.02
3	3.06	-0.016
4	3.03	-0.007
5	3.08	0.015
6	3.11	0.01
7	3.07	-0.013
8	3.02	-0.018
9	2.96	-0.02
10	2.93	-0.007
11	2.94	0.003
12	2.96	0.007
13	2.96	-0.001
14	2.93	-0.01
15	2.92	-0.002
16	2.93	0.003
17	2.94	0.001
18	2.94	0
19	2.93	0
20	2.93	0
21	2.93	0
22	2.93	0
23	2.93	0

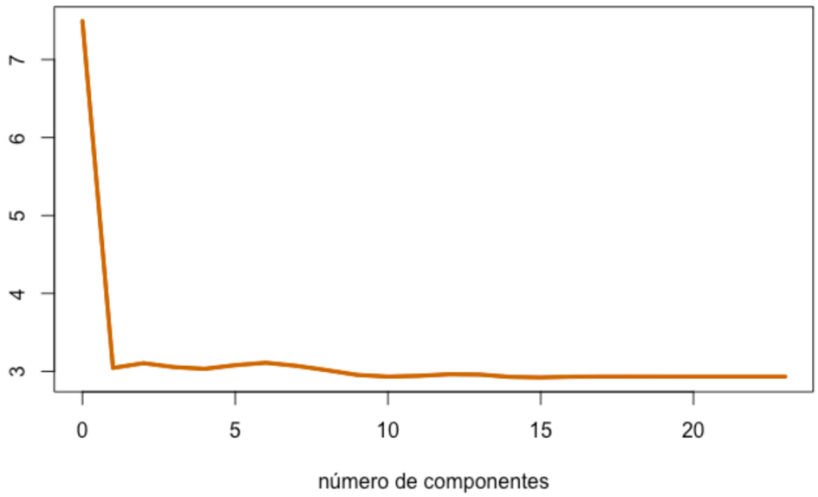


Fig. 4.6.5: evolución del RMSE de PCR en función del nº de componentes

En este caso el RMSE mínimo se obtiene con 15 componentes. Pero se deciden coger 10 componentes ya que el error es prácticamente el mismo (2.92 por 2.93). Por otra lado, se observa que se converge mucho más rápido que en el caso de PCR hacia el mínimo error. Esto es por supuesto debido a que ahora se forman las componentes de forma que se maximice la covarianza entre los regresores y la respuesta. Es decir, las primeras componentes serán las que tengan más correlación con la variable respuesta. Y por lo tanto el añadir las últimas componentes apenas mejorará el modelo. Se puede comprobar la correlación entre las componentes y las variables respuesta:

```
Comp 1 Comp 2 Comp 3 Comp 4 Comp 5 Comp 6 Comp 7 Comp 8 Comp 9 Comp 10 Comp 11 Comp 12 Comp 13
  0.92  0.183 0.0957 0.0646 0.0562 0.0471 0.0422 0.0444 0.0514  0.0433   0.028  0.0273  0.0175
Comp 14 Comp 15 Comp 16 Comp 17 Comp 18  Comp 19  Comp 20   Comp 21    Comp 22     Comp 23
 0.0144 0.00482 0.00215 0.00188 0.00159 0.000522 0.000141 0.0000316 0.00000785 0.000000232
```

Fig. 4.6.6: Correlación entre las componentes principales y la variable respuesta

De nuevo, la primera componente parece ser la única muy correlacionada con la respuesta (es ahora algo mayor que para PCR). Pero también se ve ahora que las siguientes componentes están algo más correlacionadas, y que las últimas están completamente incorrelacionadas.

Otra forma de ver que efectivamente PLS requiere menos componentes que PCR para alcanzar el mismo nivel de predicción es mirar la varianza de la variable respuesta explicada por las componentes:

TRAINING: % variance explained

	1 comps	2 comps	3 comps	4 comps	5 comps	6 comps	7 comps	8 comps	9 comps	10 comps	11 comps	12 comps
X	37.23	42.26	48.63	53.82	59.36	63.61	67.32	70.74	73.32	76.04	78.61	80.29
esperanza_vida	84.73	88.07	88.98	89.40	89.71	89.94	90.11	90.31	90.57	90.76	90.84	90.91
	13 comps	14 comps	15 comps	16 comps	17 comps	18 comps	19 comps	20 comps	21 comps	22 comps	23 comps	
X	82.66	83.65	85.76	88.85	91.11	92.40	94.49	96.03	97.51	98.46	100.00	
esperanza_vida	90.95	90.97	90.97	90.97	90.97	90.97	90.97	90.97	90.97	90.97	90.97	

