TRABAJO FIN DE GRADO

REGRESIÓN SOBRE COMPONENTES PRINCIPALES

Doble Grado en
Matemáticas y Estadística

Irene Deduy Guerra

Sevilla, Junio de 2019

Índice general

Resumen

La presencia de multicolinealidad en los Modelos de Regresión Múltiple conduce a problemas en las estimaciones y resultados poco fiables. La aplicación de la Regresión sobre Componentes Principales puede evitar estos problemas, a la vez que lleva implícito un procedimiento de selección de variables, reduciendo la dimensión del espacio predictor. El objetivo de este trabajo es la descripción teórica y metodológica de la técnica estadística y su implementación en R, con una ilustración sobre datos reales. Además, con objeto de ilustrar la aplicabilidad del método, se incluye referencias sobre trabajos científicos recientes en los que se ha hecho uso de la misma.

Palabras claves: Regresión Múltiple, Multicolinealidad, Componentes Principales.

Abstract

The existence of multicollinearity in Multiple Regression Models leads to problems in estimation and unreliable results. The application of Principal Components Regression can avoid these problems, while involving a variable selection procedure, reducing the dimension of the predictor space. The aim of this work is the theoretical and methodological description of the statistical technique and its implementation in R, with an illustration on real data. In addition, in order to illustrate the applicability of the method, references of recent scientific works in which it has been used are included.

Key words: Multiple Regression, Multicollinearity, Principal Components.

Capítulo 1

INTRODUCCIÓN

En el Modelo de Regresión Lineal Múltiple, una de las hipótesis que deben cumplir las variables predictoras es la ausencia de relación lineal entre ellas. Sin embargo, en la práctica, es muy frecuente que se presenten ciertas relaciones aproximadamente lineales entre las variables predictoras del modelo, provocando la obtención de estimadores poco precisos e inestables. Además, la relación entre las variables dificulta cuantificar con precisión el efecto que cada variable ejerce sobre la dependiente, y como consecuencia las varianzas de los estimadores son elevadas. Esta fuerte correlación entre las variables explicativas del modelo se denomina "multicolinealidad".

La Regresión sobre Componentes Principales (RCP) es un método introducido por Kendall (1957) para combatir la multicolinealidad. Con este método, las variables originales se transforman en un nuevo conjunto de variables incorreladas llamadas componentes principales. Esta transformación clasifica las nuevas variables ortogonales por orden de importancia, es decir, según la contribución que ofrezcan en el modelo, y el procedimiento implica eliminar algunas de dichas componentes para lograr una reducción de la varianza. A continuación, se realiza un análisis de regresión múltiple de la variable dependiente en función del conjunto reducido de componentes principales utilizando la estimación de mínimos cuadrados ordinarios. Como las componentes principales son ortogonales, dicho método es apropiado. Finalmente, una vez calculados los coeficientes de regresión para el conjunto reducido de variables ortogonales, se transforman en un nuevo conjunto de coeficientes que corresponden al conjunto de variables inicial.

En este Trabajo de Fin de Grado, se desarrolla el fundamento teórico de la Regresión sobre Componentes Principales y se presenta su implementación en R. Además, se ilustra la metodología con ejemplos de aplicaciones realizadas en ámbitos tan distintos como lo son la agricultura, la sociología y la dendrología (estudio de plantas arboladas).

Capítulo 2

REGRESIÓN SOBRE COMPONENTES PRINCIPALES

2.1. Hipótesis Previas

Se considera el modelo de Regresión Lineal

$$Y = X\beta + \epsilon \qquad (1)$$

donde

- Y : vector $n-$dimensional compuesto por las observaciones de la variable dependiente.
- X : matrix $(n \times p)$ cuyo (i, j)-ésimo elemento representa el valor de la j-ésima variable predictora en la i-ésima observación.
- β : vector $p-$dimensional de coeficientes de regresión a estimar.
- ϵ : vector $n-$dimensional de errores aleatorios, considerados independientes e idénticamente distribuidos según $N(0, \sigma^2 I_n)$.

En primer lugar se estandarizan las variables, tanto la dependiente como las predictoras, restándole su media y dividiendo por su desviación típica. Este proceso es necesario pues las variables pueden estar medidas en diferentes unidades, lo cual podría ocasionar confusión en la interpretación del análisis. Se hace notar que, consecuentemente, los cálculos se basan en las variables estandarizadas, luego los coeficientes de regresión obtenidos del análisis deben ser posteriormente reajustados a la escala original. Por simplicidad de notación, en adelante se asume que las variables del modelo (1) están estandarizadas.

Se utiliza el procedimiento de Mínimos Cuadrados Ordinarios para estimar las constantes desconocidas β mediante $\hat{\beta}$ de modo que minimicen la suma de cuadrados de los residuos.

- Si la matrix $(X'X)$ es no singular, el problema tiene solución y los coeficientes de regresión se estiman a través de:

$$\hat{\beta} = (X'X)^{-1}X'Y$$

El estimador de mínimos cuadrados obtenido es insesgado y de mínima varianza. En particular: $v(\hat{\beta}) = \sigma^2(X'X)^{-1}$

- Si dicha matriz tiene determinante cercano a cero, los estimadores del método de mínimos cuadrados son poco precisos y no son fiables pues presentan una elevada correlación y gran varianza. En este caso, se dice que existe "multicolinealidad", es decir, dependencia lineal entre las variables predictoras, dificultando evaluar de forma precisa el efecto que tiene cada variable predictora sobre la variable dependiente.

Existen muchas formas para detectar la multicolinealidad. Para poder tomar una decisión al respecto, se recomienda basarse en los resultados de algunas de las siguientes técnicas:

- Gráfico de dispersión matricial: permite tener una idea sobre la posible relación lineal entre cada par de variables predictoras.
- Matriz de correlación R de las variables predictoras, y su inversa R^{-1}: como las variables están estandarizadas, la matriz de covarianzas muestrales $\hat{\Sigma}$ coincide con la matriz de correlación R;

$$\hat{\Sigma} = \frac{1}{n}\sum_{i=1}^{n}(x_i - \tilde{x})(x_i - \tilde{x})' = \frac{1}{n}\sum_{i=1}^{n} x_i x_i' = \frac{1}{n}X'X$$

es decir, $R = \frac{1}{n}X'X$. La existencia de algún valor fuera de la diagonal de R que sea próximo a ± 1, indica que existe una fuerte relación lineal entre las variables predictoras. Por otro lado, los elementos de la diagonal de R^{-1} se llaman Factores de Inflación de Varianza y verifican

$$FIV(i) = \frac{1}{1 - R_i^2}$$

para cada elemento de la diagonal, donde R_i es el coeficiente de determinación de la regresión de la variable X_i en función de las demás variables predictoras. De esta forma, un valor alto de $FIV(i)$ indica multicolinealidad causada por la variable X_i. El problema está en el cálculo de la inversa R^{-1} cuando R tiene un determinante muy próximo a cero.

- Autovalores de la matriz de correlación R: en presencia de multicolinealidad, al menos uno de los autovalores de la matriz es próximo a cero. Equivalentemente, se estudia el Índice de Condición de la matriz R desarrollado por Belsley et al. (1980) y definido como

$$cond(R) = \sqrt{\frac{\lambda_s}{\lambda_1}}$$

donde los autovalores se ordenan tal que $\lambda_1 < \lambda_2 < \ ... \ < \lambda_s$. De esta forma, para valores del índice mayores que 30, se considera existencia de alta multicolinealidad, y para valores menores que 5 multicolinealidad débil. Este indicador de multicolinealidad es global para las variables predictoras del modelo, a diferencia del FIV que se trata de un indicador específico para cada variable.

Para tratar de corregir la multicolinealidad, se puede actuar de diversas formas:

- Se puede usar un subconjunto de variables predictoras en ausencia de multicolinealidad. Se pueden eliminar, basándose en un t-test, variables de la ecuación muy correlacionadas con las que se incluyen, o en lugar de eliminar directamente variables, se pueden considerar las componentes principales de dichas variables. De esta última forma, el problema de la multicolinealidad desaparece ya que las componentes principales son incorreladas, y por tanto los cálculos se simplifican. Además, este procedimiento de eliminación de variables predictoras tiene como consecuencia una reducción del número de parámetros que se han de estimar.
- Se puede introducir información externa a los datos originales mediante el uso de estimadores "contraídos" que proporciona el enfoque bayesiano.
- Se puede utilizar estrategias de "contracción" o "regularización" tratando de obtener estimadores que, aunque no sean insesgados, tengan varianzas pequeñas, es decir, menor error cuadrático medio.

En estos casos se sustituyen los estimadores de los coeficientes de regresión obtenidos a través de mínimos cuadrados por estimadores sesgados. A pesar de ello, estos estimadores presentan un error cuadrático medio considerablemente menor.

En este trabajo, se presenta la Regresión sobre Componentes Principales para solucionar el problema de multicolinealidad en los datos y reducir la dimensión del espacio de variables predictoras.

2.2. Análisis de Componentes Principales

El Análisis de Componentes Principales (ACP) es una técnica que consiste en transformar las variables correlacionadas en un número menor de variables construidas como combinaciones lineales de las originales. Este nuevo conjunto de variables incorreladas se denomina conjunto de "Componentes Principales" (CP). Esta metodología aplicada a la regresión permite hallar asociaciones entre las variables y reducir el número de éstas para facilitar su análisis e interpretación.

Se considera el modelo de regresión estandarizado (1). Los valores de las componentes principales de las variables predictoras para cada observación, vienen dados por

$$Z = XA \qquad (2)$$

donde:

- Z: matriz cuyo (i,k)-ésimo elemento representa el valor de la k-ésima componente principal en la i-ésima observación.
- A: matriz cuya k-ésima columna es el autovector unitario asociado al k-ésimo mayor autovalor de $\frac{1}{n}X'X$.

Se trata, por tanto, de hallar cada componente Z_k maximizando su varianza $v(Z_k) = a_k'\Sigma a_k$ con la restricción $a_k'a_k = 1$ y la condición de incorrelación con las demás CPs $a_k'\Sigma a_j = \lambda_k a_k'a_j = a_k'a_j = 0,\ k \neq j$, siendo a_k la k-ésima columna de A y λ_k el k-ésimo mayor autovalor de $\frac{1}{n}X'X$.

Algunas propiedades a destacar de las componentes principales muestrales:

- Las Z_i están incorreladas; $cov(Z_k, Z_j) = 0,\ k \neq j$
- $E(Z_k) = 0$
- $v(Z_k) = \lambda_k$
- $0 \leq\ v(Z_p) \leq\ ...\ \leq v(Z_1)$
- $\sum_{k=1}^{p} v(Z_k) = \sum_{k=1}^{p} v(X_k)$

2.3. Fundamento Teórico de la Regresión

Como consecuencia de las propiedades de las componentes principales, la matriz A es ortogonal y $AA' = A'A = I_p$. De esta forma, se puede expresar el Modelo de Regresión Lineal (1) como

$$Y = X\beta = XAA'\beta$$

Teniendo en cuenta (2) y tomando $\gamma = A'\beta$, se tiene (1) como

$$Y = Z\gamma + \epsilon$$

Dado que $A'(X'X)A = L^2$ donde L es una matriz diagonal cuyo elemento k-ésimo de la diagonal es $\lambda_k^{1/2}$, entonces $Z'Z = L^2$ y de esta forma el vector $\hat{\gamma}$ es

$$\hat{\gamma} = (Z'Z)^{-1}Z'Y = L^{-2}Z'Y$$

Sustituyendo dicha expresión en

$$\beta = A\gamma$$

se obtiene

$$\hat{\beta} = A(Z'Z)^{-1}Z'Y = AL^{-2}Z'Y$$

Teniendo en cuenta (2), se puede expresar $\hat{\beta}$ como

$$\hat{\beta} = \sum_{k=1}^{p} \lambda_k^{-1} a_k a_k' X'Y \qquad (3)$$

donde:

- a_k: k-ésima columna de A.

- λ_k: k-ésimo elemento diagonal de L^2.

