

UNIVERSIDAD SAN FRANCISCO DE QUITO

Paquete de R para selección de variables en regresión lineal
utilizando componentes principales

David Gonzalo Vera Alcívar

Tesis de grado presentada como requisito para la obtención del
título de Maestría en Matemáticas Aplicadas

Quito

Enero de 2007

Universidad San Francisco de Quito
Colegio de Graduados

HOJA DE APROBACION DE TESIS

Paquete de R para selección de variables en regresión lineal
utilizando componentes principales
David Gonzalo Vera Alcívar

Gonzalo Mendieta, Ph.D.
Director de Tesis	(firma)
Carlos Jiménez ,Ph.D.
Miembro del Comité	(firma)
Carlos Jiménez,Ph.D.
Director de la Maestría	(firma)
Víctor Víteri,Ph.D.
Director del Colegio de Graduados	(firma)

Quito, Enero de 2007

© Derechos de Autor
David Gonzalo Vera Alcívar
2007

Dedicatoria

A mis padres y hermanos.

A mis ahijados y padrinos.

A Cecilia, por su paciencia.

A mis verdaderos amigos.

Agradecimiento

A todos los profesores de la Maestría en Matemáticas Aplicadas de la USFQ, de manera especial a Gonzalo Mendieta, Director del presente trabajo, por su invaluable e incondicional ayuda.

David G. Vera Alcívar

Quito, Enero de 2007

Resumen

El objetivo principal de este trabajo es implementar en **R** y probar, el algoritmo por pasos de selección de variables utilizando componentes principales, propuesto por Boneh y Mendieta (1992) e implementado por primera vez en S por Walsh (1993). Los objetivos específicos son: generar un paquete de R con el método, y probar este paquete con conjuntos de datos que son usados como benchmarks para problemas de selección de variables en modelos lineales. Vamos a demostrar que este método funciona bien cuando los conjuntos de datos tienen problemas de multicolinealidad.

Abstract

The main objective of this work is to implement in **R** and prove the stepwise variable selection algorithm using principal components, proposed by Boneh and Mendieta (1992) and first implemented in S by Walsh (1993). The specific objectives are: to generate an R package with the method, and to prove this package with data sets that are used as benchmarks for variable selection in linear models problems. We are going to show that this method performs well when the data sets have multicollinearity problems.

Índice general

Dedicatoria	IV
Agradecimiento	V
Resumen	VI
Abstract	VII
1. Marco teórico	1
1.1. Introducción	1
1.2. Modelo de regresión y regresión por componentes principales (RCP)	2
1.2.1. Regresión lineal simple	3
1.2.2. Componentes principales	3
1.2.3. Regresión por componentes principales (RCP)	5
1.2.4. Métodos de selección de variables con ACP	8
2. Implementación del método de selección utilizando componentes principales	9
2.1. El algoritmo de selección	9
2.1.1. Pasos para la implementación del algoritmo	12
3. Resultados obtenidos	17
3.1. Metodología de prueba	17
3.1.1. Métodos que se usaron para comparación	18
3.2. El conjunto de datos <i>pitprop</i>	19
3.3. El conjunto de datos <i>Bodyfat</i>	20
3.4. El conjunto de datos <i>Pollution</i>	23
3.5. El conjunto de datos <i>Longley</i>	24

	IX
3.6. Conjuntos de datos simulados	25
3.6.1. Equi-correlación	25
3.6.2. Correlación markoviana	26
4. Conclusiones y Recomendaciones	28
Bibliografía	30
A. Codificación del algoritmo	31
B. Corridas en los conjuntos de datos	38
C. Corridas de la simulación	44
D. Instalación del paquete step.pc	57

Capítulo 1

Marco teórico

1.1. Introducción

La regresión lineal simple no es más que una técnica para explicar una variable cuantitativa llamada variable dependiente¹, Y , en términos de un conjunto de variables independientes² X_1, X_2, \dots, X_p por medio de una función lineal de la forma:

$$Y = \beta_0 + \beta_1 X_1 + \dots + \beta_p X_p + \epsilon$$

En forma matricial este modelo puede ser escrito de la forma $Y = X\beta + \epsilon$, donde X es la matriz $[1 \ X_1 \ \dots \ X_p]$ y β es el vector $[\beta_0 \ \beta_1 \ \dots \ \beta_p]^T$, si tenemos n observaciones entonces Y es una matriz $n \times 1$, β es una matriz $p \times 1$, X es una matriz $n \times p$, y ϵ es un vector aleatorio de errores de $n \times 1$.

Cuando un investigador desea ajustar un modelo de regresión trata de recopilar la mayor cantidad de variables que a su criterio podrían tener influencia sobre la variable dependiente, para que la mayor cantidad de variabilidad contenida en Y se explique por la variabilidad de las variables predictoras. Los problemas al agregar muchas variables al problema de regresión son:

- Muchas de estas variables podrían no tener una relación lineal con Y , en cuyo caso al agregarlas estamos restando grados de libertad innecesariamente y no estamos agregando información.
- Muchas de estas variables pueden tener una relación lineal con Y , pero a su vez también tener relación lineal con otras variables predictoras, este problema es conocido en Estadística como multicolinealidad.

¹También se la conoce como variable de respuesta

²También conocidas como variables predictoras

La idea entonces es encontrar un subconjunto de variables predictoras que den una buena estimación de la variable dependiente, que no tenga problemas de multicolinealidad, ni de variables que no expliquen a Y .

El primer problema ha sido muy estudiado, muchos algoritmos por pasos, incluidos los algoritmos de selección hacia adelante y hacia atrás basados en el valor F , en la suma cuadrada de los residuos o en el AIC, enfrentan muy bien este problema.

El segundo problema es un poco más difícil de enfrentar ya que deberíamos definir un criterio por el cual dos variables se consideren colineales, además como la idea es estimar β a partir de los datos, en la estimación por mínimos cuadrados se debe calcular $(X^T X)^{-1}$, mientras más problemas de multicolinealidad haya, el problema de invertir esa matriz se agudiza ya que ésta se vuelve casi singular.

Existen algoritmos de búsqueda exhaustiva, los cuales realizan una búsqueda sobre todo los subconjuntos posibles de variables para encontrar el mejor subconjunto de variables predictoras, estos algoritmos suelen determinar buenos subconjuntos de variables, aunque son computacionalmente exhaustivos y en ciertos casos su corrida no es factible debido a la cantidad de pasos, ya que se comportan de manera exponencial, que podría tomar³. Los algoritmos por pasos mencionados anteriormente han probado no seleccionar buenos subconjuntos cuando existen problemas de multicolinealidad en estudios de simulación[4].

Boneh y Mendieta (1992) proponen un nuevo algoritmo basado en los conceptos de componentes principales y regresión con componentes principales para seleccionar un subconjunto de variables predictoras que ha probado enfrentar muy bien el problema de la multi colinealidad, en este trabajo presentaremos el algoritmo y su justificación teórica, su implementación en \mathbf{R} , y su prueba con conjuntos de datos reales y simulados para probar la eficacia del mismo.

1.2. Modelo de regresión y regresión por componentes principales (RCP)

El modelo de regresión por componentes principales es un modelo que se ha estudiado por mucho tiempo y el objetivo es reducir la varianza del estimador de β , para introducir este modelo primero vamos a recordar algo de regresión y componentes principales.

³Si tenemos p variables predictoras, podemos tener 2^p subconjuntos posibles de variables

1.2.1. Regresión lineal simple

Como mencionamos anteriormente, el problema de regresión lineal es estimar β en el modelo $Y = X\beta + \epsilon$, para esto vamos a usar estimación de mínimos cuadrados, es decir que queremos minimizar la suma cuadrada del error:

$$SCE(\beta) = (Y - X\beta)^T(Y - X\beta)$$

Se puede demostrar fácilmente que el estimador $\hat{\beta}$ que minimiza este error viene dado por:

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

Y si se cumplen los supuestos que $E(\epsilon) = 0$ y $var(\epsilon) = \sigma^2 I$ entonces la varianza de $\hat{\beta}$ viene dada por:

$$var(\hat{\beta}) = \sigma^2 (X^T X)^{-1}$$

Este estimador tiene muchas propiedades deseadas, es insesgado, entre todos los estimadores insesgados es el de mínima varianza⁴, además si ϵ sigue una distribución normal multivariada con vector de medias 0 y matriz de covarianza $\sigma^2 I$, entonces el estimador $\hat{\beta}$ por mínimos cuadrados coincide con el estimador de máxima verosimilitud.