Fig. 4.6.7: % de varianza total y de la variable respuesta explicada acumulada

Primeramente, se observa que con 14 componentes ya se explica el máximo de varianza posible de la variable respuesta *esperanza_vida* (90.97%), mientras que para PCR hacían falta todas las 23 componentes. Si por ejemplo se cogen las dos primeras componentes de PLS, se está explicando el 88% de la varianza, mientras que en PCR hacen falta 14 componentes para llegar a este valor.

Veamos si efectivamente este modelo predice mejor en los datos de test:

PLS	**RMSE**	**MAE**	$\boldsymbol{R^2_{adj}}$
Train	2.258	1.764	0.883
Test	4.153	3.118	0.726

Fig. 4.6.8: Tabla de medidas de error para los datos de *train* y *test* del modelo PLS

Efectivamente todas las métricas de error son bien inferiores a las obtenidas en el modelo PCR. Además este modelo está consiguiendo ser el mejor hasta ahora en términos de $\mathrm{R}^2_{\mathrm{adj}}$. Aún así, en cuanto al RMSE y al MAE, la regresión lineal sigue siendo la más efectiva, aunque PLS supera a Lasso y a ElasticNet. Obviamente esto no quiere decir que este modelo sea el preferido hasta ahora (hay que tener en cuenta que PLS es mucho menos interpretable que ElasticNet o Lasso).

A continuación se hace una comparación de los modelos utilizados y se intenta llegar a una conclusión.

4.7. Comparación de modelos

Han sido varios los métodos de modelización implementados a lo largo del trabajo, cada uno con unas características y particularidades diferentes. El objetivo de este apartado es comparar la calidad de ajuste de estos métodos. A continuación se muestran las métricas de error obtenidas en los datos de test con cada modelo:

Test	RMSE	MAE	R2_adj
Regresión Lineal	4.09	3.04	0.654
Ridge	4.10	3.09	0.652
Lasso	4.29	3.34	0.703
ElasticNet	4.26	3.32	0.701
Group Lasso	4.31	3.37	0.616
PCR	4.58	3.61	0.641
PLS	4.15	3.12	0.726

Fig. 4.7.1: Métricas de error de los datos de test de cada modelo

Sorprendentemente (o no), la regresión lineal por mínimos cuadrados ordinarios ha obtenido los mejores resultados en términos de RMSE y MAE. Decimos sorprendentemente, porque uno de los objetivos de este trabajo era justamente usar métodos de regularización que mejorasen la calidad de predicción. Que esto no se haya conseguido aquí, no significa obviamente que estos métodos no funcionen o no se deban de usar. Cada conjunto de datos es único y no se puede saber de antemano que modelo utilizar.

Además las métricas de error pueden no ser siempre lo más importante. Se puede preferir por ejemplo un modelo que realice una selección de variables y que sea más interpretable, entonces se escogería Lasso o Elasticnet. Si la interpetabilidad no es algo fundamental uno podría decantarse por PLS (PLS es el mejor modelo en términos de R^2_{adj}).

En definitiva, la respuesta es que en este caso particular, si lo que se quiere es puramente maximizar la predicción se escogería la regresión lineal. Si hablamos en términos genéricos (i.e. otros datos u otro problema), será trabajo del investigador ponderar y decidir que modelo prefiere en función de las características de cada modelo, así como de por supuesto la capacidad predictiva que obtenga con cada modelo.

5. *Open Data* como impulso a la economía y al bienestar de la sociedad

Los datos utilizados a lo largo de este trabajo han sido obtenidos de la base de datos pública del Banco Mundial, los datos eran por lo tanto abiertos y libres de uso para todo el mundo. Este estudio ha servido así como buen ejemplo de las posibilidades que tienen para ofrecer los denominados *Open Data*.

¿Qué significa *Open Data*?

Open data, o datos abiertos, son datos que tanto gobiernos, empresas como individuos pueden acceder, compartir y usar libremente.