Por otra parte, dado que $X'X = AL^2A'$ entonces $(X'X)^{-1} = AL^{-2}A'$ y por tanto

$$\hat{\beta} = AL^{-2}Z'Y = AL^{-2}A'XY = (X'X)^{-1}XY$$

es decir, el estimador de mínimos cuadrados ordinarios de los coeficientes de regresión sobre las p componentes principales $\hat{\gamma}$ es una transformación lineal del estimador de mínimos cuadrados ordinarios de los coeficientes de regresión sobre las p variables originales

$$\hat{\beta} = A\hat{\gamma}, \quad \hat{\gamma} = A'\hat{\beta}$$

Además los valores ajustados y residuos coinciden dado que:

$$X\hat{\beta} = XA\hat{\gamma} = Z\hat{\gamma}$$

Suponiendo que la matriz de varianzas-covarianzas de Y es $\sigma^2 I_n$, se deduce que la matriz de varianzas-covarianzas de $\hat{\beta}$ es

$$v(\hat{\beta}) = \sigma^2 A(Z'Z)^{-1}Z'Z(Z'Z)^{-1}A' = \sigma^2 A(Z'Z)^{-1}A' = \sigma^2 AL^{-2}A' = \sigma^2 \sum_{k=1}^{p} \lambda_k^{-1} a_k a_k' \qquad (4)$$

La multicolinealidad se presenta como una componente principal con varianza muy pequeña. Teniendo en cuenta que la varianza de la k-ésima CP es λ_k, una componente con varianza muy pequeña tiene valores bajos de λ_k, y por consecuente, valores muy grandes de λ_k^{-1}. Es decir, valores de los autovalores cercanos a cero provocan gran varianza del estimador. Para evitarlo, se eliminan los términos de (3) correspondientes a valores muy pequeños de λ_k, obteniéndose el estimador

$$\tilde{\beta} = \sum_{k=1}^{m} \lambda_k^{-1} a_k a_k' X'Y$$

de forma que λ_s corresponde a autovalores muy pequeños para $m < s \leq p$.

Esto es equivalente a determinar y eliminar los $p - m$ elementos de γ que no son significativamente distintos de cero, lo cual conllevaría un proceso de selección de variables cuyas variables son las componentes principales. Sin embargo, esto no es sencillo ya que también se desea eliminar los términos de gran varianza de (4).

El estimador $\tilde{\beta}$ es sesgado, pues como

$$\tilde{\beta} = \hat{\beta} - \sum_{k=m+1}^{p} \lambda_k^{-1} a_k a_k' X'Y$$

se tiene

$$E[\tilde{\beta}] = E[\hat{\beta}] - E[\sum_{k=m+1}^{p} \lambda_k^{-1} a_k a_k' X'Y] = \beta - \sum_{k=m+1}^{p} \lambda_k^{-1} a_k a_k' X'X\beta = \beta - \sum_{k=m+1}^{p} a_k a_k'\beta$$

Por tanto, en general, $E[\tilde{\beta}] \neq \beta$ y por este motivo se dice que la Regresión sobre Componentes Principales pertenece a los denominados "Métodos Sesgados de Regresión".

La matriz de varianzas-covarianzas de $\tilde{\beta}$ es

$$v(\tilde{\beta}) = \sum_{j=1}^{m} \lambda_j^{-1} a_j a_j' X'X \sum_{k=1}^{m} \lambda_k^{-1} a_k a_k'$$

De la descomposición espectral de

$$X'X = \sum_{h=1}^{p} \lambda_h^{-1} a_h a_h'$$

se tiene

$$v(\tilde{\beta}) = \sigma^2 \sum_{h=1}^{p} \sum_{j=1}^{m} \sum_{k=1}^{m} \lambda_h \lambda_j^{-1} \lambda_k^{-1} a_j a_j' a_h a_h' a_k a_k'$$

Los únicos términos no nulos aparecen cuando $h = j = k$, ya que los vectores son ortonormales. Por tanto:

$$v(\tilde{\beta}) = \sigma^2 \sum_{k=1}^{m} \lambda_k^{-1} a_k a_k'$$

De esta forma, si los m primeros autovalores no son muy pequeños, no habrá varianzas grandes en la diagonal de $v(\tilde{\beta})$, consiguiéndose disminuir la varianza del estimador $\tilde{\beta}$, en comparación con la varianza de $\hat{\beta}$; $v(\tilde{\beta}) < v(\hat{\beta})$.

En casos de acentuada multicolinealidad, la reducción de la varianza puede ser considerable, mientras que el sesgo introducido puede ser pequeño en comparación. De hecho, si los elementos de γ correspondientes a componentes eliminadas son realmente nulas, entonces no se introducirá sesgo en el estimador.

Se concluye que el problema de Regresión sobre Componentes Principales expresado como $Y = Z\gamma + \epsilon$, es equivalente a usar el modelo lineal $Y = X\beta + \epsilon$ y estimar β a través de

$$\tilde{\beta} = \sum_{M} \lambda_k^{-1} a_k a_k' X'Y$$

donde M es un subconjunto de los enteros $\{1, 2, ..., p\}$.

2.4. Selección de Componentes Principales en el Modelo

Uno de los objetivos principales del uso de las CP es la reducción de la dimensionalidad de las variables originales. Eliminando adecuadamente componentes principales se puede conseguir reducir la varianza total del modelo y proporcionar un mejor modelo de predicción. Sin embargo, la elección del subconjunto M no es tarea fácil, pues por un lado se pretenden eliminar componentes con varianzas muy pequeñas para mitigar el efecto

de la multicolinealidad, pero se debe tener en cuenta que no interesa eliminar componentes que proporcionen buenas predicciones de la variable dependiente.

Debido a la incompatibilidad, en ciertas ocasiones, de llevar a la práctica dichas directrices, han surgido distintas opiniones y métodos de actuación. A continuación, se muestran algunos de los métodos propuestos con esta finalidad.

La manera más sencilla de elegir M es eliminando aquellas componentes cuyas varianzas sean menor que un determinado valor m^*. La elección de dicho valor es arbitraria. Una forma de elegirlo es haciendo uso del Factor de Inflación de la Varianza (FIV) para cada variable predictora. Tal y como se mencionó en el apartado anterior, el FIV representa el incremento en la varianza debido a la presencia de multicolinealidad. Desechando los últimos términos en $\hat{\beta}$ los FIVs para los estimadores sesgados resultantes se reducirán. Esta eliminación continuará hasta que todos los FIVs se encuentren por debajo de un determinado valor. Si todas las variables son incorreladas, entonces todos los FIVs serían 1, y equivalentemente $R_i = 0$ para todo i. Según Belsley et al. (1980) existe multicolinealidad severa si el FIV es mayor que 10 ($R^2 = 0.9$), aunque Kleinbaum et al. (1988) rebajan el umbral de FIV a 5 ($R^2 = 0.8$).

El problema con esta técnica es que no siempre una componente con varianza pequeña resulta ofrecer poca contribución en el modelo. En el lado opuesto, en lugar de basar la elección enteramente en el tamaño de la varianza, se plantea el uso de t-tests permitiendo medir la contribución de cada componente en el modelo de regresión. Sin embargo, esta metodología también presenta sus inconvenientes, pues los t-tests para componentes con varianzas pequeñas tienen una potencia reducida en comparación con aquellos para componentes con varianzas grandes, por lo que son menos propensas a ser seleccionadas (Mason y Gunst, 1985). Además, las fuertes hipótesis bajo las que se obtienen estos t-tests los hacen prácticamente inviables en la práctica.

Un punto intermedio entre basar la elección en el tamaño de la varianza y en el resultado del t-test, es ir eliminando componentes comenzando con la que tenga menor varianza, y así sucesivamente hasta alcanzar el primer t-valor significativo. No obstante, no se trata de una técnica muy eficiente pues tiende a eliminar muy pocas componentes principales.

Se han planteado otros criterios más sofisticados para la elección de las componentes en el modelo. Algunos autores, como Hill et al. (1977), se fundamentan en el uso del error cuadrático medio.

- Si el objetivo primordial es que el estimador $\tilde{\beta}$ "difiera" lo menos posible de β, se plantean dos criterios basados en el error cuadrático medio para que se tenga en cuenta tanto la varianza como el sesgo. Por un lado, se plantea elegir $\tilde{\beta}$ en lugar de $\hat{\beta}$ si el valor esperado de la distancia euclídea con β es menor para $\tilde{\beta}$. Por otro lado, un criterio más fuerte para la elección de $\tilde{\beta}$ es determinar si el error cuadrático medio de $c'\tilde{\beta}$ es menor que para $c'\hat{\beta}$ para cada vector no nulo c.

- Si por el contrario, el objetivo principal es conseguir una buena predicción de Y, entre los criterios propuestos destacan dos basados nuevamente en el error cuadrático medio. El criterio débil estipula elegir $\tilde{\beta}$ en lugar de $\hat{\beta}$ si el valor esperado de la distancia euclídea entre $X\tilde{\beta}$ y $X\beta$ es menor que entre $X\hat{\beta}$ y $X\beta$. Un criterio más fuerte sugiere usar el verdadero valor de Y en lugar de su valor esperado ($X\beta$). De esta forma, se elige $\tilde{\beta}$ en lugar de $\hat{\beta}$ si el valor esperado de la distancia euclídea entre $X\tilde{\beta}$ e Y es menor que para $X\hat{\beta}$ e Y.

Mediante validación cruzada también se puede comparar los valores observados y los ajustados de Y. El criterio propuesto por Mertens et al. (1995) es a través de $\sum_{i=1}^{n}(y_i - \hat{y}_{M(i)})^2$ donde $\hat{y}_{M(i)}$ es el estimador de y_i obtenido a través de regresión sobre componentes principales basado en un subconjunto M y usando la matrix $X_{(i)}$, que es X eliminando su $i-$ésima fila.

Otra consideración que tiene en cuenta la varianza y el sesgo, dada por Lott (1973), es mediante el cálculo del coeficiente múltiple ajustado de determinación, basado en el coeficiente de determinación usual R^2 hallado para cada conjunto M de interés. De esta forma, se elige el subconjunto M que maximice

$$\bar{R^2} = 1 - \frac{(n-1)(1-R^2)}{(n-p-1)}$$

Sin embargo, se ha demostrado que este procedimiento funciona de forma limitada.

Se observa, tras lo expuesto, la complejidad en el intento de proporcionar una regla general para la elección del subconjunto de componentes principales M, pues no se puede basar la decisión únicamente en el tamaño de la varianza, pero tampoco priorizar el valor predictivo de las componentes si éstas tienen varianzas pequeñas.

En la práctica, habitualmente se usan otros criterios como son:

- Eliminar las componentes del modelo asociadas a autovalores muy pequeños hasta que el resto de las CPs expliquen un determinado porcentaje del total de la varianza, generalmente un 80 %. El porcentaje de información que proporcionan las k ($k < p$) primeras componentes principales viene dado por: $\frac{\sum_{i=1}^{k}\lambda_i}{\sum_{i=1}^{p}\lambda_i} \times 100$. Este porcentaje varía en función del campo en que se trabaje, ya que por ejemplo en un estudio de ámbito social el porcentaje se encontraría sobre el 60 %, mientras que en un estudio de ámbito científico el porcentaje es más próximo a un 80 %.
- Regla de Kaiser (1960): eliminar las componentes asociadas con autovalores de valores menor que 1, que es el valor típico para las matrices de correlación. Así se garantiza que las componentes seleccionadas explican más variabilidad que una variable original.

2.5. Eliminación de Variables en Regresión usando Componentes Principales

Como ya se ha mencionado anteriormente, una alternativa para combatir la multicolinealidad es usar únicamente un subconjunto de variables predictoras. Suprimiendo adecuadamente variables del modelo se mejora la precisión de los parámetros estimados, aunque se introduce sesgo en los estimadores de los coeficientes de las variables predictoras y la variable dependiente. Sin embargo, el error cuadrático medio de los estimadores sesgados es menor que la varianza de los estimadores insesgados, es decir, la cantidad de sesgo introducido es menor que la reducción de la varianza.

Existen muchos métodos de selección de variables basados en la búsqueda sobre todos los subconjuntos posibles de variables con objetivo de encontrar el mejor subconjunto de variables predictoras. Estos algoritmos suelen determinar buenos subconjuntos de variables aunque, a veces, son costosos computacionalmente debido a la cantidad de pasos que se deben realizar. Se procede a exponer algunos de estos métodos clásicos mencionados en Navarro (2009):

- Todas las regresiones posibles (TRP)

La opción más natural, pero laboriosa, consiste en crear todos los posibles subconjuntos de variables predictoras y realizar todas las regresiones posibles. Posteriormente, se analiza la ganancia o pérdida de eficiencia de cada modelo con introducción de alguna variable. El problema con esta técnica se encuentra en el gran número de cálculos a realizar; si el número de variables predictoras es p, serían necesarias realizar $2^p - 1$ regresiones.

- Método de selección escalonada

Estas técnicas de selección se basan en añadir o suprimir una única variable predictora en cada paso según ciertos criterios. Este proceso secuencial finaliza cuando se cumple una determinada regla de parada. Se trata de un proceso más sencillo en cálculo y eficiente que el anterior.