El problema con este estimador es que, a pesar de ser el de mínima varianza entre todos los insesgados, puede llegar a tener una varianza muy grande, recordemos que el error cuadrático esperado de un estimador siempre viene dado por:

$$Err(\hat{\theta}) = var(\hat{\theta}) + sesgo^2(\hat{\theta})$$

a veces es deseable buscar estimadores que tengan algo de sesgo pero menor varianza, a veces la reducción de la varianza puede ser tan dramática que compensa perfectamente el aumento en el sesgo ([3]).

1.2.2. Componentes principales

Las componentes principales pertenecen a un conjunto de técnicas multivariadas conocidas como análisis de interdependencia, en las componentes principales no se busca explicar variables dependientes en términos de variables independientes, sino descubrir estructuras de correlación entre las variables, estructuras que nos permitan encontrar las direcciones fundamentales de los datos, direcciones que son hiperplanos que son combinaciones lineales de las variables originales, direcciones que son ortogonales

⁴Teorema de Gauss-Markov

y tratan de explicar la mayor cantidad de variabilidad contenida en las variables originales.

Supongamos que tenemos una matriz de datos $X_{n \times p}$, entonces cada observación o fila de la matriz representa un punto en R^p , donde cada eje representa una variable, el análisis de componentes principales encuentra un hiperplano, W_1 , en este espacio tal que la dispersión de los n puntos cuando se los proyecta sobre este plano es máxima, una vez obtenida esta línea se obtiene un segundo hiperplano, W_2 tal que la dispersión de los n puntos cuando se les proyecte sobre esta línea sea máxima pero con la restricción de que esta línea sea ortogonal a la primera línea, el método sigue hasta que se hayan encontrado p direcciones mutuamente ortogonales, por la construcción de estas líneas, el primer hiperplano tendrá la mayor cantidad de varianza, el segundo hiperplano tendrá la siguiente mayor cantidad de varianza y así sucesivamente. Los hiperplanos $W_j, j = 1, \dots, p$ son llamadas las componentes principales de X .

Vamos a escribir cada componente $W_j, j = 1, \dots, p$ como una combinación lineal de las variables originales:

$$W_j = v_{j1}X_1 + v_{j2}X_2 + \dots + v_{jp}X_p$$

donde los coeficientes v_{ji} son determinados por la restricción de que la varianza de W_j es máxima y que W_j debe ser ortogonal a $W_k, (k < j)$. Se puede probar que los vectores v_j que cumplen esta restricción son los vectores propios de la matriz de covarianzas de X , si la matriz X estuviese estandarizada (es decir, cada columna restada de su media y dividida para su desviación estándar) entonces los vectores v_j son los vectores propios de $X^T X$ ([9]).

Propiedades de las componentes principales

Para las siguientes propiedades vamos a asumir que X es una matriz de rango completo de $n \times p$, donde cada fila representa una observación y cada columna una variable, vamos a asumir también que las variables han sido estandarizadas, es decir que cada variable ha sido restada de su media y dividida para su desviación estándar, lo que significa que $X^T X$ es la matriz de correlación de X . Sea $V_{p \times p}$ la matriz cuyas columnas son los valores propios normalizados de $X^T X$, y $\{\lambda_1, \lambda_2, \dots, \lambda_p\}$ sus valores propios asociados ordenados ($\lambda_1 > \lambda_2 > \dots > \lambda_p$) y sea Λ la matriz diagonal de dichos valores propios. Entonces $W_{n \times p} = XV$ es la matriz de componentes principales de X , donde cada columna W_j representa una componente principal. Algunas propiedades

importantes de W y V son:

1. Las W_j son no correlacionadas con media 0 y desviación estándar $s(W_j) = \sqrt{\lambda_j}$.
2. La matriz $W^T W$ es la matriz diagonal Λ
3. La matriz V es una matriz ortonormal, es decir que $V^T V = V V^T = I$
4. Si $W = X V$, entonces $X = W V^T$, es decir que los datos son fácilmente recuperados a partir de V y W .
5. Una medida de la variabilidad total de las W_j es:

$$\tau(W) = \sum_{j=1}^p \text{var}(W_j) = \sum_{j=1}^p \lambda_j = \text{tr}(X^T X) = \tau(X)$$

Esto demuestra que la cantidad de variabilidad total incluida en las variables originales es la misma que la incluida en las componentes principales.

6. De la propiedad anterior se deduce que la proporción de variabilidad atribuida a cada componente principal está dada por:

$$\frac{\text{var}(W_j)}{\tau(W)} = \frac{\lambda_j}{\sum_{j=1}^p \lambda_j}$$

De la última propiedad se puede ver que el estudio de la variabilidad en las X_j puede ser reducido a aquellas componentes cuya contribución es considerada grande, es decir las componentes asociadas con valores propios grandes, y como los valores propios están ordenados de mayor a menor, esto significa elegir las primeras c componentes. De aquí se deduce uno de los primeros usos que se les dio a las componentes principales: reducir la dimensionalidad del espacio de variables.

1.2.3. Regresión por componentes principales (RCP)

Consideremos el modelo $Y = X\beta + \epsilon$. Sean $V_{p \times p}$ la matriz de vectores propios normalizados de $X^T X$, Λ la matriz diagonal de los valores propios correspondientes y $W = X V$ la matriz de componentes principales de X . Vamos a escribir el modelo de regresión entonces de la siguiente manera:

$$Y = X V V^T \beta + \epsilon$$

$$Y = W \gamma + \epsilon$$

donde $\gamma = V^T\beta$ Bajo esta forma de escribir el modelo, el estimador de mínimos cuadrados de γ es:

$$\hat{\gamma} = V^T\hat{\beta} = V^T(X^T X)^{-1}X^T Y = V^T(X^T X)^{-1}V V^T X^T Y$$

$$\hat{\gamma} = (W^T W)^{-1}W^T Y = \Lambda^{-1}W^T Y$$

Entonces se define el estimador por componentes principales de β como:

$$\tilde{\beta} = V\hat{\gamma} = V\Lambda^{-1}W^T Y$$

Hasta ahora no hemos hecho mucho, solo hemos reescrito el estimador de mínimos cuadrados $\hat{\beta}$ de tal manera que no sea necesaria X sino su descomposición espectral y la variable dependiente Y , es más bajo estas condiciones se puede ver claramente que $\hat{\beta}$ coincide con $\tilde{\beta}$.

Pero la idea de las componentes principales es reducir dimensionalidad desechando las últimas componentes principales, componentes que no aporten mayormente a explicar variabilidad contenida en las variables originales. Este enfoque tiene varios problemas:

- A veces las últimas componentes no explican mucha variabilidad pero si son informativas, por ejemplo, Venables y Ripley (1999) analizan un ejemplo sobre cangrejos donde una componente con porcentaje de explicación bajo el 1 % determina el sexo de los cangrejos
- Las componentes son nuevas variables que se forman como una combinación lineal de las variables originales, estas variables, salvo casos con pocas variables, son difíciles de interpretar y además no son directamente medibles.
- Imaginemos un caso donde se midan 10 variables y se utilicen 4 componentes para explicar estas 10 variables, a pesar de que se reduce el espacio a 4 variables, el problema sigue siendo completo en las variables originales ya que para calcular las 4 componentes necesito las 10 variables.

Vamos a ver como se ve el modelo de regresión si eliminamos las componentes no significativas. Para esto vamos a asumir que solo conservamos las primeras c componentes principales ($c < p$), entonces podríamos dividir a la matriz W en $W = [W_{(c)} \ W_{(c)}^\perp]$ donde $W_{(c)}$ es una submatriz de $n \times c$ de las c primeras componentes principales y $W_{(c)}^\perp$ es una submatriz de $n \times (p - c)$ de las componentes que vamos a eliminar. Si hacemos esta

división de W , esta división también se trasladará a V y γ , es decir que $V = [V_{(c)} \ V_{(c)}^\perp]$ y $\gamma = [\gamma_{(c)} \ \gamma_{(c)}^\perp]$.