Para que unos datos se consideren abiertos deben cumplir:

- Disponibilidad: los datos deben estar completos y gratuitamente descargables de Internet
- Accesibilidad: deben proporcionarse en un formato conveniente sin necesidad de registro
- Actualizados: deben publicarse los más rápido posible cada vez que estos cambien
- Sin restricciones: que cualquiera pueda usar, modificar y compartir los datos, sin importar su propósito (comercial, no comercial, educacional, etc)

Beneficios de los *Open Data*

Los datos abiertos tienen el potencial de estimular significativamente el crecimiento económico de un país así como el bienestar de sus habitantes. A lo largo de los últimos años ha habido un movimiento favorable a los *Open data*. Especialmente por parte de gobiernos y de organismos, que han comenzado a ver las ventajas que ofrecen estos en múltiples ámbitos, tanto en términos de efectividad y eficiencia.

Primeramente, el uso de datos abiertos aumenta la transparencia de las entidades. En efecto, la tendencia hacia los datos abiertos permite al público mantenerse informado, y seguir el día a día de las operaciones llevadas a cabo por su gobierno local. Se desarrolla así una mayor confianza, credibilidad y reputación de la administración. De esta manera se fomenta también la participación de los ciudadanos en la formulación de políticas. Debido a la mayor transparencia, los ciudadanos pueden estudiar mejor los diferentes reportes y formar su propia opinión.

Por otro lado, los *Open Data* no solo mejoran la imagen y la eficiencia de estos organismos, sino que también son un gran motor para el crecimiento económico. El uso de datos abiertos ayuda a las empresas y a las administraciones públicas a reducir costes debido a que aumenta

la eficiencia de trabajo y a generar más beneficios ya que se crean nuevos servicios. Es decir, los *Open Data* promueven el progreso y la innovación. Se brindan nuevas oportunidades para aplicaciones comerciales, se reduce el tiempo de investigación, mejora la calidad de la información, etc. Por supuesto, también se crea empleo a raíz de las necesidades de los datos abiertos (recolectores de datos, científicos de datos, ingenieros de bases de datos, etc). De hecho, el Portal Europeo de Datos, en un informe publicado sobre el valor económico de los datos abiertos en Europa, calcula en un marco optimista, que el sector *Open Data* podría llegar a alcanzar los 1.97 millones de empleados directos e indirectos, y un tamaño de mercado de hasta 334.20 millones de euros.

Los *Open Data* son por lo tanto muy positivos para la sociedad, y no solo en términos económicos. Multitud de sectores se ven beneficiados:

- Los efectos sobre la sanidad y la ciencia son claros, el libre uso e intercambio de datos médicos y científicos fomenta la colaboración entre investigadores, acelerando así las investigaciones
- Ayuda a la educación ya que dota a los enseñantes de material más ilustrativo y contrastado
- Los servicios públicos son más eficientes. Se optimizan servicios y recursos como el transporte público, atención de emergencias, recogida de basura, provisión de electricidad etc
- El medioambiente también se ve beneficiado. Los datos permiten concienciar y 'convencer' a los ciudadanos del daño que se está generando a la naturaleza y al planeta en general. Además la optimización en sectores como el transporte, mencionado anteriormente, y el uso más eficiente de los recursos conlleva un claro beneficio para todos
- Otros muchos sectores como el ocio, el turismo o los servicios informativos se ven por supuesto también beneficiados

Ejemplos de uso de los *Open Data*

A continuación se lista una serie de ejemplos de proyectos y aplicaciones de distintos ámbitos que ilustran las posibilidades de los datos abiertos:

- Sanidad

PADRIS: El *Programa d'analítica de dades per a la recerca i la innovació en salut* (PADRIS), es un programa de análisis de datos para la investigación y la innovación en salud. Su misión principal es poner a disposición de la comunidad científica los datos sanitarios generados por el sistema sanitario integral de utilización pública de Cataluña (SISCAT).

Consorcio de Genómica Estructural: El Consorcio de Genómica Estructural (SGC), es una asociación público-privada que investiga las enfermedades de todas las proteínas codificadas por el genoma humano. El SGC pone todos sus resultados a disposición de la comunidad científica, de esta manera se crea una red colaborativa abierta de científicos de diferentes instituciones académicas y compañías farmacéuticas.