Los tres algoritmos más usados son los siguientes:

1. Procedimiento de selección hacia delante: “Forward”. Se parte de un modelo muy sencillo y en cada paso se añade la variable más significativa según cierto criterio hasta una cierta regla de parada. Una vez que una variable ha sido añadida al modelo nunca se elimina.
2. Procedimiento de eliminación hacia atrás : “Backward”. Al contrario que en el anterior, en este método se parte de un modelo muy complejo, el que contiene todos los términos, y en cada etapa se desecha la variable menos significativa, hasta que se considere no suprimir ningún término más. Una vez eliminada, ninguna variable se puede volver a incorporarse al modelo.

3. Método escalonado por pasos: "Stepwise". Se trata de una combinación de las dos técnicas anteriores. Se parte de un modelo muy sencillo como en "forward", en cada paso se introduce una variable y posteriormente se cuestiona si todas las variables introducidas deben permanecer en el modelo.

Para estas técnicas se necesita establecer la condición para añadir o suprimir un término. Se suelen considerar el Criterio de Información de Akaike (Akaike, 1974), el Criterio de Información Bayesiano (Schwarz, 1978) o la significación de cada coeficiente.

Sin embargo, la experiencia indica que estos algoritmos por pasos mencionados anteriormente tienden a no seleccionar buenos subconjuntos de variables predictoras cuando se presentan problemas de multicolinealidad. La Regresión sobre Componentes Principales también puede ser usada de forma iterativa para seleccionar variables. A continuación se muestran algunos procedimientos que utilizan componentes principales para proceder a la selección de variables.

Considérese el problema de Regresión sobre Componentes Principales anterior tal que el estimador propuesto para β es

$$\tilde{\beta} = \sum_{M} \lambda_k^{-1} a_k a_k' X'Y$$

Se pretende estudiar si hay subconjuntos de elementos de $\tilde{\beta}$ que sean significativamente distintos de cero, y aquellas variables cuyos coeficientes no resultan significativos pueden ser eliminados del modelo.

Siempre que los verdaderos coeficientes de las CPs eliminadas sean cero y que los supuestos de normalidad sean válidos, el estadístico apropiado es el F-estadístico, usando t-estadísticos si sólo una variable es considerada a la vez.

Este procedimiento se puede iterar de manera que el siguiente paso sería aplicar regresión sobre componentes principales en el subconjunto reducido de variables y observar si pueden ser eliminadas más variables de dicho subconjunto usando el mismo razonamiento anterior. Esto se repetiría hasta que no se eliminen más variables.

Existen dos variaciones de este procedimiento descritas por Mansfield et al. (1977).

- La primera variación busca una única variable a eliminar. A continuación, el mejor par de variables a eliminar, siendo una de ellas la variable del primer paso. Luego, buscar el mejor trío de variables a eliminar, que incluiría el mejor par anterior... y así sucesivamente.
- La segunda variación sugiere eliminar sólo una variable en cada paso y luego se vuelve a calcular las CPs utilizando el conjunto reducido de variables, en lugar de permitir la eliminación de varias variables antes de que se vuelvan a calcular las CPs.

Boneh y Mendieta (1994) propusieron un procedimiento iterativo diferente que resulta ser computacionalmente más eficiente. Siguiendo con la misma notación hasta el momento, el algoritmo se describe como sigue:

1. Selección de la primera variable

- Se realiza regresión sobre las componentes principales: $Y = Z\gamma + \epsilon$
- Se elige un subconjunto M de CPs que se considere que contribuyen significativamente a la regresión. Dicha significación se basa en el uso de t-tests, aunque dado que se ha mostrado que los t-tests tienen una potencia reducida para las CPs con varianzas pequeñas (Mason y Gunst, 1985), se ha introducido una modificación para tales CPs.
- Si $M = \emptyset$, el proceso de selección termina concluyendo que no se deben incluir variables predictoras en el modelo. En caso contrario, sean $SSE_i,\ i = 1, ..., p$ las sumas de cuadrados del error cuando se hace la regresión de X_i sobre el subconjunto de CPs incluidas en M, se selecciona como primera variable X_i que tiene asociada SSE mínima.

Sin pérdida de generalidad, se asume que la primera variable seleccionada es X_1.

2. Selección de la segunda variable

- Se realiza regresión de X_k sobre X_1, $k = 2, ..., p$.
- Se obtienen las componentes principales $(Z_2', ..., Z_p')$ de $(\epsilon_2, ..., \epsilon_p)$ siendo $\epsilon_i,\ i = 2, ..., p$ el vector de residuos estandarizados de la regresión de X_i sobre X_1.
- Se hace regresión de Y sobre $(Z_2', ..., Z_p')$ y, siguiendo un criterio similar al anterior, se selecciona un subconjunto M' óptimo de $(Z_2', ...Z_p')$
- Sea M' un subconjunto de $Z_2', ..., Z_p'$; si $M' = \emptyset$, el proceso de selección termina con la conclusión de que solo la variable X_1 debe ser incluida en el modelo. En caso contrario, sean $SSE_i',\ i = 1, ..., p$ las sumas de cuadrados del error cuando se hace la regresión de ϵ_i en M', la segunda variable predictora seleccionada es aquella para la cual la SSE_i' sea mínima.

La selección de variables continúa de forma similar hasta obtener un subconjunto de componentes principales M^* que sea vacío.

Además de estos pasos de selección hacia adelante, el procedimiento de Boneh y Mendieta incluye la opción de volver hacia atrás, ya que las variables seleccionadas previamente pueden eliminarse y nunca podrán volver a ser seleccionadas. La eliminación de una variable se produce si su contribución es suficientemente disminuida por la inclusión posterior de otras variables.

Daling y Tamura (1970) propusieron otra forma de seleccionar variables basándose en una modificación de la idea de asociar una variable con cada una de las primeras (últimas)

componentes y luego retener (eliminar) aquellas variables asociadas con estas primeras (últimas) componentes. El método de Daling y Tamura sostiene eliminar primero las últimas CPs, luego rotar las CPs restantes usando varimax y finalmente seleccionando una variable asociada con cada CPs rotada que tenga una correlación significativa con la variable dependiente.

La principal desventaja de este método se encuentra en el primer paso del mismo, ya que la eliminación de CPs con baja varianza puede hacer desaparecer información importante con respecto a la relación entre Y y las variables predictoras.

2.6. Ventajas y Desventajas de la Técnica

Al aplicar el método de mínimos cuadrados para estimar γ en la ecuación $Y = Z\gamma + \epsilon$, el estimador de β calculado a través de $\hat{\beta} = A\hat{\gamma}$ es equivalente al obtenido aplicando directamente mínimos cuadrados a la ecuación $Y = X\beta + \epsilon$. Sin embargo, una de las ventajas de la regresión sobre componentes principales radica en la sencillez del cálculo; como las columnas de Z son ortogonales, calcular $\hat{\gamma}$ de $Y = Z\gamma + \epsilon$ es mucho más directo que hallar $\hat{\beta}$ de $Y = X\beta + \epsilon$.

En caso de incluir todas las componentes principales en el modelo de regresión, éste sería equivalente al obtenido por mínimos cuadrados, presentando multicolinealidad, cuya erradicación era el objetivo primordial de esta técnica.

No obstante, en caso de eliminar componentes principales del modelo de regresión, los estimadores obtenidos serán sesgados, aunque reducen considerablemente las varianzas de los estimadores de los coeficientes de regresión provocados por la multicolinealidad.

Además, si se aplica regresión a las componentes principales en lugar de las variables predictoras, las contribuciones de las variables transformadas en el modelo de regresión pueden ser interpretadas más fácilmente que las contribuciones de las variables originales.

Más aún, la contribución y los coeficientes estimados de una componente principal no se ven afectados según las componentes principales que también sean incluidas en la regresión. Esto es debido a la incorrelación, ya que para las variables originales, tanto las contribuciones como los coeficientes pueden cambiar considerablemente cuando alguna otra variable es modificada en la ecuación. Es decir, si se considera el modelo de regresión Y sobre las CPs $Z_1, Z_2, ..., Z_p$ con coeficientes de regresión estimados $\hat{\gamma} = [\hat{\gamma_1}, ..., \hat{\gamma_p}]'$ y el modelo de regresión Y sobre las CPs $Z_1, Z_2, ..., Z_m$ con $m < p$ y coeficientes de regresión estimados $\hat{\gamma}^* = [\hat{\gamma_1^*}, ..., \hat{\gamma_p^*}]'$ entonces

$$\hat{\gamma_j^*} = \hat{\gamma_j}, \quad j = 1, ..., m$$

Incluso cuando la multicolinealidad no es un problema, aplicar regresión sobre las componentes principales en lugar de usar las variables originales, tiene ventajas en la interpretación y computación.

Estos hechos se acentúan en presencia de multicolinealidad, siendo por tanto de gran ventaja el uso de la Regresión sobre Componentes Principales. En este caso, se obtiene un estimador de β mucho más estable eliminando un subconjunto de las componentes principales con menor varianza.

Capítulo 3

APLICABILIDAD DEL MÉTODO

3.1. Agricultura: Cambio Climático y Cultivos

Los factores climáticos juegan un papel muy importante a la hora de explicar determinados procesos. Estos factores son frecuentemente utilizados como variables independientes para explicar la variabilidad de ciertos procesos sin tener en cuenta la posibilidad de que algunos valores estén muy relacionados entre sí.

Uno de los principales campos de estudio de los factores ambientales es la agricultura. Desde el punto de vista agronómico, las altas temperaturas afectan a los niveles de humedad del suelo, lo que podría disminuir los rendimientos de los cultivos si el suministro de agua para riego no es suficiente. Por otro lado, la precipitación mantiene la humedad del suelo necesaria para el crecimiento del cultivo. Diferentes temperaturas pueden cambiar la duración de la temporada de crecimiento induciendo variaciones en el rendimiento de los cultivos. Por ejemplo, las altas temperaturas tienden a acortar muchas temporadas de cultivo y los cultivos están expuestos a menos radiación solar necesaria para la fotosíntesis. Resumiendo, a largo plazo, el cambio de temperatura y precipitación puede alterar los patrones de cultivo.

En 2011, la Universidad de Georgia publicó un estudio que surgió con el objetivo de comparar los efectos del cambio climático en los rendimientos de los cultivos en diferentes regiones de Estados Unidos (véase Cai et al. (2011)). Se pretendía estimar las relaciones entre el clima y los rendimientos de los cultivos de maíz, soja, algodón y cacahuetes para varios estados del norte (Minnesota, Nebraska, Indiana, Illinois y Iowa) y sur (Georgia, Alabama y Texas) de EE.UU., considerados los principales estados generadores de dichos productos.

El estudio contempla el periodo $1960 - 2009$ y se seleccionó la temperatura mensual y la precipitación durante la temporada de crecimiento como variables climáticas para el modelo de rendimiento. Para predecir el cambio climático futuro con escenarios alternativos de gases de efecto invernadero, los científicos han desarrollado muchos modelos de cambio

climático. Los diferentes modelos proporcionan distintas proyecciones de cambio climático basadas en diferentes enfoques y escenarios posibles. En este estudio, se consideraron tres proyecciones de cambio climático (frío, cálido y medio).

Una temporada de crecimiento típica es de siete u ocho meses, lo que supondría considerar alrededor de 14 a 16 variables de predicción tanto para la temperatura como para la precipitación en el modelo, lo que llevó a plantear la necesidad de reducir variables. A lo largo de la historia, se han utilizado muchas metodologías para esto en el ámbito de la agricultura. Por ejemplo, en lugar de usar los meses del calendario, algunos investigadores dividen la temporada de crecimiento por las etapas del crecimiento de los cultivos. De esta forma, como la temporada de crecimiento activo para el maíz en el estado de Georgia es de abril a octubre, se consideran siete variables de temperatura mensuales. Al dividir la temporada de crecimiento por las cuatro etapas generales de crecimiento para el maíz, se podría reducir el número de variables de temperatura a cuatro.

Desde el punto de vista agrícola, este método tiene más sentido que el uso de meses de calendario. Sin embargo, es difícil especificar el límite exacto entre dos etapas de crecimiento del cultivo, además de que el calendario de crecimiento del cultivo varía de un año a otro y de una región a otra.

Alternativamente, se decidió usar técnicas de selección de variables estadísticas para reducir el número de meses como variables predictoras. Sin embargo, las variables meteorológicas están correlacionadas, luego un enfoque directo de selección de variables conduciría a resultados inestables.