Recordemos que el modelo de regresión por componentes principales es:

$$Y = W\gamma + \epsilon$$

donde $\gamma = V^T\beta$. Al eliminar $W_{(c)}^\perp$ lo que estamos haciendo prácticamente es $\gamma_{(c)}^\perp = 0$, por la ortogonalidad de W , el estimador de mínimos cuadrados de $\gamma_{(c)}$ no es afectado por este cambio, así que nuestro estimador de γ es:

$$\tilde{\gamma} = [\tilde{\gamma}_1 \ \tilde{\gamma}_2 \ \dots \ \tilde{\gamma}_c \ 0 \ 0 \ \dots \ 0]^T$$

Entonces ahora nuestro estimado por componentes principales es:

$$\tilde{\beta} = V\tilde{\gamma} = [V_{(c)} \ V_{(c)}^\perp][\tilde{\gamma}_{(c)} \ 0]^T = V_{(c)}\tilde{\gamma}_{(c)}$$

Hay que observar que a pesar de que solo se estiman c parámetros en $\tilde{\gamma}_{(c)}$ se estiman p parámetros en $\tilde{\beta}$, como se mencionó anteriormente si se utilizan todas las componentes principales el estimador por componentes principales coincide con de estimador de mínimos cuadrados de β .

Cuando solo se elige un subconjunto de las componentes principales para hacer el ajuste, se da que $\text{var}(\tilde{\beta}) < \text{var}(\hat{\beta})$, pero por el otro lado $\tilde{\beta}$ es sesgado, es decir que $E(\tilde{\beta}) \neq \beta$. Vamos a probar esto a continuación:

Primero recordemos que $\text{var}(\hat{\beta})$, la matriz de varianzas y covarianzas de $\hat{\beta}$ viene dada por:

$$\sigma^2 V(W^T W)^{-1} V^T = \sigma^2 \sum_{k=1}^p \lambda_k^{-1} v_k v_k^T = \sigma^2 \left(\sum_{k=1}^c \lambda_k^{-1} v_k v_k^T + \sum_{k=c+1}^p \lambda_k^{-1} v_k v_k^T \right)$$

donde λ son los elementos de la diagonal de Λ . De aquí podemos deducir que:

$$\text{var}(\tilde{\beta}) = \sigma^2 \sum_{k=1}^c \lambda_k^{-1} v_k v_k^T$$

Y como la matriz de varianzas y covarianzas es definida positiva, entonces $\lambda_i > 0$, con lo que se concluye que $\text{var}(\tilde{\beta}) < \text{var}(\hat{\beta})$. Para probar que $\tilde{\beta}$ es sesgado, del resultado anterior tenemos que:

$$\tilde{\beta} = \hat{\beta} - \sum_{k=c+1}^p \lambda_k^{-1} v_k v_k^T X^T y, \quad E(\hat{\beta}) = \beta$$

y

$$E\left(\sum_{k=c+1}^p \lambda_k^{-1} v_k v_k^T X^T y\right) = \sum_{k=c+1}^p \lambda_k^{-1} v_k v_k^T X^T X \beta = \sum_{k=c+1}^p v_k v_k^T \beta$$

Dado que este último término es, en general, diferente de cero, se concluye que $E(\tilde{\beta}) \neq \beta$

Por este motivo la regresión por componentes principales pertenece a la clase de los llamados *métodos sesgados de regresión*.

1.2.4. Métodos de selección de variables con ACP

A pesar de que la regresión por componentes principales ha demostrado ser muy eficiente para reducir varianza en la estimación de β , este método adolece de ciertos problemas mencionados anteriormente, como el de que el problema sigue siendo completo en X . A pesar de esto, se han desarrollado ciertos procedimientos que utilizan componentes principales para dar una idea de selección de variables.

El primer procedimiento, desarrollado por Jeffers (1972), recomienda identificar con cada componente principal W_j la variable X_i que más carga tenga en dicha componente. Una vez que se elige un subconjunto $W_{(c)}$ de componentes significativas, las variables X_i correspondientes a $W_{(c)}^\perp$ son eliminadas. Existen varias críticas a este modelo, algunas de ellas es que no consideran las correlaciones entre las X_i 's y la respuesta Y .

Un segundo método, expuesto en Mansfield, Webster y Gunst (1977), está basado en un procedimiento de eliminación hacia atrás. En este procedimiento se selecciona un subconjunto (c) de componentes principales, se obtiene el estimado $\tilde{\beta}$ de β y elimina la variable X_i menos significativa (aquella variable que contribuye al menor incremento en la suma cuadrática de los residuos), de ahí examina las $p-1$ componentes principales y repite los pasos. Este proceso continua hasta que ninguna de las variables remanentes sea considerada significativa en la etapa de selección.

El método que vamos a estudiar propuesto por Boneh y Mendieta es por pasos y es basado en invertir repetidamente las componentes principales a las variables originales. Este método ofrece la eficiencia computacional de un procedimiento por pasos combinado con las principales fortalezas de la regresión por componentes principales: reducción de la dimensionalidad y variables predictoras no correlacionadas (por la ortogonalidad de las componentes principales).

Capítulo 2

Implementación del método de selección utilizando componentes principales

2.1. El algoritmo de selección

En este capítulo vamos a describir el método de selección propuesto por Boneh y Mendieta. En el siguiente capítulo daremos las justificaciones teóricas y los detalles computacionales. Vamos a asumir un modelo de regresión lineal normal, aunque el modelo puede ser extendido a modelos lineales generales utilizando otras pruebas. Para todo este capítulo vamos a asumir que estamos utilizando para todas las pruebas un nivel de significancia α , y que la matriz de datos es la matriz X y el vector de respuestas es el vector Y .

1. Selección de la primera variable

- a) Obtenga las principales componentes, es decir la matriz W a partir de X
- b) Haga la regresión de Y como función de W , es decir ajuste el modelo $Y = W\gamma + \epsilon$
- c) Sea $W_{(s)}$ el subconjunto de las componentes principales para las cuales el coeficiente de regresión $\hat{\gamma}_j$ sea significativo al nivel α . En el modelo normal la prueba es la típica prueba t , pero Boneh y Mendieta recomiendan que se modifique la prueba en caso de valores propios pequeños.

- d) Si $W_{(s)}$ es un conjunto vacío, entonces el proceso de selección termina con la conclusión que no se deben incluir variables predictoras en el modelo, es decir que el modelo de regresión sería $y = \beta_0$, si no es este el caso, sean $SSE_j, j = 1, \dots, p$ las sumas de cuadrado del error cuando se hace la regresión de X_j en $W_{(s)}$. La primera variable predictora seleccionada es aquella para la cual SSE_j sea mínima. Sin pérdida de generalidad vamos a asumir que se selecciona la variable X_1

2. Selección de la segunda variable

- a) Haga la regresión de (X_2, \dots, X_p) sobre X_1 y sea $\epsilon_j, j = 2, \dots, p$ el vector de residuos estandarizados $(\frac{\epsilon_j}{\sigma_{\epsilon_j}})$ obtenidos de la regresión de X_j sobre X_1
- b) Obtenga las componentes principales $[W'_2, \dots, W'_p]$ de $[\epsilon_1, \dots, \epsilon_p]$.
- c) Haga la regresión de Y sobre $[W'_2, \dots, W'_p]$, es decir ajuste el modelo $Y_i = \sum_{j=2}^p \delta_j W'_{ij} + e_i$.
- d) Sea $W'_{(s)}$ el subconjunto de W'_2, \dots, W'_p conteniendo aquellas componentes principales para las cuales el coeficiente de regresión $\hat{\delta}_j$ es significativo al nivel α (con la misma observación para valor es propios pequeños).
- e) Si $W'_{(s)}$ es vacío, entonces el proceso de selección termina con la conclusión que solo la variable X_1 debe ser incluida en el modelo, de no ser este el caso, sea $SSE'_j, j = 1, \dots, p$ las sumas de cuadrado de los errores cuando se hace la regresión de ϵ_j sobre $W'_{(s)}$. La segunda variable predictora es aquella correspondiente al SSE'_j más pequeño. Sin pérdida de generalidad asumamos que se selecciona la variable X_2 .