OpenVaccine, COVID-19 mRNA Vaccine Degradation Prediction: Publicado en setiembre de 2020 por la Universidad de Stanford en la popular comunidad online de *machine learning* Kaggle, OpenVaccine fue un concurso abierto cuyo objetivo era construir modelos que predijesen lo mejor posible las tasas de degradación probables en cada base de una molécula de ARN. Siendo el objetivo primordial el de mejorar la estabilidad de las vacunas ARNm contra el virus SARS-CoV-2.

- Territorio y Medio ambiente

PlatgesCat: Aplicación que permite consultar la calidad, la características y el estado de las playas y las zonas de baño de Cataluña

Ecofacts: Ecofacts es una aplicación que provee información, de cualquier país del mundo, sobre el consumo de energía y el cambio climático. Su objetivo principal es concienciar a los ciudadanos y las organizaciones de cómo sus acciones pueden afectar al medio ambiente.

- Educación

Escuelas Comciencia: Impulsado por el grupo de investigación Ciberimaginario, escuelas Comciencia es una iniciativa que buscar acercar a estudiantes de Secundaria y Bachillerato el conocimiento y la investigación científica. Su método se basa en que los estudiantes aprendan y comprendan los conceptos científicos a través del tratamiento y el uso de datos abiertos.

College Affordability and Transparency Center: Diseñado por el departamento de educación de Estados Unidos, el CATC tiene por objetivo ofrecer información sobre el coste de las universidades. A través de datos abiertos, esta aplicación da a sus usuarios las herramientas necesarias para comparar matrículas, precios y otras características.

- Economía y empresa

Eixos: Eixos ofrece datos e informes económicos a los agentes económicos interesados (inversores, particulares, emprendedores, administraciones públicas, etc). Opera principalmente en Cataluña pero su actividad se extiende a otras ciudades como Madrid o Sevilla. Los datos que ofrece se actualizan periódicamente y provienen de las administraciones públicas y de entidades privadas.

Indicadores de contratación pública de la gencat: Serie de aplicativos sostenidos por datos públicos que permiten analizar y visualizar diferentes ámbitos de la contratación pública de la *Generalitat de Catalunya* y de su sector público.

Open data en números

Cada año, el Portal Europeo de Datos publica un informe que recopila una serie de indicadores comparativos para analizar y evaluar el nivel de desarrollo de los datos abiertos de los países europeos. Estos datos nos permiten comparar países en términos de desarrollo de *Open Data,* así como medir la evolución de estos a nivel europeo. Los datos se recolectan a través de un cuestionario enviado a los representantes nacionales colaboradores con la Comisión Europea y con el Grupo de Expertos en Información del Sector Público. Estos datos se publican en la web del Portal Europeo en forma de *dashboard*.
Los indicadores que ahí se encuentran miden el grado de madurez de los países europeos en términos de *Open Data*. El grado de madurez se divide en cuatro ámbitos: políticas, portal, impacto y calidad. Cada uno de estos ámbitos esta compuesto de varios indicadores. El valor en estos indicadores serán los que a su vez determinen la puntuación de un país en cada uno de los ámbitos mencionados. La explicación de los indicadores y el sistema de puntuación viene explicado en este documento.

Del informe se destaca que España ocupa la tercera plaza en este ranking, por detrás de Dinamarca y Francia. Además el Portal Europeo de Datos publica un informe personalizado de cada país en la lista, donde entre otras cosas se detallan puntos en los que mejorar. En el caso de España, se destaca que se debe seguir trabajando en armonizar el trabajo entre los diferentes niveles de la administración (central, regional, local).

6. Conclusiones

Seguidamente, se exponen las conclusiones del trabajo de fin de grado y se proponen líneas de interés para seguir trabajando en la misma dirección. Además, y ya como punto final, se escribe un pequeño párrafo donde daré mi opinión más personal, tanto de lo positivo como de lo negativo, de lo que ha sido la realización de este proyecto.