Para reducir el número de variables sin pérdida significativa de información y eliminar un posible problema grave de multicolinealidad, se aplicó la Regresión sobre Componentes Principales. De esta forma, cada componente principal contiene información sobre todas las variables meteorológicas y ningún mes de la temporada de crecimiento se omitirá por completo, independientemente de la técnica de selección de variables que se aplique. En lugar de simplemente tomar las primeras componentes principales como variables predictoras, se usaron técnicas de selección de variables estadísticas para seleccionar un subconjunto apropiado de CPs.

A partir de un único conjunto de datos meteorológicos, se crearon tres modelos de regresión sobre componentes principales diferentes para pronosticar la respuesta del rendimiento del cultivo basados en tres conjuntos de datos de proyección del cambio climático y se generó un Índice de Impacto del Cambio Climático (CCII) que permitiese comparar los efectos del cambio climático en diferentes regiones.

Las estimaciones de cambio climático de estos tres modelos climáticos proporcionan una proyección mensual de la temperatura y la precipitación hasta el año 2100. Como en general, a medida que aumenta el horizonte temporal, la predicción de la respuesta del rendimiento de los cultivos será menos confiable, sólo se usaron los datos del cambio

climático hasta el año 2050.

Dado que el objetivo principal del análisis era el estudio de la predicción del rendimiento, no se hace mención sobre las cargas de las variables meteorológicas en las CP específicas y los coeficientes estimados en el modelo de regresión sobre componentes principales.

Para proporcionar una evidencia más sólida, se replicaron análisis similares sobre los productos a nivel de condado para todos los condados disponibles en los ocho estados esudiados.

Los resultados finales del estudio indicaron que el clima más cálido en el futuro tendrá un impacto negativo en los condados del sur de los EE.UU., mientras que tendrá un impacto insignificante en los condados del norte de los EE.UU. en las próximas cuatro décadas. Es decir, los resultados sugieren que los agricultores del sur podrían hacer frente al cambio climático cambiando a cultivos con mejor tolerancia al calor.

3.2. Dendrología: Datos "Pitprop"

En esta sección se considera el conjunto de datos conocido como "*Pitprop*", dado originalmente por Jeffers (1967) y el cual ha sido analizado por varios autores posteriormente. Se trata de un estudio de la fuerza de compresión de los puntales cortados a partir de madera cultivada, con el objetivo de determinar las características, o combinación de variables, que hacen que tales puntales sean lo suficientemente fuertes para su uso en minas.

Los datos consisten en 180 tajos cortados de madera de pino corso del este de Anglia (Reino Unido) a los cuales se midieron las siguientes variables:

X_1 : $TOPDIAM$ – *diámetro superior*

X_2 : $LENGTH$ – *longitud*

X_3 : $MOIST$ – *humedad*

X_4 : $TESTSG$ – *gravedad específica en el momento de la prueba*

X_5 : $OVENSG$ – *gravedad específica secada al horno*

X_6 : $RINGTOP$ – *número de anillos anuales en la parte superior*

X_7 : $RINGBUT$ – *número de anillos anuales en la parte base*

X_8 : $BOWMAX$ – *arco máximo*

X_9 : $BOWDIST$ – *distancia del punto de arco máximo a lo alto del puntal*

X_{10} : $WHORLS$ – *número de espirales de nudo*

X_{11} : $CLEAR$ – *longitud del puntal claro desde la parte superior del puntal*

X_{12} : $KNOTS$ – *número medio de nudos por espiral*

X_{13} : $DIAKNOT$ $-$ *diámetro medio de los nudos*

El objetivo es, a partir de los valores de estas variables, construir una ecuación de predicción para la variable

$$Y : STRENGTH \; - \; fuerza \; de \; compresión$$

En la siguiente tabla se muestran los coeficientes de correlación entre cada una de las 13 variables y entre cada una de las variables y la respuesta. De esta tabla resulta evidente la presencia de multicolinealidad entre las variables. En particular, muchas variables están correlacionadas con la longitud del puntal y con el número de anillos anuales en la base del puntal. Por ello, es claro que el uso de las 13 variables en un modelo de regresión es inadecuado.

TOPDIAM													
0,954	LENGTH												
0,364	0,297	MOIST											
0,342	0,284	0,882	TESTSG										
-0,129	-0,118	-0,148	0,22	OVENSG									
0,313	0,291	0,153	0,381	0,364	RINGTOP								
0,496	0,503	-0,029	0,174	0,296	0,813	RINGBUT							
0,424	0,419	-0,054	-0,059	0,004	0,09	0,372	BOWMAX						
0,592	0,648	0,125	0,137	-0,039	0,211	0,465	0,482	BOWDIST					
0,545	0,569	-0,081	-0,014	0,037	0,274	0,679	0,557	0,526	WHORLS				
0,084	0,076	0,162	0,097	0,091	-0,036	-0,113	0,061	0,085	-0,319	CLEAR			
-0,019	-0,036	0,22	0,169	-0,145	0,024	-0,232	-0,357	-0,127	-0,368	0,029	KNOTS		
0,134	0,144	0,126	0,015	-0,208	-0,329	-0,424	-0,202	-0,076	-0,291	0,007	0,184	DIAKNOT	
-0,419	-0,338	-0,728	-0,543	0,247	0,117	0,110	-0,253	-0,235	-0,101	-0,055	-0,117	-0,153	STRENTGH

Figura 3.1: Coeficientes de correlación. Fuente: Jolliffe (2002)

En la siguiente tabla se muestran los coeficientes de las variables para cada una de las CPs, las varianzas de cada componente, el porcentaje de varianza total explicada por cada componente, los coeficientes γ_k en la regresión de Y sobre las CPs y los valores de t-estadísticos que miden la importancia de cada CP en la regresión.

		COMPONENTES PRINCIPALES												
		1	2	3	4	5	6	7	8	9	10	11	12	13
COEFICIENTES	X1	-0,40	0,22	-0,21	-0,09	-0,08	0,12	-0,11	0,14	0,33	-0,31	0,00	0,39	-0,57
	X2	-0,41	0,19	-0,24	-0,10	-0,11	0,16	-0,08	0,02	0,32	-0,27	-0,05	-0,41	0,58
	X3	-0,12	0,54	0,14	0,08	0,35	-0,28	-0,02	0,00	-0,08	0,06	0,12	0,53	0,41
	X4	-0,17	0,46	0,35	0,05	0,36	-0,05	0,08	-0,02	-0,01	0,10	-0,02	-0,59	-0,38
	X5	-0,06	-0,17	0,48	0,05	0,18	0,63	0,42	-0,01	0,28	0,00	0,01	0,20	0,12
	X6	-0,28	-0,01	0,48	-0,06	-0,32	0,05	-0,30	0,15	-0,41	-0,10	-0,54	0,08	0,06
	X7	-0,40	-0,19	0,25	-0,07	-0,22	0,00	-0,23	0,01	-0,13	0,19	0,76	-0,04	0,00
	X8	-0,29	-0,19	-0,24	0,29	0,19	-0,06	0,40	0,64	-0,35	-0,08	0,03	-0,05	0,02
	X9	-0,36	0,02	-0,21	0,10	-0,10	0,03	0,40	-0,70	-0,38	-0,06	-0,05	0,05	-0,06
	X10	-0,38	-0,25	-0,12	-0,21	0,16	-0,17	0,00	-0,01	0,27	0,71	-0,32	0,06	0,00
	X11	0,01	0,21	-0,07	0,80	0,34	0,18	-0,14	0,01	0,15	0,34	-0,05	0,00	-0,01
	X12	0,12	0,34	0,09	-0,30	-0,60	-0,17	0,54	0,21	0,08	0,19	0,05	0,00	0,00
	X13	0,11	0,31	-0,33	-0,30	0,08	0,63	-0,16	0,11	-0,38	0,33	0,04	0,01	-0,01
	VARIANZA	4,22	2,38	1,88	1,11	0,91	0,82	0,58	0,44	0,35	0,19	0,05	0,04	0,04
	% DE VARIANZA TOTAL	32,50	18,30	14,40	8,50	7,00	6,30	4,40	3,40	2,70	1,50	0,40	0,30	0,30
	COEFS. REGRESIÓN	0,13	-0,37	0,13	-0,05	-0,39	0,27	-0,24	-0,17	0,03	0,00	-0,12	-1,05	0,00
	VALOR T-ESTADÍSTICO	68,60	14,39	4,38	1,26	9,23	6,19	4,50	2,80	0,46	0,00	0,64	5,26	0,01

Figura 3.2: Tabla resumen. Fuente: Jolliffe (2002)

Teniendo en cuenta únicamente el tamaño de la varianza, parece que las últimas tres o cuatro componentes deben eliminarse de la regresión. Sin embargo, estudiando los valores de γ_k y los t-estadísticos correspondientes, se observa que la duodécima componente es relativamente importante como variable predictora, a pesar de que representa solo el 0.3 % de la variación total en las variables predictoras. Jeffers (1967) solo retuvo la primera, segunda, tercera, quinta y sexta CP en su ecuación de regresión, mientras que Mardia et al. (1979) sugieren que también deberían incluirse las séptima, octava y duodécima CP.

Se pueden encontrar más detalles de este estudio en Boneh y Mendieta (1992).

3.3. Sociología: Desperdicio de Alimentos en el Hogar

Aproximadamente un tercio de los alimentos del mundo se desperdician anualmente. Aunque está presente en toda la cadena de suministro posterior a la cosecha, gran parte de la responsabilidad es atribuible a los consumidores, por lo que resulta fundamental conocer los comportamientos sociales relacionados con la conciencia, las actitudes y las opiniones que tienen los individuos para explicar el alto nivel de desperdicio de alimentos en el hogar, permitiendo a las instituciones tomar medidas en función de los patrones de conducta de la población.

Con tal objetivo, la Universidad de Ohio realizó en 2015 un estudio sobre el desperdicio de alimentos utilizando las respuestas de una encuesta realizada a nivel nacional de residentes en EE.UU. (véase Qi y Roe (2016)).

Para comprender los patrones comunes entre las respuestas obtenidas se extrajeron las componentes principales y aquellas con autovalores menores que uno se eliminaron del análisis. Posteriormente, se realizó un análisis de regresión sobre las componentes principales para encontrar las asociaciones entre las actitudes de los encuestados y las características personales y familiares.

De dicho análisis destaca una clara asociación entre el desperdicio de alimentos y el ingreso familiar en el hogar, de forma que en hogares con ingresos más altos se tiende a tirar más alimentos priorizando beneficios personales como son la garantía de la calidad y frescor del alimento y la reducción de las enfermedades transmitidas por los mismos. También se concluye un mayor desperdicio de aquellos alimentos que se compran a granel.

Para conseguir reducir el desperdicio de alimentos, la información proporcionada por este estudio resulta relevante y muy útil a la hora de orientar políticas de actuación por parte de las instituciones.

Capítulo 4

IMPLEMENTACIÓN EN R

4.1. Introducción

Para la implementación de la metodología de Regresión sobre Componentes Principales en R existen varias opciones. En este trabajo se presentan ciertos paquetes estadísticos incluidos en el software que ejecutan un análisis completo de una forma directa y rápida, y posteriormente se realizará, de forma detallada, el desarrollo de la técnica a través de las funciones básicas del programa.

Uno de los paquetes a destacar es el denominado *PCovR* (Vervloet et al., 2015), que permite analizar un conjunto de datos de regresión en presencia de multicolinealidad, reduciendo a su vez las variables predictoras a un número limitado de componentes.

Este paquete contiene la función *pcovr*(). Ésta ejecuta un análisis completo de regresión sobre componentes principales de un conjunto de datos y proporciona varias opciones de preprocesamiento, selección de modelo y rotación.

Se usa la orden:

pcovr(X, Y, modsel="*seq*", Rmin=1, Rmax=$ncol(X)/3$, R=$NULL$, weight=$NULL$, rot="*varimax*", target=$NULL$, prepX="*stand*", prepY="*stand*", ratio="*estimation*", fold="*LeaveOneOut*", zeroloads=$ncol(X)$)

con argumentos de entrada:

	ARGUMENTOS DE ENTRADA DE LA ORDEN "PCOVR"
X	Dataframe que contiene las variables predictoras.
Y	Dataframe que contiene las variables dependientes.
modsel	Procedimiento de selección del modelo (*seq, seqRcv, seqAcv* o *sim*).
Rmin	Número mínimo de componentes consideradas.
Rmax	Número máximo de componentes consideradas.
R	Número de componentes (anula Rmin y Rmax).
weight	Ponderación considerada.
rot	Criterio de rotación (*varimax, quartimin, targetT, targetQ, wvarim, promin* o *none*).
target	Matriz para la rotación (componentes x variables predictoras).
prepX	Preprocesamiento de las variables predictoras: estandarizar (*stand*) o centrar (*cent*).
prepY	Preprocesamiento de las variables dependientes: estandarizar (*stand*) o centrar (*cent*).
ratio	Relación de las varianzas de error estimadas del bloque predictor y del bloque de dependientes.
fold	Valor de *k* al realizar *k*-validación cruzada. Por defecto, se realiza validación cruzada dejando uno fuera.
zeroloads	Número de cargas cercanas al cero de la rotación simplimax.
x	Objeto de tipo producido por pcovr.
cpal	Vector de colores usado para los gráficos de selección de modelos.
lpal	Vector de líneas usado para los gráficos de selección de modelos.