3. Verificación de X_1

Después de seleccionar la segunda variable se revisa si X_1 debe permanecer en el modelo.

- a) Haga la regresión de X_1 sobre X_2 y obtenga el vector de residuos estandarizados ϵ^* .
- b) Haga la regresión de Y sobre ϵ^* para ajustar el modelo de regresión simple $Y_i = \theta + \eta \epsilon_i^* + e_i$ (en realidad en el paso general se hace la regresión de Y sobre las componentes principales de los residuos estandarizados).
- c) Si $\hat{\eta}$ es significativo al nivel α , concluimos que X_1 debe permanecer en el modelo, en otro caso X_1 es sacada del modelo (técnicamente X_1 podría

entrar nuevamente al modelo en otra iteración, aunque para esta implementación vamos a quitar X_1 de toda la lista de variables, para evitar posibles ciclos).

4. **Paso general de selección (hacia adelante).** Continúe seleccionando variables de manera similar, de acuerdo a los siguientes pasos generales:

- a) Sean $X_{(s)}$ y $X_{(r)}$ los conjuntos de variables predictoras seleccionadas y no seleccionadas respectivamente. Haga la regresión de cada variable de $X_{(r)}$ sobre todas las variables en $X_{(s)}$ y obtenga el correspondiente vector de residuos estandarizados $\{\epsilon_j, j \in (r)\}$.
- b) Obtenga las componentes principales $[W]$ de los vectores ϵ_j , haga la regresión de Y sobre $[W]$.
- c) Sea $W_{(s)}$ el subconjunto de $[W]$ conteniendo las componentes principales para las cuales el coeficiente de regresión es significativo al nivel α , con la respectiva modificación para valores propios pequeños.
- d) Si $W_{(s)}$ es vacío el proceso de selección se detiene, en otro caso, sea $SSE_j, j \in (r)$ la suma cuadrática de los errores cuando se hace la regresión de ϵ_j sobre $W_{(s)}$. La siguiente variable seleccionada es la que corresponda a la SSE_j mínima.

5. **Paso general de verificación (hacia atrás).** Después de la selección de cada variable, las variables previamente seleccionadas son verificadas de acuerdo a los siguientes pasos, los cuales son esencialmente el reverso de los pasos de selección. Vamos a asumir que X_k es la variable recientemente seleccionada.

- a) Sea $X_{(c)}$ el conjunto de variables previamente seleccionadas, haga la regresión de cada una de las variables en $X_{(c)}$ sobre X_k y obtenga los correspondientes residuos estandarizados $\epsilon_{(c)}$.
- b) Obtenga las componentes principales $W_{(c)}$ de $\epsilon_{(c)}$ y haga la regresión de Y sobre $W_{(c)}$.
- c) Si todos los coeficientes de regresión son significantes al nivel α , entonces no debemos eliminar variables del modelo.
- d) Si no es este el caso, entonces una variable en $X_{(c)}$ debe ser excluida del modelo. Para detectar cual es la variable, sea $W_{(n)}$ el subconjunto de $W_{(c)}$

conteniendo las componentes principales con los coeficientes no significativos. Haga la regresión de cada vector de residuos $\epsilon_{(c)}$ sobre $W_{(n)}$, y obtenga $SSE_j, j \in (c)$.

- e) La variable en $X_{(c)}$ correspondiente a la mínima SSE_j es excluida del modelo. Si la SSE_j coincide en algunas variables, todas esas variables son excluidas. Después de que se completa este paso, se procede a seleccionar la siguiente variable. El proceso termina cuando no se encuentren componentes principales con coeficientes significativos de regresión en el paso de selección, o cuando ya no haya más variables para seleccionar.

2.1.1. Pasos para la implementación del algoritmo

A continuación vamos a describir la implementación del algoritmo. Cabe destacar que esta implementación está basada en la implementación realizada por Walsh (1993) y sigue básicamente el mismo guión, la diferencia se da en la codificación donde se lo hace en **R** en forma de paquete.

Primera y segunda selección

Para la selección de la primera variable, y en general para todo el programa, todo lo que se necesita como entrada son las matrices de correlación $X^T X$ y $X^T Y$, y el número de observaciones n (por el asunto de los grados de libertad). En estos pasos se necesita encontrar los $\hat{\gamma}$ descritos en el algoritmo y los correspondientes valores t . Para encontrar $\hat{\gamma}$ se necesita calcular: los vectores propios de $X^T X$, es decir la matriz V , y Λ^{-1} , la matriz diagonal de los valores propios inversos.

Con estos cálculos, $\hat{\gamma} = \Lambda^{-1} V^T (Y^T X)^T$, y los valores t correspondientes son:

$$t^* = \frac{\hat{\gamma}}{\hat{\sigma}(\hat{\gamma})} = \frac{1}{\sqrt{MSE} \sqrt{\Lambda^{-1}}}(\hat{\gamma})$$

donde la división es división de matrices término por término.

En este caso, $MSE = [1 - Y^T X (V \Lambda^{-1} V^T) X^T Y] / (n - p - 1)$

Entonces una componente W_j es seleccionada si $t_j > t_{n-p-1, \alpha}$ y $\lambda_j > 0,1$. Si no hay ninguna componente principal seleccionada, el proceso termina con la conclusión que no debe haber variables en el modelo. De lo contrario, proyectamos cada X_j sobre el espacio generado por las $W_{(c)}$ y consideramos la distancia $SSE(X_j) = \sum_{j \notin (c)} v_{ij}^2 \lambda_j$. Se selecciona la variable predictora correspondiente a la SSE_j mínima, la cual asumimos sin pérdida de generalidad que es X_1

Para la selección de la segunda variable, encontramos las componentes principales de los residuos estandarizados que resultan de hacer la regresión de X_j en X_1 , $j = 2, \dots, p$ y se procede como ya se indicó arriba. Es decir, realizamos los mismos pasos pero usamos como matriz de datos X la matriz $e = [e_2, e_3, \dots, e_p]$. Más adelante, en el paso general, se mostrará como calcular los e_j (los vectores del error). La suma cuadrática del error (MSE) usada en la selección de la segunda variable será la misma usada en la selección de la primera variable, de hecho, el MSE será constante hasta que una variable sea desechada del pool de variables. Cuando una variable es desechada, se recalcula el MSE para el pool de variables reducido. Recordemos que el pool de variables está conformado por las variables seleccionadas y aquellas que no han sido seleccionadas todavía, y que una vez que una variable se desecha del pool, nunca vuelve a ser considerada (para evitar posibles ciclos en el programa).

Verificación de la primera selección

Primero vamos a obtener e^* , el vector de residuos estandarizados de la regresión de X_1 en X_2 . Realizamos la regresión de Y en e^* y en X_2 , ajustando el modelo $Y_i = \theta X_{i2} + \eta e_i^* + \epsilon_i$ (en realidad no se necesita obtener e^* y ajustar el modelo explícitamente). El estadístico t usado para probar la significancia de η se calcula de la siguiente manera:

$$t^* = \frac{1}{\sqrt{MSE}}(\hat{\eta})$$

donde MSE es

$$MSE = \frac{1 - Y^T e(V\Lambda^{-1}V^T)e^T Y - Y^T X_2(X_2^T X_2)^{-1} X_2^T Y}{n - 2}$$

y donde $V\Lambda^{-1}V^T = (e^T * e^*) - 1$.

Técnicamente se hace la regresión de Y en X_2 y en las *componentes principales* de e^* , pero dado que los e^* son vectores simples, ellos son sus propias componentes principales. Por lo tanto $\hat{\eta}$ se encuentra con la misma fórmula de los $\hat{\gamma}$ ya descrita, excepto que Λ^{-1} y V^T son 1, con lo que la fórmula se reduce a $\hat{\eta} = (Y^T e^*)^T$. Más aún, dado que e^* y X_2 están estandarizados, entonces $(e^T * e^*)^{-1}$ y $(X_2^T X_2)^{-1}$ serán la unidad.