Primeramente, se espera que este trabajo haya servido como modelo de los pasos más habituales que se suelen seguir en un proyecto típico de *Machine Learning*. Aunque obviamente puedan existir infinidad de procedimientos diferentes, incluso el lector menos especializado debería haber visto que las etapas suelen ser algo parecido a esto: definir el problema, extraer y preprocesar los datos, explorar los datos, entrenar el/los modelos (ajustando los posibles parámetros), probar los modelos sobre datos desconocidos, evaluar la capacidad predictiva y finalmente utilizar el modelo para realizar predicciones futuras.

Por otro lado, en lo referente a los datos y a las técnicas utilizadas, se ha querido dar al lector una herramienta para que el mismo explore y 'juegue' con estos datos y técnicas. Esta herramienta es por supuesto la aplicación web mencionada varias veces con anterioridad. En esta, el lector puede navegar y descubrir características de los países en referencia a los indicadores de desarrollo utilizados. La aplicación debe servir también como herramienta para comprender mejor las técnicas utilizadas a lo largo del trabajo. Es decir, los datos utilizados son muy entendibles por la gran mayoría del público, por lo que efectivamente, este hecho combinado con el uso de la aplicación interactiva debería ser de gran ayuda para comprender mejor ciertos conceptos.

En cuanto a las técnicas de regularización, la realización de este trabajo ha servido para entender mejor el funcionamiento de tales técnicas. Desde la búsqueda de los parámetros óptimos por validación cruzada, hasta la evaluación de la calidad de predicción. El objetivo no era tanto lograr la máxima capacidad de predicción posible de la esperanza de vida, sino utilizar, comparar y comprender los métodos utilizados. Cada conjunto de datos es único, así como las preferencias del investigador, por lo que cada problema puede tener una solución distinta. Uno debe evaluar si dar más o menos importancia a la selección de variables, a la interpretabilidad del modelo, al coste computacional, etc. Y a partir de ahí elegir el modelo que mejor le convenga para realizar sus predicciones futuras.
Así por ejemplo, en este caso, yo hubiese elegido la regresión lineal por mínimos cuadrados. Es verdad que Lasso y ElasticNet eliminaban variables del modelo, haciéndolo más simple, sin empeorar demasiado las predicciones. Pero seguramente hubiésemos conseguido resultados similares simplemente eliminando las variables menos significativas de la regresión lineal. En situaciones donde la cantidad de variables fuese bien superior, entonces si que recomendaría Lasso y ElasticNet para seleccionar variables, ya que estos realizan la selección automáticamente, mientras que con la regresión lineal sería un trabajo manual. También se ha visto que la regresión por mínimos cuadrados parciales obtiene muy buenos resultados, aunque por supuesto se pierde mucho en interpretabilidad. En sectores donde esto no es lo primordial, como el farmacéutico o el químico, PLS si que podría llegar a ser muy útil. En definitiva, que en este caso en particular no hayan funcionado bien ciertos métodos, no quiere decir que estos no sean válidos o peores.

Por otra parte, se espera que este trabajo haya servido para mostrar la gran desigualdad que aún existe en el mundo. Como se ha visto en el *profiling*, nosotros nos encontramos en el grupo de países que sin duda mejor viven. Este grupo tan solo representa el 23% de la población mundial. Esto quiere decir que más de tres cuartos de la población viven en inferioridad de condiciones. Si se tiene en cuenta el casi 15% de la población mundial que conforma uno de los *clusters*, estamos hablando de que más de 1000 millones de personas viven en países donde las condiciones son pésimas. En estos países la esperanza de vida media es apenas de 61 años. Esto es alrededor 20 años menos que en por ejemplo, España.

El trabajo ha servido también como gran ejemplo del potencial que tienen los datos abiertos, tanto en la economía como en la sociedad en general. Estos pueden efectivamente impulsar numerosos sectores como la sanidad, la educación, los servicios públicos o el medio ambientes entre otros.

El *Machine Learning* es un campo muy amplio, cubre multitud de técnicas y modelos diferentes, aquí solo se ha visto una pequeña parte de las posibilidades que presenta. Una posible dirección en la que seguir trabajando sería por lo tanto probar otros métodos de regularización (e.g. *sparse group lasso*, Bridge, L-PLS, etc). Sería también interesante probar las técnicas utilizadas en otros datos, de esta manera veríamos como se comportan los modelos en diferentes situaciones.