Figura 4.1: Argumentos de entrada

y argumentos de salida:

	ARGUMENTOS DE SALIDA DE LA ORDEN "PCOVR"
Px	Matriz de cargas (componentes x variables predictoras).
Py	Matriz de pesos de regresión (componentes x variables dependientes).
Te	Matriz de puntuaciones de las componentes (observaciones x componentes).
W	Matriz de ponderaciones de las componentes (variables predictoras x componentes)
Rx2	Proporción de varianza explicada en X.
Ry2	Proporción de varianza explicada en Y.
Qy2	Ajuste de validación cruzada (parámetro de ponderación x número de componentes)
VAFsum	Suma ponderada de la varianza contabilizada en X y en Y (1 x número de componentes).
alpha	Valor del parámetro de ponderación seleccionado .
R	Número de componentes seleccionadas.
modsel	Modelo de procedimiento de selección usado.
rot	Criterio de rotación usado.
prepX	Método de preprocesamiento usado para las variables predictoras.
prepY	Método de preprocesamiento usado para las variables dependientes.
Rvalues	Número de componentes que fueron consideradas.
Alphavalues	Valores del parámetro de ponderación que fueron considerados.

Figura 4.2: Argumentos de salida

Además de la librería *PCovR*, la cual es destinada exclusivamente a la técnica desarrollada en este trabajo, se encuentran disponibles otros paquetes estadísticos en R que contienen, entre otras, funciones que permiten realizar regresión por componentes principales. El paquete *pls* (Mevik y Wehrens, 2007), destinado a métodos de regresión multivariante, dispone de algunas funciones útiles como $pcr()$, $mvr()$ o $svdpc.fit()$. El paquete *spcr* (Kawano, 2016) incluye funciones para el desarrollo de la técnica de regresión por componentes principales "dispersa", la cual puede considerarse una modificación de la regresión por componentes principales cuya finalidad es conseguir que gran parte de los coeficientes de la matriz de cargas sean nulos.

4.2. Ilustración

Pese a la existencia de los paquetes estadísticos de R mencionados anteriormente, en esta memoria se procede a realizar el desarrollo de la técnica a través de las funciones básicas del sistema.

Para la implementación manual de la técnica de Regresión sobre Componentes Principales en R se usa un conjunto de datos denominado "*Boston*" incluido en la librería *MASS* (Venables y Ripley, 2002) de R (R Core Team, 2018).

Esta base de datos contiene diversas medidas de calidad de 506 distritos municipales de Estados Unidos con el objetivo de predecir el valor medio de las viviendas. Las variables consideradas en el estudio son:

$X_1 : crim$ – *tasa de criminalidad per cápita*

$X_2 : zresid$ – *porcentaje de zonas residenciales*

$X_3 : indus$ – *porcentaje de zonas industriales*

$X_4 : jr$ – 1 *si el distrito es limítrofe con el río,* 0 *en caso contrario*

$X_5 : nox$ – *concentración de óxido de nitrógeno*

$X_6 : nmhv$ – *número medio de habitantes por vivienda*

$X_7 : antig$ – *porcentaje de viviendas anteriores a* 1940

$X_8 : dis$ – *distancia media a los* 5 *centros de trabajo principales*

$X_9 : circunv$ – *índice de accesibilidad a vías de circunvalación*

$X_{10} : tax$ – *indicador de impuestos de la administración*

$X_{11} : alprof$ – *proporción alumno* $-$ *profesor*

$X_{12} : ipn$ – *índice de población negra*

$X_{13} : porcb$ – *porcentaje de población de clase baja*

La variable objetivo es

$$Y : medv \; - \; valor\ medio\ de\ las\ viviendas$$

Dado que la información sobre el precio de las viviendas es de gran importancia en el mercado y la economía, se pretende encontrar una regla de estimación de la misma a través de ciertas medidas que caracterizan a los distritos.

A continuación, se muestran los 5 primeros casos del conjunto.

```
   crim zresid indus jr   nox  nmhv antig    dis circunv tax alprof
1 0.00632     18  2.31  0 0.538 6.575  65.2 4.0900       1 296   15.3
2 0.02731      0  7.07  0 0.469 6.421  78.9 4.9671       2 242   17.8
```

```
3 0.02729      0  7.07  0 0.469 7.185  61.1 4.9671       2 242   17.8
4 0.03237      0  2.18  0 0.458 6.998  45.8 6.0622       3 222   18.7
5 0.06905      0  2.18  0 0.458 7.147  54.2 6.0622       3 222   18.7
     ipn porcb medv
1 396.90  4.98 24.0
2 396.90  9.14 21.6
3 392.83  4.03 34.7
4 394.63  2.94 33.4
5 396.90  5.33 36.2
```

Para evitar confusiones y facilitar la interpretación, en primer lugar se estandarizan las variables.

```
datos = as.data.frame(scale(datos))
```

Para poder evaluar la capacidad predictiva del modelo, se dividen las observaciones disponibles en dos grupos: uno de entrenamiento para ajustar el modelo (70 % de los datos) y uno de test (30 % de los datos).

```
training<-datos[1:354,]
test<-datos[355:506,]
```

El Análisis de Regresión Clásico se realiza utilizando la función $lm()$.

```
RegClas<- lm(formula = medv ~ .+0,
             data = training)
summary(RegClas)
```

```
Call:
lm(formula = medv ~ . + 0, data = training)

Residuals:
    Min      1Q  Median      3Q     Max
-0.8193 -0.1882 -0.0383  0.1914  1.3701

Coefficients:
        Estimate Std. Error t value Pr(>|t|)
crim     0.21453    0.17568   1.221  0.22289
zresid   0.03089    0.02399   1.287  0.19880
indus    0.03550    0.03177   1.117  0.26462
jr       0.02480    0.01769   1.402  0.16178
nox     -0.05971    0.04029  -1.482  0.13926
nmhv     0.70746    0.02939  24.068  < 2e-16 ***
antig   -0.13572    0.02968  -4.572 6.76e-06 ***
```

```
dis      -0.19682    0.03249  -6.057 3.66e-09 ***
circunv  0.04531    0.10013   0.453  0.65115
tax      -0.26538    0.05329  -4.980 1.01e-06 ***
alprof   -0.15975    0.02215  -7.213 3.57e-12 ***
ipn       0.12605    0.04642   2.716  0.00695 **
porcb    -0.07316    0.03768  -1.941  0.05304 .
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.3357 on 341 degrees of freedom
Multiple R-squared:  0.8813,    Adjusted R-squared:  0.8767
F-statistic: 194.7 on 13 and 341 DF,  p-value: < 2.2e-16
```

Al estudiar el resumen del modelo, destaca el valor de R cuadrado 0.8767, indicando la obtención de un buen modelo. Sin embargo, analizando los coeficientes de regresión individuales se aprecia que muchas de las variables no son estadísticamente significativas a un nivel de significación del 0.05.

La ecuación de regresión se obtiene a través de:

$$Y = 0.21453\ crim + 0.03089\ zresid + 0.03550 indus + 0.02480\ jr$$

$$-0.05971\ nox + 0.70746\ nmhv - 0.13572\ antig - 0.19682\ dis$$

$$+0.04531\ circunv - 0.26538\ tax - 0.15975\ alprof + 0.12605\ ipn - 0.07316\ porcb$$

Los intervalos de confianza de los coeficientes de regresión a un nivel del 95 % se muestran a continuación:

```
(ICClas<-confint(RegClas,level=0.95))
```

```
                2.5 %        97.5 %
crim     -0.131033202  0.5600922416
zresid   -0.016300018  0.0780721589
indus    -0.026988003  0.0979786878
jr       -0.009988527  0.0595850425
nox      -0.138963759  0.0195390415
nmhv      0.649641653  0.7652728838
antig    -0.194111148 -0.0773387575
dis      -0.260737133 -0.1329063371
circunv  -0.151630198  0.2422579908
tax      -0.370199659 -0.1605659906
alprof   -0.203309319 -0.1161816309
ipn       0.034750786  0.2173485770
porcb    -0.147283357  0.0009643296
```

Todas las suposiciones para regresión simple (linealidad, homocedasticidad, independencia y normalidad) también se aplican para regresión múltiple con otra condición más; la ausencia de multicolinealidad en los datos.

Se asumen las hipótesis básicas con objeto de centrar este trabajo en la detección y corrección de cierta dependencia entre algunas variables predictoras. Se procede, por tanto, a analizar la posible existencia de multicolinealidad:

- Gráficos de dispersión matricial para cada par de variables:

```
par(mfrow = c(1,3))
plot(training$nmhv, training$medv, xlab="Nº medio de hab. por vivienda",
     ylab="Valor medio de la vivienda", col="dark blue")
abline(lm(training$medv ~ training$nmhv))
plot(training$alprof, training$porcb, xlab="Profesores/alumno",
     ylab="% de pob. de clase baja", col="dark blue")
abline(lm(training$porcb ~ training$alprof))
plot(training$nmhv, training$porcb, xlab="Nº medio de hab. por vivienda",
     ylab="% de pob. de clase baja", col="dark blue")
abline(lm(training$porcb ~ training$nmhv))
```

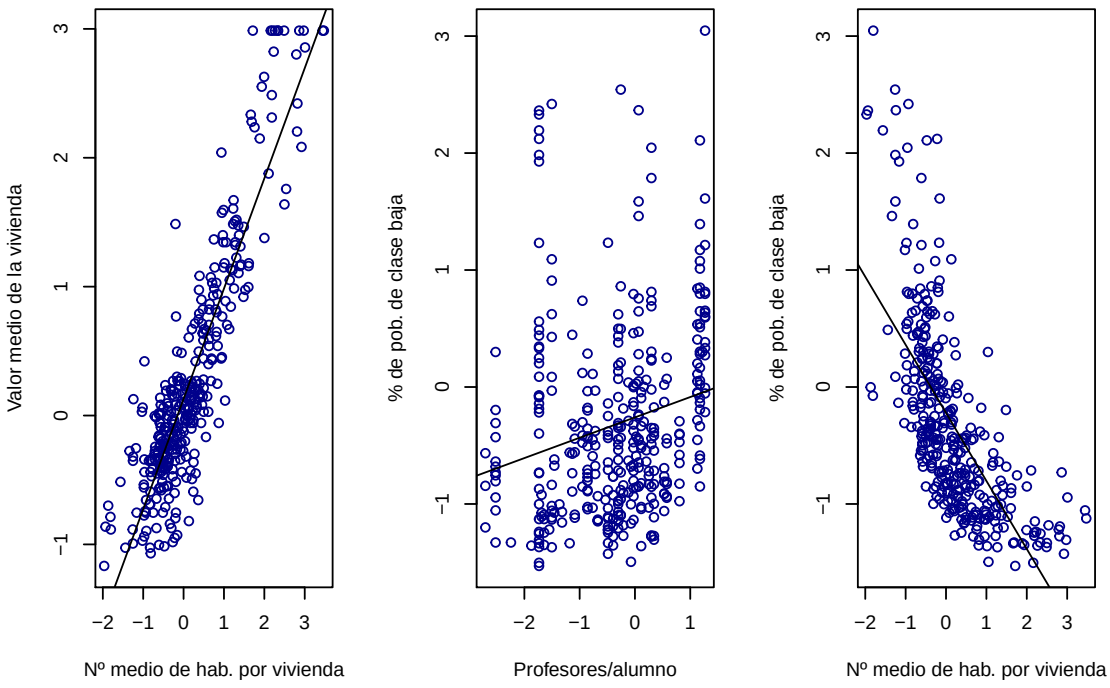

Figura 4.3: Gráficos de dispersión matricial

De estos gráficos se puede extraer un primer indicio de presencia de multicolinealidad en los datos.