En el caso más general, solo se selecciona una variable en cada paso. Para permitir más de una selección, se necesita reemplazar X_2 en el cálculo de la media cuadrática MSE con $X_{(k)}$, donde (k) es el subconjunto que contiene los índices de dos o más variables, y ajustar los grados de libertad apropiadamente. El número de nuevas

selecciones debería ser pequeño así que el cálculo de $(X_{(k)}^T X_{(k)})^{-1}$ no deberá ser tan largo ni costoso desde el punto de vista del tiempo del proceso.

Paso general de selección (hacia adelante)

La forma del paso general hacía adelante es muy parecido al de la selección de la primera variable, con la diferencia de que X , la matriz de datos, es reemplazada por la matriz e , cuyas columnas son los vectores de residuos estandarizados $\{e_j, j \in (r)\}$. Para entender esto, vamos a usar la siguiente notación: $X_{(s)}$ es el subconjunto de variables seleccionadas; $X_{(r)}$ es el subconjunto de variables no seleccionadas o restantes. Vamos a llamar también D a $X^T X$, y F a $X^T Y$.

El primer paso es ajustar $X_{(r)}$ con $X_{(s)}$, es decir vamos a calcular los residuos de la regresión de $X_{(r)}$ sobre $X_{(s)}$:

$$e * (X_{(r)}|X_{(s)}) = (I - X_{(s)}(X_{(s)}^T X_{(s)})^{-1} X_{(s)}^T) X_{(r)}$$

Una de las fortalezas de este método es que solo se debería necesitar la información de correlación para correrlo, esto es, D y F . Bajo esta condición, $e * (X_{(r)}|X_{(s)})$ no puede ser calculado. Sin embargo esto no importa ya que lo que necesitamos es $e *^T e *$, que está dado por:

$$D(r, r) - D(r, s)[D(s, s)^{-1}]D(s, r)$$

donde $D(i, j)$ es la submatriz de las i filas y las j columnas de D .

Esta matriz $e *^T e *$ debe ser estandarizada, para lo cual lo hacemos de la forma clásica, con división término por término de matrices.

$$e^t e = \frac{e *^T e *}{\sqrt{(\text{diag}(e *^T e *)[\text{diag}(e *^T e *)]^T)}}$$

Ya con el vector de residuos estandarizados obtenemos las componentes principales $W *$ de $e(X_{(r)}|X_{(s)})$ y hacemos la regresión de Y sobre $X_{(s)}$ y $W *$, es decir que queremos ajustar el modelo $Y = X_{(s)}\beta + W * \gamma + \epsilon$, bajo la condición de que $X_{(s)}$ es ortogonal a W (esto significa que γ y β pueden ser estimados independientemente¹). El objetivo de este paso es obtener los coeficientes de regresión de $W * \hat{\gamma}$, el cual se lo obtiene de la siguiente manera:

$$\hat{\gamma} = \Lambda^{-1} V^T e^T Y$$

¹Para una nota sobre regresión ortogonal consulte el capítulo 3 de Hastie, Tibshirani y Friedman (2001)

Esta formula para $\hat{\gamma}$ hace referencia a los datos originales, en particular a Y , pero se puede demostrar que:

$$e^T Y = F^T(r) - D(r, s)[D(s, s)^{-1}]F(s)$$

Con lo que $\hat{\gamma}$ ya puede ser calculado solo con la información de correlación de los datos.

El siguiente paso es seleccionar un subconjunto $W_{*(c)}$ de W^* cuyos coeficientes en la regresión anterior sean significantes. Para esto podemos usar el siguiente estadístico de prueba:

$$t^* = \frac{\hat{\gamma}}{\sqrt{\frac{diag(\Lambda^{-1})(1 - F_{cf}^T V_{cf} \Lambda_{cf}^{-1} V_{cf}^T F_{cf})}{n - p_{cf} - 1}}}$$

donde $diag(\Lambda^{-1})$ es el vector de elementos de la diagonal de Λ^{-1} y la raíz es una operación término por término en la matriz. El subíndice cf se usa para indicar que F_{cf} , V_{cf} , Λ_{cf} y p_{cf} corresponden al modelo completo actual, recuérdese que en el paso de eliminación se pueden sacar variables las cuales nunca más vuelven a entrar al proceso, por lo que se vuelve necesario elegir las filas y columnas correspondientes a las variables que están en el pool de variables actuales². t^* entonces es un vector de $r \times 1$ cuya elemento j -ésimo es el estadístico para la prueba de hipótesis $H_0 : \gamma_j = 0$, es decir que seleccionamos una componente W_j si $t_j > t_{n-p_{cf}-1, \alpha}$, $j \in (r)$. El trabajo original de Boneh y Mendieta sugiere que en los casos en que el correspondiente λ_j sea menor a 0.1, la prueba debe ser modificada para seleccionar componentes con valor t_j mayor a 1 en valor absoluto, debido a problemas de condicionamiento en la matriz Λ . En esta implementación, debido a que sigue la implementación de Walsh, no se implementa, aunque su aplicación es directa si se lo quisiera hacer.

Una vez más se tienen varias componentes W_{*j} , se debe elegir la nueva variable predictora cuya suma cuadrática del error sea menor (recuérdese que se está utilizando la SSE como una norma). La suma cuadrática es del error con respecto a las componentes, es decir $SSE(e(X_j|X_{(s)})|W_{(c)})$ y se calcula de la misma manera que en la selección de la primera variable como $\sum_{k \notin (c)} v_{jk}^2 \lambda_k$.

Paso general de eliminación (hacia atrás)

Vamos a usar la siguiente notación como hicimos en el paso general hacia adelante, sea X_k la variable recientemente seleccionada y sea $X_{(s)}$ el conjunto de las

²Es decir cada vez que se elimina una variable, se debe recalculan estas matrices para que concuerden con el nuevo modelo completo

variables previamente seleccionadas. Muchas de las fórmulas dadas bajo el paso general hacia adelante son usadas aquí.

Primero debemos ajustar X_k en $X_{(s)}$ y calcular los residuos de dicha regresión $e * (X_{(s)}|X_k)$, sin embargo estos residuos no pueden ser calculados sin referencia a los datos, pero solo se necesita $e *^T e*$, que de igual manera que en el paso anterior se puede calcular solo con $D = X^T X$:

$$e *^T e* = D(s, s) - D(s, k)[D(k, k)]^{-1}D(k, s)$$

Y la estandarización sigue de la misma manera que en el paso anterior:

$$e^t e = \frac{e *^T e*}{\sqrt{(diag(e *^T e*)[diag(e *^T e*)]^T)}}$$

Se sigue obteniendo las componentes principales $W_{(s|k)}$ de $e(X_{(s)}|X_k)$ y hacer la regresión de Y en X_k y $W_{(s|k)}$. El objetivo de esta regresión es el mismo del paso anterior, estimar $\hat{\gamma}$, los coeficientes de regresión de $W_{(s|k)}$. La fórmula es exactamente la misma que en el paso anterior:

$$\hat{\gamma} = \Lambda^{-1}V^T(F^T(r) - D(r, s)[D(s, s)]^{-1}F(s))$$

Con la diferencia que ahora $D = e^T e$, $F = e^T Y$ y V y Λ son encontrados para $e(X_{(s)}|X_k)$.

Y de la misma manera se calcula el estadístico t para probar la significancia de los coeficientes $\hat{\gamma}$ de la regresión con la siguiente fórmula:

$$t* = \frac{\hat{\gamma}}{\sqrt{\frac{(diag(\Lambda^{-1})(1 - F^T V \Lambda^{-1} V^T F - Y^T X_k (X_k^T X_k)^{-1} X_k^T Y))}{n-p-1}}}$$

En este caso especial $p=1$ ya que se asume que solo se hizo una selección en el paso anterior³, y de nuevo $F = e^T Y$ y V y Λ son encontrados para $e(X_{(s)}|X_k)$. Se consideran las componentes principales con valor t no significativo, y finalmente se elimina la variable con la mínima suma cuadrática del error. Si no existen componentes significativas se considera que no hay que eliminar variables del modelo. La suma cuadrática del error se calcula con la siguiente fórmula:

$$SSE(e(X_j|X_k)|W_{(ns|k)}) = \sum_{i \notin (ns|k)} v_{ji}^2 \lambda_i, j \in (s)$$

Donde $W_{(ns|k)}$ es el subconjunto de $W_{(s|k)}$ conteniendo las componentes principales con los coeficientes no significativos.