Me gustaría concluir el trabajo de fin de grado con una valoración personal de lo que ha sido esta experiencia. Primeramente, me gustaría destacar que mi carrera profesional va encaminada hacia el mundo del dato. De hecho, estos últimos meses he compaginado la realización del TFG con unas prácticas a tiempo completo como *Data Scientist*. Esto me ha permitido intercambiar conocimientos entre ambas cosas, y me ha motivado para realizar lo mejor posible este trabajo de fin de grado. Es decir, lo he visto como una oportunidad para seguir formándome y no como una obligación.
Entrando en cuestiones más técnicas, decir que lo que me ha resultado más interesante y placentero ha sido el desarrollo de la aplicación web. A lo largo de la carrera no había hecho nada parecido. Además, durante el desarrollo de la aplicación he tenido la ocasión de aprender nuevos lenguajes de programación como CSS y HTML.
Por supuesto, también he encontrado complicaciones durante estos meses. Principalmente, tuve dificultades en encontrar una base de datos que me gustase y que no estuviese ya muy explotada. Obviamente, como ya sabemos, la cantidad de datos abiertos es enorme, pero quería una base de datos sin preprocesar y que estuviese relacionada con la economía y el desarrollo mundial.

7. Bibliografía

Bevans, R. (2020, February 20). *An introduction to multiple linear regression*. Retrieved from Scribbr: https://www.scribbr.com/statistics/multiple-linear-regression/

Brownlee, J. (2018, May 23). *A Gentle Introduction to k-fold Cross-Validation*. Retrieved from Machine Learning Mistery: https://machinelearningmastery.com/k-fold-cross-validation/

Cheng, J. (s.f). *leaflet*. Retrieved from RDocumentation: https://www.rdocumentation.org/packages/leaflet/versions/2.0.4.1

Escobar, M., Cruz, C. D., & Rincón, C. (Diciembre de 2003). Partial least squares (PLS) regression and its application to coal analysis. *Revista Técrnica de la Facultad de Ingenieria, Universidad de Zulia*. Obtenido de Scielo.

Friedman, J., Hastie, T., & Tibshirani, R. (2010, February 11). *A note on the group lasso and a sparse group lasso.* Retrieved from https://statweb.stanford.edu/~tibs/ftp/sparse-grlasso.pdf

Garbade, D. M. (2018, September 12). *Understanding K-means Clustering in Machine Learning*. Retrieved from Towards Data Science: https://towardsdatascience.com/understanding-k-means-clustering-in-machine-learning-6a6e67336aa1

Generalitat de Catalunya. (s.f). *Open Data*. Retrieved from gencat: https://web.gencat.cat/en/actualitat/reportatges/dades-obertes/

Hastie, T. (s.f). *glmnet*. Retrieved from RDocumentation: https://www.rdocumentation.org/packages/glmnet/versions/4.1/topics/glmnet

Jaadi, Z. (2019, September 4). *A STEP-BY-STEP EXPLANATION OF PRINCIPAL COMPONENT ANALYSIS*. Retrieved from BuiltIn, Expert Contributor Network: https://builtin.com/data-science/step-step-explanation-principal-component-analysis

James, G., Hastie, T., Witten, D., & Tibshirani, R. (2017). *An Introduction to Statistical Learning: with Applications in R.* Springer Science .

Kant, A., & Sekhri, D. G. (2021). Data can be an asset for governance, growth and public welfare. *Hindustan Times*.

Rstudio. (s.f). *Shiny*. Obtenido de Rstudio: https://shiny.rstudio.com/

The World Bank Group. (s.f). *DataBank*. Retrieved from The World Bank: https://databank.worldbank.org/home.aspx

The World Bank Group. (s.f). *Starting an Open Data Initiative*. Retrieved from The World Bank: http://opendatatoolkit.worldbank.org/en/starting.html

Tibshirani, R. (1996). *Regression Shrinkage abd Selection via the Lasso*. Retrieved from Journal of the Royal Statistical Society. Series B, Volume 58 Issue 1, pp 267-288: https://statweb.stanford.edu/~tibs/lasso/lasso.pdf

Zou, H., & Hastie, T. (2003, August). *Regularization and Variable Selection via the Elastic Net.*

8. Anexo

Todo el código se encuentra disponible en el siguiente repositorio:
github.com/AlvaroGarnica/TFG