- Matriz de correlación R:

```
R1=cor(training)
library(corrplot)
corrplot(R1)
```

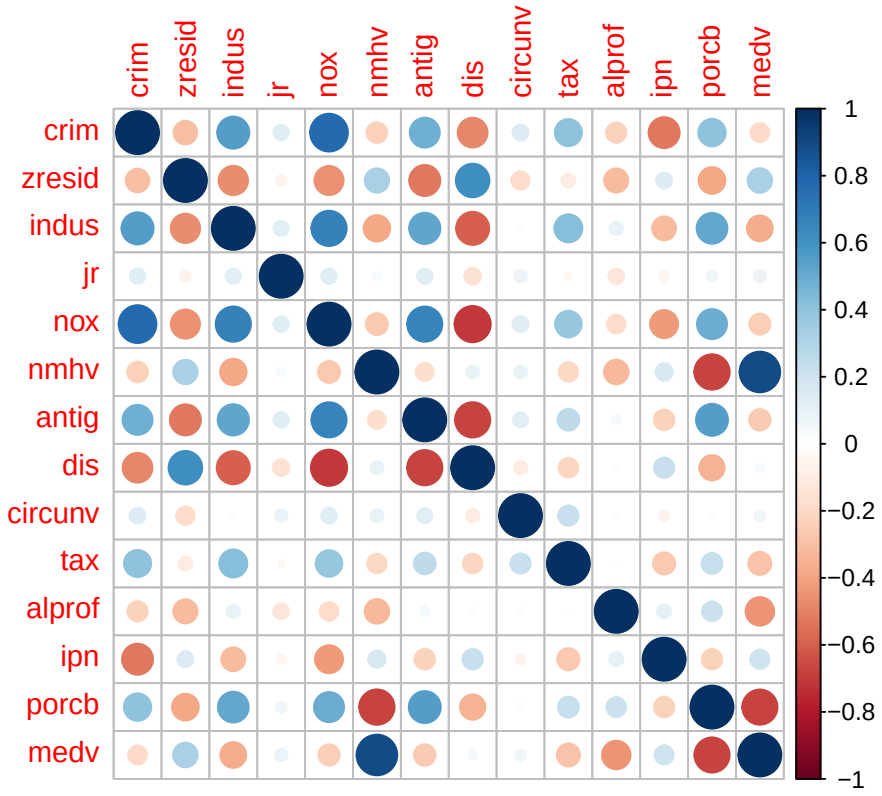

Figura 4.4: Matriz de correlación

En efecto, en la matriz R se observan valores próximos a ± 1 fuera de la diagonal. Más aún, el determinante de la matriz de correlación de las variables predictoras es próximo a cero:

```
R2=cor(training[,1:13])
det(R2)
```

```
[1] 0.001806082
```

- Cálculo del factor de inflación de la varianza:

```
library(car)
round(vif(RegClas),3)
```

```
   crim  zresid   indus      jr     nox    nmhv   antig     dis circunv
 14.076   2.308   3.005   1.074   4.720   2.571   3.038   3.505  11.780
    tax  alprof     ipn   porcb
  4.445   1.755   1.775   3.471
```

Estudiando los elementos de la diagonal de R^{-1}, o equivalentemente, los FIV_i, se aprecian valores considerablemente altos (> 10) para 2 de las 13 variables predictoras, lo que indica alta multicolinealidad causada por dichas variables.

- Índice de Condición de la matriz R:

```
autov= eigen(R2)$`values`
round(sqrt(autov[1]/autov[13]),3)
```

```
[1] 5.812
```

De nuevo se presentan indicios de multicolinealidad al obtener $cond(R) > 5$.

4.2.1. Análisis de Regresión sobre Componentes Principales

Para tratar de corregir el problema de la multicolinealidad en este conjunto de datos, se procede al análisis de Regresión sobre Componentes Principales de los mismos.

Primero, se hallan las componentes principales y posteriormente se decide el número de ellas a mantener en el modelo:

```
cps=princomp( ~ crim + zresid + indus + jr + nox + nmhv + antig + dis +
                circunv + tax + alprof + ipn + porcb,
              data = training,cor=TRUE)
summary(cps)
```

```
Importance of components:
                          Comp.1    Comp.2     Comp.3     Comp.4
Standard deviation     2.2001514 1.3267248 1.12700070 1.04349281
Proportion of Variance 0.3723589 0.1353999 0.09770235 0.08375979
Cumulative Proportion  0.3723589 0.5077588 0.60546118 0.68922097
                           Comp.5     Comp.6     Comp.7     Comp.8
Standard deviation     0.96384547 0.85630483 0.81177267 0.69158815
Proportion of Variance 0.07146139 0.05640446 0.05069037 0.03679186
Cumulative Proportion  0.76068236 0.81708682 0.86777719 0.90456905
                          Comp.9    Comp.10    Comp.11    Comp.12
Standard deviation     0.5982098 0.56327294 0.48704562 0.43005766
Proportion of Variance 0.0275273 0.02440588 0.01824719 0.01422689
Cumulative Proportion  0.9320963 0.95650223 0.97474941 0.98897631
                          Comp.13
Standard deviation     0.37856046
Proportion of Variance 0.01102369
Cumulative Proportion  1.00000000
```

Las representaciones gráficas son de gran ayuda a la hora de proporcionar una idea

sobre el comportamiento de las CPs. A continuación se muestran unos gráficos con esta finalidad.

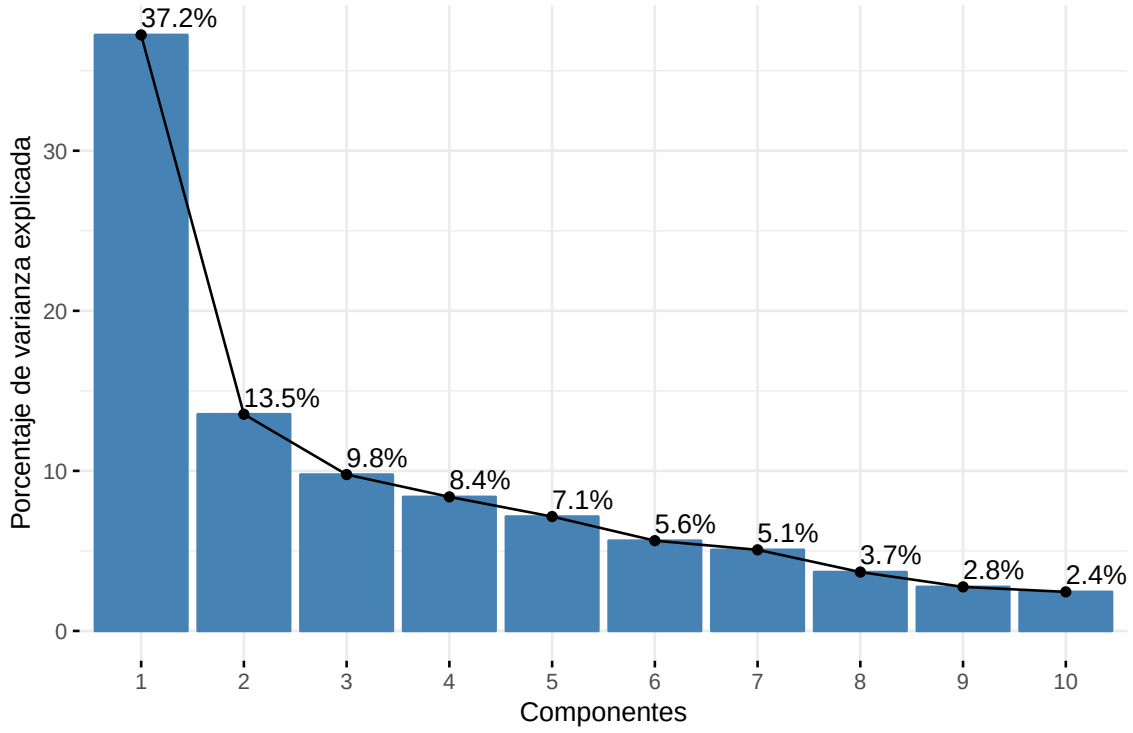

Figura 4.5: Proporción de varianza explicada por componentes

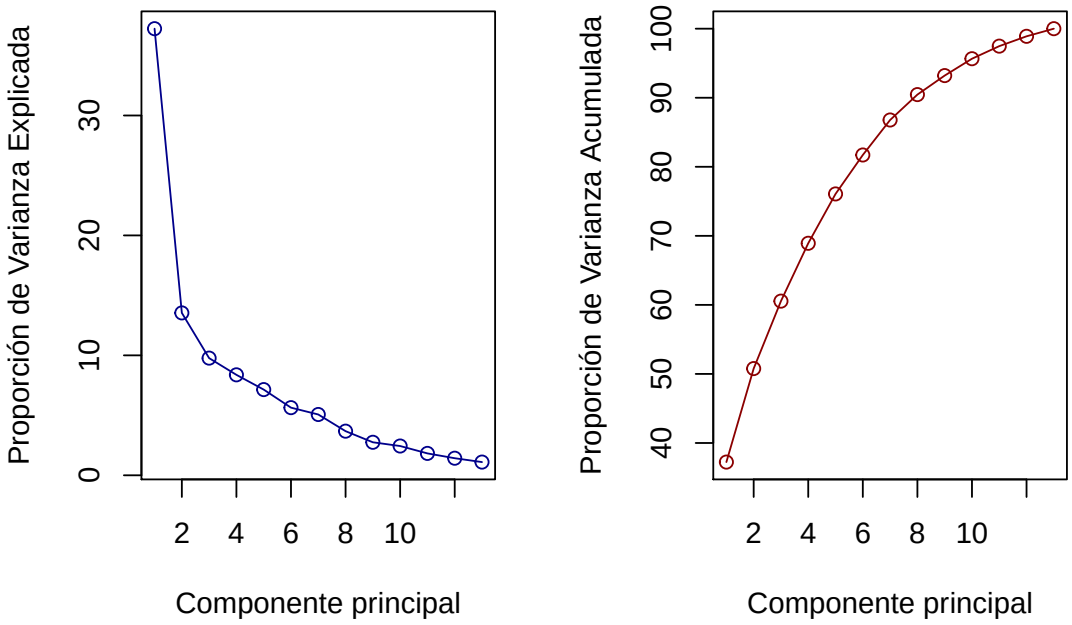

Figura 4.6: Proporción de varianza explicada y acumulada por componentes

El estudio de la proporción de varianza explicada muestra que la primera componente recoge la mayor parte de la información (37.23 %), decayendo drásticamente la varianza en las sucesivas componentes.

Para tomar una decisión sobre las componentes a retener basándose en el porcentaje de información explicada, se plantea emplear las seis primeras componentes pues de esta forma se consigue que, en conjunto, expliquen al menos el 80 % de varianza. Acorde a la Regla de Kaiser, se aconseja mantener las cuatro primeras componentes pues son las asociadas a autovalores de valores mayor que 1, como se aprecia en el siguiente gráfico:

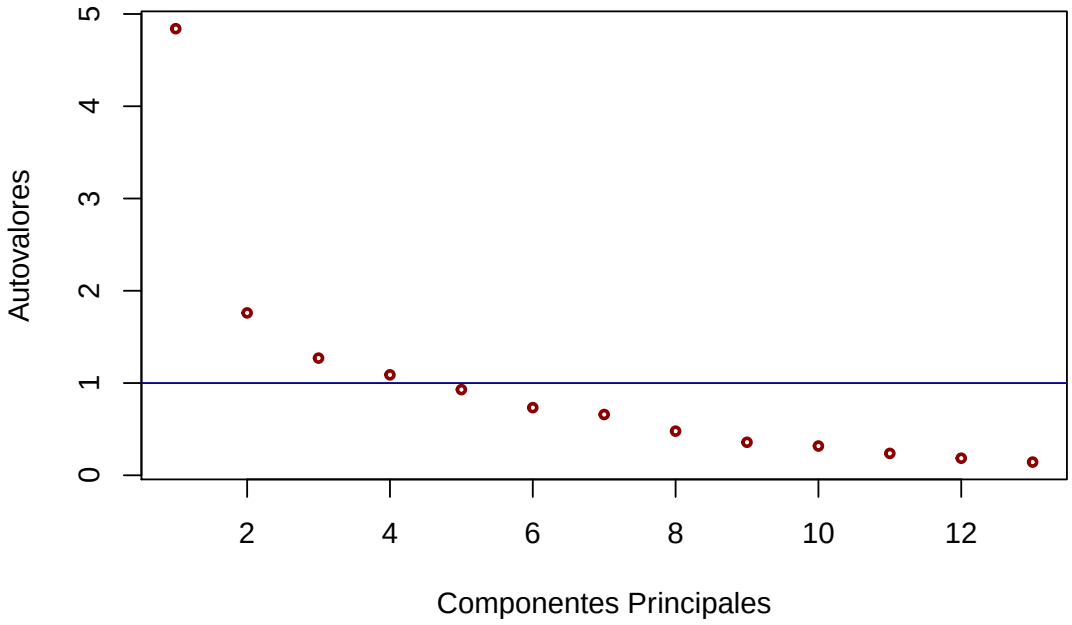

Figura 4.7: Autovalores asociados a las CPs

En la siguiente representación se visualiza la contribución de las variables en una determinada componente, es decir, la importancia de las variables para las componentes principales. La línea discontínua roja muestra el valor medio de contribución, de forma que una variable con una contribución que sobrepase esta línea puede considerarse importante a la hora de contribuir en la componente en cuestión.