³En el caso en que dos variables tenga la misma SSE se pueden hacer selecciones múltiples, en cuyo caso se reemplaza X_k por $X_{(k)}$

Capítulo 3

Resultados obtenidos

3.1. Metodología de prueba

En este capítulo vamos a probar el método. Una vez codificado el programa fue probado inicialmente con el famoso conjunto de datos *pitprop*, con el que fue probado también en el trabajo de Walsh (1993). Una vez que se comprobó la igualdad de resultados en este conjunto, se procedió a probar el programa con tres conjuntos de datos que son utilizados como benchmark para métodos de selección y que fueron ejecutados en implementaciones previas del algoritmo, además de con datos simulados bajo dos estructuras diferentes de correlación.

Se ejecutó sobre los 3 conjuntos de datos el método por ACP detallado en esta tesis, la función *step* del **R** que ejecuta un método de selección por AIC, y finalmente el método de búsqueda exhaustiva utilizando el paquete *leaps* del **R**, desarrollado por Thomas Lunley, con código en Fortran desarrollado por Alan Miller.

Sobre los datos simulados se ejecutó el método por ACP y el método basado en AIC, ya que en ellos se tiene más control del resultado que se espera.

En este informe se van a exponer los resultados de 3 conjuntos de datos reales, donde se tiene:

1. Un conjunto donde el método basado en ACP da mejores resultados que el método basado en AIC.
2. Un conjunto donde el método basado en AIC da mejores resultados que el método basado en ACP.
3. Un conjunto donde los dos métodos dan resultados iguales.

Con respecto a los datos simulados, se generaron n datos de una distribución normal multivariada con vector de medias 0 y covarianza dada por la matriz de $(p + 1) \times (p + 1)$

$$C = \begin{pmatrix} \rho_x & \rho_{xy} \\ \rho_{xy} & 1 \end{pmatrix}$$

Vamos a usar en este caso particular $p = 8$ y $n = 500$. Las correlaciones ρ_{xy} entre la respuesta y los predictores son tales que corresponden a las especificaciones particulares de las pendientes en el modelo $Y = X\beta + \epsilon$. Se utilizó un modelo con dos pendientes diferentes de cero (0.43,0.62) y la varianza del error σ igual a 1, conforme lo recomendado en [2].

Sobre la estructura de ρ_x , [2] propone correlaciones de la forma

$$\rho_x = \begin{pmatrix} A_q & 0 \\ 0 & I_{p-q} \end{pmatrix}$$

donde I es la matriz identidad de orden $p - q$ y A es una de las siguientes matrices:

Equi-correlación,

$$A = \begin{pmatrix} 1 & \rho & \rho & \rho \\ \rho & 1 & \rho & \rho \\ \rho & \rho & 1 & \rho \\ \rho & \rho & \rho & 1 \end{pmatrix}$$

Markoviana,

$$A = \begin{pmatrix} 1 & \rho & \rho^2 & \rho^3 \\ \rho & 1 & \rho & \rho^2 \\ \rho^2 & \rho & 1 & \rho \\ \rho^3 & \rho^2 & \rho & 1 \end{pmatrix}$$

3.1.1. Métodos que se usaron para comparación

Vamos a hacer una breve descripción de los métodos que se usaron para comparar los resultado del algoritmo de Boneh y Mendieta.

Método basado en AIC

El algoritmo basado en AIC no es más que un algoritmo de selección por pasos, el cual puede ser hacia adelante o hacia atrás, es decir que puede ir quitando variables de un modelo completo o ir agregando variables al modelo $y = \beta_0$. La

diferencia con los algoritmos por pasos tradicionales que usan el valor F o el coeficiente de determinación R^2 para decidir la variable que entra o sale del modelo, este método utiliza el criterio de información de Akaike, el cual para la regresión lineal simple está definido por:

$$AIC = \frac{SSE}{MSE} - n + 2(p + 1)$$

Donde n es el número de casos y p es el número de variables¹. Nótese que en el caso de la regresión lineal con el supuesto de distribución normal del error, el AIC coincide con el famoso coeficiente de Mallows C_p ([3])

El método de búsqueda exhaustiva

El método de búsqueda exhaustiva es obviamente el mejor ya que busca sobre todos los subconjuntos de variables posibles el mejor subconjunto, el problema de este método es su ineficiencia, e inclusive su no factibilidad, en conjuntos de datos grandes. En **R**, la función *leaps* realiza este procedimiento mediante un algoritmo de branch-and-bound que es muy eficiente, y a pesar de esto no puede ser ejecutado en conjuntos de datos de más de 32 variables. Este método elige el mejor subconjunto basado en tres criterios: el coeficiente de Mallows C_p , el coeficiente R^2 o el coeficiente R^2 ajustado. Para las pruebas aquí realizadas se utilizó el criterio del R^2 ajustado.

3.2. El conjunto de datos *pitprop*

El conjunto de datos *pitprop* son sobre un estudio de la resistencia a presión ciertos de troncos con el objeto de determinar si estos troncos son lo suficientemente fuertes como para usarlos en minas como soportes. Para una muestra de 180 troncos se midió la fuerza máxima a la que los mismos se partían, esta es la variable de respuesta, las variables predictoras fueron 13, cada una representando una característica especial del tronco:

1. El máximo diámetro del tronco
2. La longitud del tronco
3. La humedad del tronco
4. La gravedad específica de los tablones al momento de la prueba

¹Se asume que en el modelo se estima la constante, de lo contrario fuese p y no $p + 1$

5. La gravedad específica de los tablones al ser secados al horno
6. El número de anillos anuales en la parte superior del tronco
7. El número de anillos anuales en la base del tronco
8. El arco máximo
9. La distancia del punto del máximo arco desde la parte superior del tronco
10. El número de marcas de ramas salidas del tronco
11. La longitud de tronco sin marcas desde la parte superior
12. El número promedio de marcas por tronco
13. El diámetro promedio de las marcas

El resultado obtenido por el programa fue que se tuvo que seleccionar las variables 1, 2, 3, 4, 6, 8, 11 y 12, el cual coincide con el resultado obtenido por Walsh. Para más detalles consultar [9], el objetivo era simplemente mostrar la igualdad de resultados.

3.3. El conjunto de datos *Bodyfat*

El conjunto de datos bodyfat proviene de un experimento donde se pretendía estimar el porcentaje de grasa corporal. La medición exacta de la grasa corporal es inconveniente y costosa. Se trata de tener métodos fáciles de estimar esta grasa corporal. Un método propuesto era tomar ciertas medidas del cuerpo (altura, medida de los brazos, ancho del abdomen, etc.) y a partir de estos encontrar una regla de estimación de la grasa.

Un método preciso de estimación de la grasa corporal es el de *densidad bajo el agua*, el cual calcula el volumen del cuerpo como la diferencia entre el peso del cuerpo medido en el aire y durante inmersión en el agua, una estimación del volumen lleva directamente a la densidad del cuerpo y por tanto a la densidad de la grasa corporal. Se midió en 252 individuos la densidad bajo el agua (DUW), la cual será la respuesta. Adicionalmente a estos 252 individuos se les midió 13 características antropométricas:

1. Edad (en años)
2. Peso (en lb.)

3. Altura (en cm.)
4. Circunferencia del cuello (en cm.)
5. Circunferencia del pecho (en cm.)
6. Circunferencia del abdomen (en cm.)
7. Circunferencia de la cadera (en cm.)
8. Circunferencia del muslo (en cm.)
9. Circunferencia de la rodilla (en cm.)
10. Circunferencia de la canilla (en cm.)
11. Circunferencia del biceps extendido (en cm.)
12. Circunferencia del antebrazo (en cm.)
13. Circunferencia de la muñeca (en cm.)

Se utilizó un nivel de significancia del 5%. Vamos a describir en una forma más gráfica los resultados del experimento, las corridas detalladas del algoritmo pueden ser consultadas en el apéndice B, note que la numeración de las variables es la misma que se mencionó anteriormente:

Paso	Variable que entra	Variable que sale	Conjunto actual
1	3	-	{3}
2	8	-	{3, 8}
3	6	3, 8	{6}
4	10	-	{6, 10}
5	2	-	{6, 10, 2}
6	11	-	{6, 10, 2, 11}
7	12	-	{6, 10, 2, 11, 12}
8	13	10	{6, 2, 11, 12, 13}
9	-	-	Final de la corrida

Según la corrida del programa, para estimar la cantidad de grasa corporal, solo sería necesario medir el peso, la circunferencia del abdomen, la circunferencia del biceps extendido, la circunferencia del antebrazo y la circunferencia de la muñeca.

Vamos a mostrar ahora los resultados utilizando la función `step` y `leaps` del R, que realizan selección por AIC y búsqueda exhaustiva respectivamente.

Función	Mejor subconjunto	Valor F	Valor R^2 ajustado
Modelo completo	Todas	51.6	0.7238
<code>step</code>	{1, 2, 4, 6, 7, 8, 11, 12, 13}	74.79	0.7257
<code>leaps</code>	{1, 2, 4, 6, 7, 8, 10, 11, 12, 13}	67.68	0.7265
<code>step.pc</code>	{6, 2, 11, 12, 13}	129.1	0.7185

Nótese que el conjunto obtenido por `step.pc` es el de R^2 ajustado más bajo, aunque solo con menos de una décima. Pero sin embargo da el valor más alto del estadístico F. Las cinco variables pueden predecir el nivel de masa corporal con una pérdida mínima y no significativa en el valor R^2 .

En este conjunto de datos la función `step.pc` originó un subconjunto aceptable de variables predictoras, por sobre los otros métodos

A pesar de que el método por componentes principales genera el conjunto de variables que mejor ajusta al modelo lineal, surge la duda si será también el conjunto que mejor predice valores de Y para nuevos valores de X. Para probar esto dividimos al conjunto de datos en: conjunto de entrenamiento y conjunto de prueba, con el 70 % y el 30 % de las observaciones respectivamente. Corrimos el algoritmo `step` y `step.pc` en el conjunto de entrenamiento y usamos las variables seleccionadas para ajustar el modelo de regresión en el conjunto de prueba y calculamos el error de estimación como $\sum_{i=1}^n (y_i - \hat{y}_i)^2$ para cada algoritmo.

Los resultados que se obtuvieron fueron los siguientes:

1. El Conjunto seleccionado de variables varió bajo las nuevas circunstancias, ahora el conjunto seleccionado para `step.pc` fue {2,3,4,6,11,13}, es decir entraron las variables 3 y 4 pero salió la variable 12, es decir aumentó una variable
2. En cambio para el algoritmo `step` se redujo una variable, la variable 12, aún así `step` seleccionó 8 variables y `step.pc` 6

3. Corrimos la regresión lineal y aplicamos el modelo al conjunto de prueba, la suma cuadrática del error del modelo generado con step fue 0.0614, mientras que con step.pc 0.015, con lo cual quedó demostrado que el algoritmo step.pc, además de ajustar bien a los datos, tiene un poder predictivo mayor que el algoritmo step.

3.4. El conjunto de datos *Pollution*

Este conjunto de datos data de 1973, el objetivo era probar si la contaminación ambiental incide directamente en la mortalidad de la población. El experimento en cuestión consistía en medir ciertas medidas del ambiente en determinados lugares de EE.UU. y realizar la regresión sobre la tasa de mortalidad de los habitantes del lugar. Obviamente la variable de respuesta es la tasa de mortalidad ajustada por la edad por cada 100,000, y las mediciones ambientales (15) son:

1. Precipitación anual promedio en pulgadas
2. Temperatura promedio de Enero en grados Fahrenheit
3. Temperatura promedio de Julio en grados Fahrenheit
4. Porcentaje de 1960 de la población SMSA de 65 o más años
5. Número promedio de personas en el hogar
6. Mediana de los años de educación completados por las personas mayores a 22 años
7. Porcentaje de unidades habitacionales que son estrechas y con todas las facilidades
8. Población por millas cuadradas en áreas urbanizadas en 1960
9. Porcentaje de población no blanca en áreas urbanizadas en 1960
10. Porcentaje de la PEA ocupada en empleos de oficina
11. Porcentaje de familias con ingresos menores a \$3000
12. Potencial relativo de la polución por hidrocarbano
13. Potencial relativo de la polución por óxidos nítricos

14. Potencial relativo de la polución por bióxido de azufre
15. Porcentaje anual promedio de humedad relativa a la 1pm

Se utilizó un nivel de significancia del 5%. Vamos a mostrar ahora los resultados utilizando las funciones `step.pc`, `step` y `leaps` del R.

Función	Mejor subconjunto	Valor F	Valor R^2 ajustado
Modelo completo	Todas	9.542	0.6847
<code>step</code>	{1, 2, 3, 4, 5, 6, 9, 12, 13}	17.36	0.7138
<code>leaps</code>	{1, 2, 3, 4, 5, 6, 8, 9, 12, 13}	15.73	0.714
<code>step.pc</code>	{1, 3, 5, 8, 9}	17.21	0.6288

En este caso vemos que a pesar de que la función `step` da el mejor resultado, la función `step.pc` lo hace con casi la mitad de las variables con una pérdida de algo más de 8% en el valor R^2 . La búsqueda exhaustiva da un valor R^2 parecido al de la función `step`. En este conjunto de datos podríamos decir que no hay diferencia significativa entre los dos métodos porque a pesar de que `step.pc` da menos variables, sacrifica R^2 , y en `step` es lo contrario.

3.5. El conjunto de datos *Longley*

Este conjunto de datos fue obtenido de la base del StatLib, un repositorio en el Internet de conjuntos de datos de prueba para benchmarks de modelos. En este caso se trataba de explicar el empleo total derivado en los Estados Unidos², en base a las siguientes variables:

1. Producto interno bruto estandarizado a 1954
2. Producto interno bruto
3. Desempleo
4. Tamaño de las Fuerzas Armadas
5. Población no institucional de 14 años o más

²Este conjunto de datos es muy estudiado en econometría

6. Año

Note que esta es una serie de tiempo.

Se utilizó un nivel de significancia del 5%. Veamos ahora los resultados:

Función	Mejor subconjunto	Valor F	Valor R^2 ajustado
Modelo completo	Todas	330.3	0.9925
step	{2, 3, 4, 6}	589.8	0.9937
leaps	{2, 3, 4, 6}	589.8	0.9937
step.pc	{4, 6}	115.6	0.9286

Vemos que en este conjunto de datos el step.pc da el peor subconjunto de variables, mientras que el step y la búsqueda exhaustiva dan el mejor subconjunto.

3.6. Conjuntos de datos simulados

3.6.1. Equi-correlación

De acuerdo a la simulación planteada anteriormente, se mostrarán los resultados para el diseño de equi correlación con $\rho = 0,8$, en el cual las dos variables con pendiente diferente de cero se pondrán en dos variables independientes y en dos variables correlacionadas. Los resultados se resumen a continuación:

1. Se construyó la variable de respuesta y la matriz de correlación de los predictores de tal manera que las cuatro primeras variables (de la 1 a la 4) estén correlacionadas y las cuatro últimas (de la 5 a la 8) sean independientes. Se asignó las pendientes diferentes de cero a los predictores 1 y 4. Bajo este enfoque, la selección óptima debería ser una de las dos variables (1 o 4). Vamos a escribir los resultados, si el lector desea revisar el detalle de las corridas, puede consultar el apéndice C.

Se observó que el primer método (ACP) da como resultado que solo se debe seleccionar la variable 4, mientras que el segundo método (AIC) elige las variables 1, 2, 4 y 6. En este experimento ACP da el mejor subconjunto.