En este gráfico se muestran las 10 variables que más contribuyen en la primera CP, y en la tabla la carga de cada una de las variables en dicha CP.

```
fviz_contrib(cps, choice = "var", axes = 1, top = 10)
```

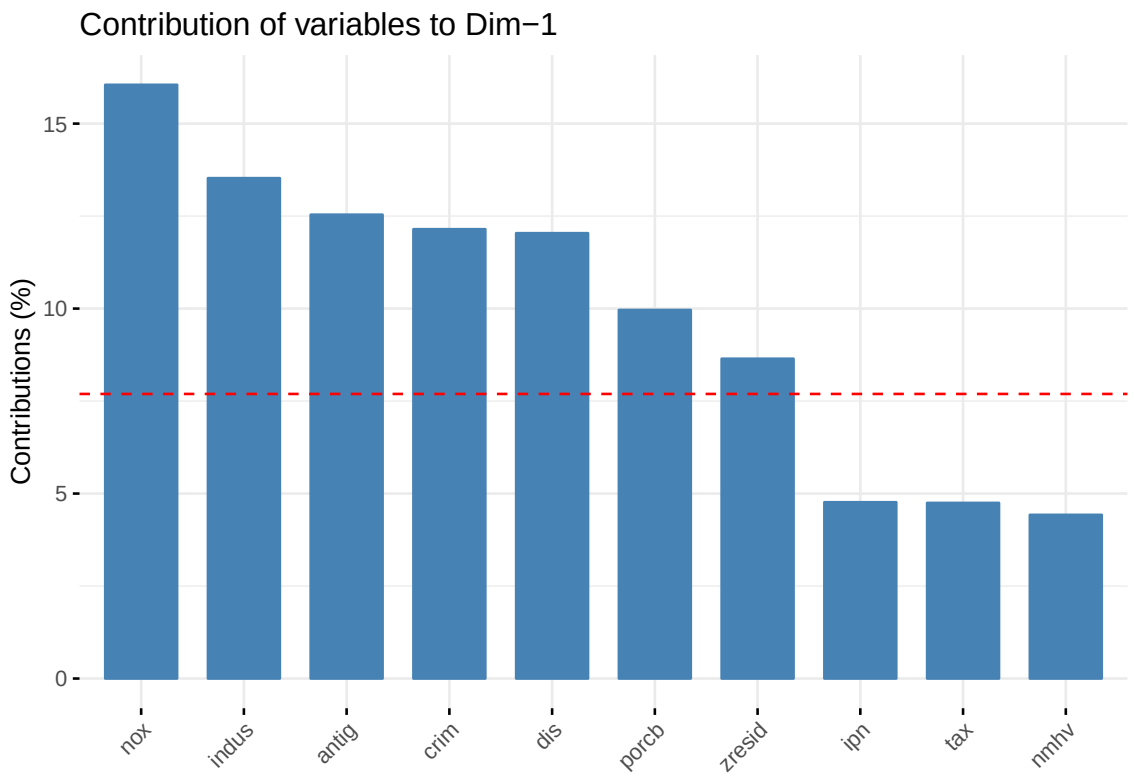

Figura 4.8: Contribuciones en la primera componente principal

```
round(cps$loadings[,1:6],3)
```

```
         Comp.1 Comp.2 Comp.3 Comp.4 Comp.5 Comp.6
crim      0.349 -0.259  0.221  0.028  0.011  0.105
zresid   -0.294 -0.228  0.417  0.101 -0.096 -0.259
indus     0.368  0.053  0.044  0.047  0.020 -0.218
jr        0.075 -0.208 -0.271  0.321 -0.830 -0.047
nox       0.401 -0.193  0.015  0.076  0.118 -0.044
nmhv     -0.210 -0.477 -0.336 -0.047  0.265 -0.002
antig     0.354 -0.010 -0.257  0.069  0.131 -0.106
dis      -0.347  0.099  0.367 -0.108 -0.249  0.038
circunv   0.074 -0.202 -0.252 -0.739 -0.310  0.143
tax       0.218 -0.101  0.343 -0.481 -0.055 -0.498
alprof    0.022  0.600 -0.152 -0.247 -0.051  0.193
ipn      -0.218  0.206 -0.396  0.013  0.030 -0.737
porcb     0.316  0.326  0.164  0.140 -0.187 -0.092
```

Tanto de la representación gráfica como de la tabla, se observa que la variable con más peso en la primera componente es *nox*. De esta forma:

$$Z1 \;=\; 0.349\; crim \;-0.294\; zresid \;+\; 0.368\; indus \;+\; 0.075\; jr \;+\; 0.401\; nox$$

$$- 0.210\ nmhv\ +\ 0.354\ antig\ -\ 0.347\ dis\ +\ 0.074\ circunv\ +\ 0.218\ tax$$
$$+ 0.022\ alprof\ -\ 0.218\ ipn\ -\ 0.316\ porcb$$

El peso asignado en la quinta componente a la variable jr (-0.830) es considerablemente superior al asignado al resto de variables. Además, la variable *circunv* es la segunda con mayor peso en dicha componente (-0.310). Todo esto implica que la quinta componente recoge principalmente la información correspondiente a la ubicación de la vivienda. Sin embargo, la interpretación de las otras componentes no es tan clara.

Finalmente, se procede a realizar Regresión sobre Componentes Principales, tomando únicamente las seis primeras CPs, que acumulan el 81.80 % de la variabilidad:

```
z1<- cps$scores[,1]
z2<- cps$scores[,2]
z3<- cps$scores[,3]
z4<- cps$scores[,4]
z5<- cps$scores[,5]
z6<- cps$scores[,6]

RegCP=lm(medv~z1+z2+z3+z4+z5+z6+0,data=training)
summary(RegCP)
```

```
Call:
lm(formula = medv ~ z1 + z2 + z3 + z4 + z5 + z6 + 0, data = training)

Residuals:
     Min       1Q   Median       3Q      Max
-0.70958 -0.00093  0.25505  0.46807  1.84222

Coefficients:
   Estimate Std. Error t value Pr(>|t|)
z1 -0.19030    0.01183 -16.093  < 2e-16 ***
z2 -0.44842    0.01961 -22.866  < 2e-16 ***
z3 -0.29431    0.02309 -12.749  < 2e-16 ***
z4  0.05207    0.02493   2.088   0.0375 *
z5  0.19836    0.02699   7.348 1.44e-12 ***
z6 -0.01465    0.03038  -0.482   0.6299
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.4895 on 348 degrees of freedom
Multiple R-squared:  0.7424,    Adjusted R-squared:  0.738
```

```
F-statistic: 167.2 on 6 and 348 DF,  p-value: < 2.2e-16
```

La ecuación de regresión se obtiene a través de:

$$Y = -0.19030\ Z1 + 0.44842\ Z2 - 0.29431\ Z3$$

$$+0.05207\ Z4 + 0.19836\ Z5 - 0.01465\ Z6$$

En este modelo las variables predictoras no presentan multicolinealidad ya que los FIVs son todos iguales a 1:

```
vif(RegCP)
```

```
z1 z2 z3 z4 z5 z6
 1  1  1  1  1  1
```

Realizando la transformación para obtener los coeficientes de regresión originales, se obtiene finalmente la ecuación de regresión:

```
round((coefsCP<-RegCP$coefficients[1:6] %*% t(cps$loadings[,1:6])),3)
```

```
       crim zresid  indus    jr   nox  nmhv antig    dis circunv    tax
[1,] -0.013  0.025 -0.097 0.011 0.034 0.403 0.044 -0.142   0.049 -0.126
     alprof   ipn  porcb
[1,] -0.255 0.083 -0.283
```

$$Y = -0.013\ crim + 0.025\ zresid - 0.096\ indus + 0.011\ jr + 0.034\ nox$$

$$+0.403\ nmhv + 0.043\ antig - 0.141\ dis + 0.048\ circunv - 0.125\ tax$$

$$-0.254\ alprof + 0.083\ ipn - 0.283\ porcb$$

De esta forma, los distritos cercanos al río (jr) aumentan el valor de sus viviendas en 0.011 en comparación con los que no lo están, y el aumento en una unidad de la tasa de criminalidad de la zona ($crim$) supone una disminución de 0.013 del valor medio de las viviendas.

La obtención de intervalos de confianza fiables en la regresión sobre componentes principales resulta compleja debido a la sesgadez del estimador $\tilde{\beta}$. Una alternativa es el uso de la técnica bootstrap (véanse Babamoradi y Rinnan (2013), Zumel (2016)). Otra opción es hallar intervalos de confianza asintóticos como sigue: sea $\tilde{\beta}_{CP}$ la estimación de los coeficientes de la regresión sobre componentes principales transformados a las variables originales, se tiene que $v(\tilde{\beta}_{CP}) = v(A\tilde{\gamma}) = Av(\tilde{\gamma})A' = \tilde{V}_{CP}$. De esta forma $(\tilde{\beta}_{CP}) \sim N(\beta, \tilde{V}_{CP})$ asintóticamente y el intervalo de confianza será:

$$IC(\beta, 1-\alpha) = \tilde{\beta}_{CP} \pm SE(\tilde{\beta}_{CP})\ Z_{1-\frac{\alpha}{2}}$$

con $SE()$ el error de estimación, es decir, $(\tilde{V}_{CP})^{\frac{1}{2}}$.

Se procede a la comparación de las características de este modelo con el obtenido a través de un análisis clásico de regresión. Para ello, se estima el error de predicción mediante el Error Cuadrático Medio:

$$ECM = \frac{1}{n}\sum_{i=1}^{n}(\hat{y}_i - y_i)^2$$

```
predClas <- predict(RegClas,newdata = test)
ecmClas <- mean((predClas - test$medv)^2)
```

```
El ECM para el modelo clásico de regresión es 1.137293
```

```
scorestest<-predict(cps, newdata=test)
test$z1<-scorestest[,1]
test$z2<-scorestest[,2]
test$z3<-scorestest[,3]
test$z4<-scorestest[,4]
test$z5<-scorestest[,5]
test$z6<-scorestest[,6]
predCP <- predict(RegCP, newdata = test)
ecmCP <- mean((predCP - test$medv)^2)

ecm<- c(ecmClas,ecmCP)
```

```
El ECM para el modelo de regresión sobre CPs es 1.080388
```

```
plot(ecm,col=c("dark blue","dark red"),type="p",cex=1.5,
     lwd=2,xlim=c(0.5,2.5),ylim=c(1.05,1.2),
     xlab=" ",ylab="ECM")
legend("topright", inset=.02, title="Modelo de regresión",
       c("Clásico","CP"), fill=c("dark blue","dark red"), horiz=TRUE,
       cex=0.9)
```

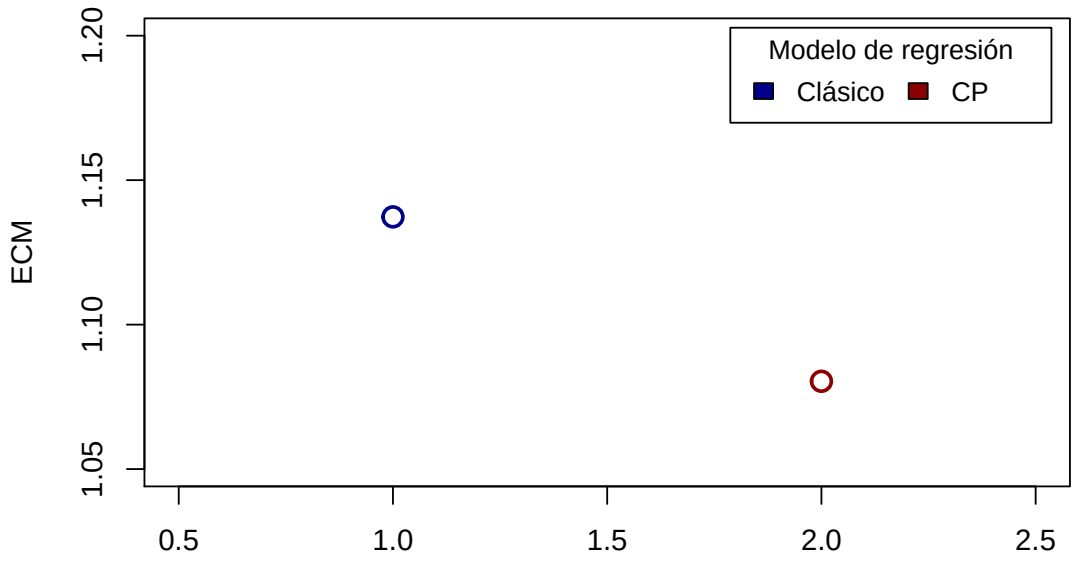

Figura 4.9: Comparación de ECM

El error de predicción de ambos modelos es bajo, siendo levemente inferior para el que hace uso de las componentes principales como variables predictoras. De esta forma, se ha conseguido reducir el error cuadrático medio además de la dimensionalidad del espacio de variables predictoras.

Sin embargo, la estimación de la varianza $\hat{\sigma}^2$ para el modelo de regresión sobre componentes principales es mayor que para el modelo de regresión clásico:

```
varmod<-c(summary(RegClas)[[6]]^2, summary(RegCP)[[6]]^2)
plot(varmod,col=c("dark blue","dark red"),type="p",cex=1.5,
     lwd=2, xlim=c(0.5,2.5), ylim=c(0.10,0.26),
     xlab=" ",ylab="Estimación de la varianza")
legend("bottomright", inset=.02, title="Modelo de regresión",
       c("Clásico","CP"), fill=c("dark blue","dark red"),
       horiz=TRUE, cex=0.9)
```

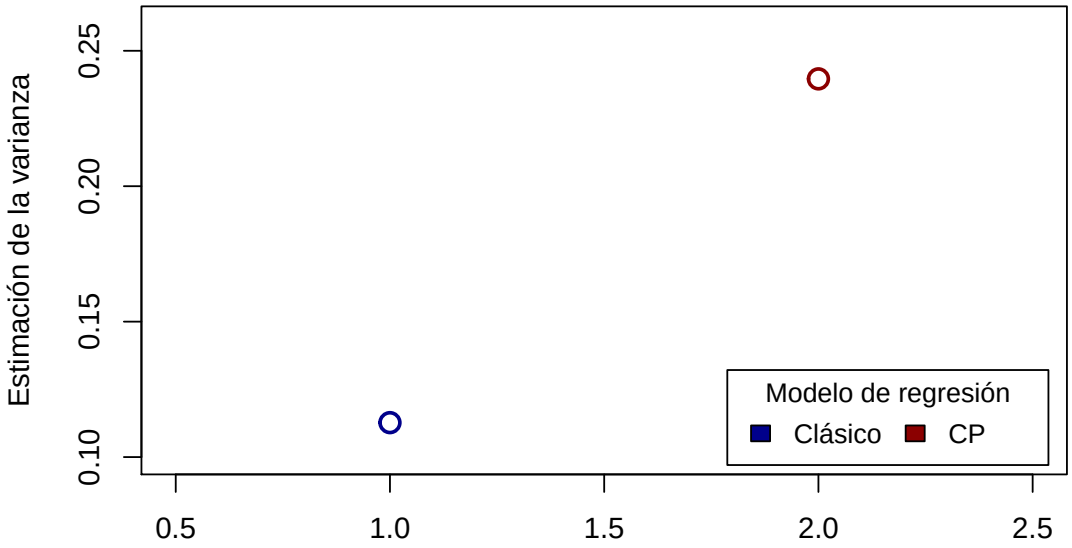

Figura 4.10: Comparación de las estimaciones de la varianza

Por último, resulta de interés el estudio de los residuos de los modelos:

```
residuosestud<- rstudent(RegClas)
residuosestudCP<- rstudent(RegCP)

par(mfrow = c(2,2))
plot(residuosestud, main ="R. Studentiz. Reg. Clásica", col="dark blue",
     xlab=" ", ylab=" ")
plot(residuosestudCP, main ="R. Studentiz. Reg. sobre CPs",
     col="dark blue", xlab=" ", ylab=" ")
hist(resid(RegClas),main='Residuos Reg. Clásica',
     xlab="Residuos estandarizados", ylab="Frecuencia",col="dark red",
     xlim=c(-1,2))
hist(resid(RegCP),main='Residuos Reg. sobre CPs',
     xlab="Residuos estandarizados", ylab="Frecuencia",col="dark red",
     xlim=c(-1,2))
```

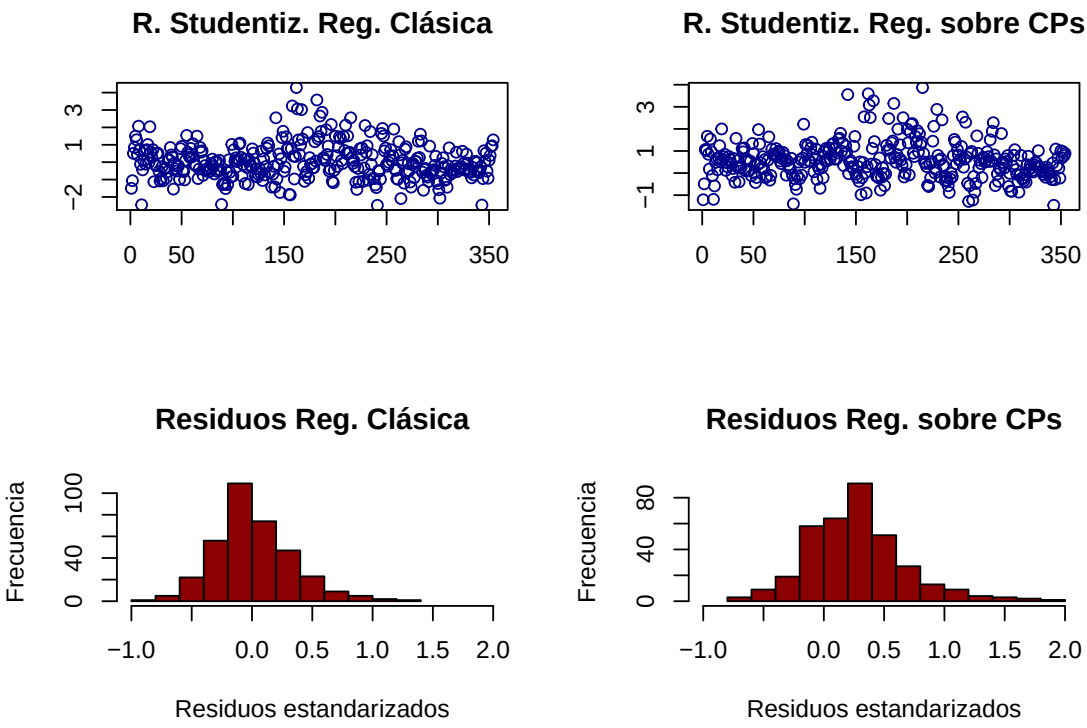

Figura 4.11: Residuos

Además, no se aprecian diferencias relevantes entre los residuos de ambos modelos.

Tras la exposición de las características del modelo de Regresión sobre Componentes Principales se concluye su calidad e idoneidad para el conjunto de datos utilizado en este trabajo, ya que se ha conseguido reducir el error cuadrático medio y el número de variables predictoras con respecto al modelo de Regresión Clásico, a pesar de haber ganado un leve aumento de la varianza.

Bibliografía

Akaike, H. (1974). A new look at the statistical model identification. *IEEE Transactions on Automatic Control*, 716–723.

Alibuhtto, M.C., Peiris, T.S.G. (2015). Principal component regression for solving multicollinearity problem. *IntSym, SEUSL.*

Allaire, J., Xie, Y., McPherson, J., Luraschi, J., Ushey, K., Atkins, A., Wickham, H., Cheng, J. & Chang, W. (2017). *Rmarkdown: Dynamic documents for r.*

Babamoradi, H., Rinnan, A. (2013). Bootstrap based confidence limits in principal component analysis: A case study. *Chemometrics and Intelligent Laboratory Systems*, **120**, 97–105.

Belsley, D.A., Kuh, E., Welsch, R.E. (1980). *Regression diagnostics. Identifying influential data and sources of collinearity.* John Wiley & Sons, New York.

Boneh, S., Mendieta, G.R. (1992). Regression modeling using principal components. *Manual on Applied Statistics in Agriculture.*

Boneh, S., Mendieta, G.R. (1994). Variable selection in regression models using principal components. *Commun. Statist. - Theor. Meth.*, **23**, 197–213.

Cai, R., Bergstrom, J.C., Mullen, J.D., Wetzstein, M.E., Shurley, W.D. (2011). Principal component analysis of crop yield response to climate change. *FS11-01. University of Georgia.*

Daling, J.R., Tamura, H. (1970). Use of orthogonal factors for selection of variables in a regression equation - an illustration. *Appl. Statist.*, **19**, 260–268.

Fekedulegn, B.D., Colbert, J.J., Hicks, R.R, Schuckers, M.E. (2002). Coping with multicollinearity: An example on application of principal components regression in dendroecology. *Manual on Applied Statistics in Agriculture.*

Hill, R.C., Fomby, T.B., Johnson, S.R. (1977). Component selection norms for principal components regression. *Commun. Statist.*, **A6**, 309–334.

Jeffers, J.N.R. (1967). Two case studies in the aplication of principal component analysis.

Appl. Statist., **16**, 225–236.

Jolliffe, I.T. (2002). *Principal component analysis*. Springer-Verlag.

Kaiser, H. (1960). The application of electronic computers to factor analysis. *Educational and Psychological Measurement*, **20**, 141–151.

Kawano, S. (2016). *spcr: Sparse Principal Component Regression*.

Kendall, M. (1957). *A course in multivariate analysis*. Griffin, London.

Kleinbaum, D.G., Kupper, L.L., Muller, K.E. (1988). *Applied regression analysis and other multivariable methods*. PWS-Kent, Boston.

Lott, W.F. (1973). The optimal set of principal component restrictions on a least squares regression. *Commun. Statist.*, **2**, 449–464.

López-González, E. (1998). Tratamiento de la colinealidad en regresión múltiple. *Psicothema*, **10(2)**, 491–507.

Luque-Calvo, P.L. (2017). *Escribir un trabajo fin de estudios con R markdown*. Disponible en http://destio.us.es/calvo.

Mansfield, E.R, Webster, J.T., Gunst, R.F. (1977). An analytic variable selection technique for principal component regression. *Appl. Statist.*, **26**, 34–40.

Mardia, K.V., Kent, J.T., Bibby, J.M. (1979). *Multivariate analysis*. London: Academic Press.

Mason, R.L., Gunst, R.F. (1985). Selecting principal components in regression. *Stat. Prob. Lett.*, **3**, 299–301.

Mertens, B., Fearn, T., Thompson, M. (1995). The efficient cross-validation of principal components applied to principal component regression. *Statist. Comput.*, **5**, 227–235.

Mevik, B.H., Wehrens, R. (2007). The pls package: principal component and partial least squares regression in R. *Journal of Statistical Software*, **18(2)**, 1–23.

Navarro, O. (2009). Selección de variables en regresión componentes principales. *LACCEI*.

NCSS, LLC. Principal component regression. *NCSS, LLC*, **34**, 1–17.

Novau, J.C., Zozaya, M.B. (1996). La multicolinealidad de los datos climáticos. La regresión en componentes principales. *Instituto Pirenaico de Ecología*.

Qi, D., Roe, B.E. (2016). Household food waste: Multivariate regression and principal components analyses of awareness and attitudes among U.S. consumers. *PLOS ONE, Public Library of Science*, **11(7)**, 1–19.

R Core Team. (2018). *R: A lenguage and environment for statistical computing*. R

Foundation for Statistical Computing, Austria.

RStudio Team. (2015). *RStudio: Integrated development environment for R*. RStudio, Inc., Boston, MA.

Schwarz, G. (1978). Estimating the dimension of a model. *The Annals of Statistics*, **6(2)**, 461–464.

Vega-Vilca, J.C., Guzmán, J. (2011). Regresión PLS y PCA como solución al problema de multicolinealidad en regresión múltiple. *Revista de matemática: Teoría y aplicaciones*, **18(1)**, 9–20.

Venables, W.N., Ripley, B.D. (2002). *Modern applied statistics with S*. Springer, New York.

Vervloet, M., Kiers, H.A.L., Van den Noortgate, W., Ceulemans, E. (2015). PCovR : An R package for principal covariates regression. *Journal of Statistical Software*, **65**, 1–14.

Zumel, N. (2016). Principal components regression: Picking the number of components. *R - Win-Vector blog*.