Sin embargo esta fue una sola corrida, corrimos la simulación 500 veces y se obtuvieron los siguientes resultados:

- a) 184 veces el algoritmo step seleccionó el conjunto de variables correcto, mientras que el algoritmo step.pc lo hizo 186 veces, lo cual no es una diferencia significativa.
2. Vamos a ejecutar el mismo experimento cambiando a los predictores 1 y 6 como los de pendiente significativa. Dado que estos dos predictores no son correlacionados, la selección óptima debería ser los dos predictores.

Se observó que en este caso que hay ausencia de multicolinealidad, el segundo método (AIC) eligió las variables 1, 6 y 8. A pesar que eligió una variable más (la 8), atinó en elegir las variables correctas, mientras que el primer método (ACP solo eligió la variable 1.

Sin embargo esta fue una sola corrida, corrimos la simulación 500 veces y se obtuvieron los siguientes resultados:

- a) 197 veces el algoritmo step seleccionó el conjunto de variables correcto, mientras que el algoritmo step.pc lo hizo 254 veces, es decir que en este caso el desempeño de step.pc fue mucho mejor.

3.6.2. Correlación markoviana

La corrida con correlación markoviana será ejecutado de la misma manera que con equi-correlación, con $\rho = 0,8$ y alternando dos variables como de pendiente significativa.

Si el lector desea más detalles sobre la corrida, puede consultar el apéndice C. Un resumen de los resultados se muestran a continuación:

1. En la primera corrida con las variables 1 y 4 como significativas, se obtuvo con la corrida con ACP que el conjunto debía ser la variable 1 y la variable 4, y en la corrida con AIC se obtuvo que el conjunto son las variables 1, 3, 4 y 8. Recordemos que el conjunto ideal debería ser la variable 1 o la 4, aunque en este caso las correlaciones son $\rho, \rho^2 y \rho^3$, es decir que van decreciendo. En este caso podríamos decir que aunque ninguno de los dos métodos dio el resultado que se esperaba, el método con ACP dio lo más cercano al resultado deseado.
2. En la segunda corrida con las variables 1 y 6 como significativas, el método ACP dio como conjunto las variables 1, 3 y 6 y el método AIC el conjunto de las

variables 1, 5 y 6. Recordemos que el resultado ideal debería ser la variable 1 y la 6. Aquí los dos métodos tuvieron un desempeño muy parecido, obtuvieron las variables que debían obtener y una más.

Capítulo 4

Conclusiones y Recomendaciones

En esta tesis se implementó el algoritmo desarrollado por Boneh y Mendieta para selección de variables en regresión lineal utilizando componentes principales. Este algoritmo utiliza inversiones repetidas entre los subespacios generados por las variables originales y por las componentes principales, en cada inversión va determinando las variables que quedan más “cerca” y más “lejos” de dichos subespacios con el fin de ir seleccionando las variables que entran y salen del modelo. Para la definición de distancia se utiliza la suma cuadrática del error como norma.

El resultado final del algoritmo es un subconjunto de las variables originales. Este algoritmo tiene una gran ventaja y es que puede ser ejecutado en su totalidad sin hacer referencia directa a los datos, solo se necesita la información de correlación, es decir $X^T X$ y $X^T Y$ y el número de datos (para información de los grados de libertad). Otra ventaja es que no realiza inversas de matrices, salvo en un solo paso, y dicha inversión solo es de submatrices pequeñas.

Una vez ejecutada la función con tres conjuntos de prueba y con dos conjuntos de datos simulados (véase el capítulo 3 para más detalle) se llegó a las siguientes conclusiones.

1. El algoritmo `step.pc` es muy bueno con ciertos conjuntos de datos (reales y simulados), especialmente aquellos que presentaban clara multicolinealidad en sus predictores.
2. Sin embargo este algoritmo presenta problemas cuando los coeficientes de la regresión tienen alta varianza o en la regresión con series de tiempo.
3. Una ventaja del `step.pc` sobre los demás métodos es su velocidad al trabajar solo

con la matriz de correlación de los predictores y el vector de correlación de los predictores con la respuesta.

4. La eficacia del método puede diferir si se varía el método de prueba, por ejemplo se podría comparar con el subconjunto de variables que de mejor valor F en la búsqueda exhaustiva y no el R^2 ajustado.
5. Una de las principales conclusiones de este trabajo es que ningún método de selección de variables, sean estos los algoritmos por pasos clásicos o el método basado en AIC o el método por componentes principales, funciona siempre mejor por sobre los demás métodos en todos los casos, el programa `step.pc` da resultados especialmente buenos en el caso de matrices con multicolinealidad, por lo que podríamos concluir que el método es solo una herramienta más para el analista, el cual basado en la particularidad de cada conjunto de datos, decidirá que algoritmo le da el mejor subconjunto de variables bajo sus criterios particulares de evaluación.

Como recomendación a este trabajo, la función puede implementarse para modelos lineales generales haciendo cambios en los estadísticos de prueba usados, y se pueden programar muchas funciones internas en C y unirlos con el programa en R mediante API para hacerlo más eficiente.

También se podría realizar el experimento completo de simulación diseñado en [2] para sacar conclusiones más fehacientes sobre la corrida de la función bajo situaciones de alto control.

Bibliografía

- [1] Mendieta G. Boneh S. Variable selection in regression models using principal components. *Department of Mathematics and Statistic, The Wichita State University.*
- [2] Walsh R. Boneh S., Mendieta G. A simulation study to evaluate the performance of a new variable selection method in regression. *Computational intensive methods, Interfaces, 1996.*
- [3] Tibshirani R. Hastie T. and Friedman J. . *The Elements of Statistical Learning.* Springer, first edition, 2001.
- [4] J. Edward Jackson. *A user's guide to Principal Components.*
- [5] R Development Core Team. *R: A language and environment for statistical computing.* R Foundation for Statistical Computing, Vienna, Austria, 2005. ISBN 3-900051-07-0.
- [6] Rice J. A. *Mathematical Statistics and Data Analysis.* Duxbury Press, second edition, 1995.
- [7] Venables W. and Ripley B. *Modern Applied Statistics with S-Plus.* Springer-Verlag, first edition, 1994.
- [8] Thomas Lumley using Fortran code by Alan Miller. *leaps: regression subset selection.* R package version 2.7.
- [9] Walsh R. A stepwise variable selection algorithm using principal components. Master's thesis, Department of Mathematics and Statistics, The Wichita State University, 1993.

Apéndice A

Codificación del algoritmo

El código presentado a continuación corresponde a la codificación de la función principal en *R* para la selección de variables utilizando componentes principales, la cual ya fue codificada para *S* por Walsh(1993).

```
step.pc=function (x, y, alfa = 0.05, stdz = T, n = 180)
{
  first <- function(x, y, stdz, n) {
    if (stdz == F) {
      xx <- cor(x)
      yx <- cor(x, y)
      yx <- as.vector(yx)
      tam <- dim(x)[1]
      p <- dim(x)[2]
    }
    else {
      xx <- x
      yx <- y
      tam <- n
      p <- dim(x)[1]
    }
    return(list(xx = xx, yx = yx, n = tam, p = p))
  }
  fmse <- function(yx, xx, df) {
    fdv <- eigen(xx, symmetric = T)
```

```

d <- fdv$values
v <- fdv$vectors
mse <- (1 - t(yx) %*% v %*% diag(1/d) %*% t(v) %*% yx)[1,
      1]/(df - 1)
mse
}
fres <- function(x, y, s, i) {
  g <- fg(x, s)
  a <- fa(x, s, i[-s], g)
  ee <- fee(a)
  ye <- fye(y, x, s, i[-s], g, a)
  yee <- list(yres = ye, xres = ee)
  yee
}
fg <- function(x, s) {
  g <- solve(x[s, s])
  g
}
fa <- function(x, s, ns, g) {
  a <- x[ns, ns] - x[ns, s] %*% g %*% x[s, ns]
  a
}
fee <- function(a) {
  ee <- a/sqrt(as.matrix(diag(a)) %*% t(as.matrix(diag(a))))
  ee
}
fye <- function(y, x, s, ns, g, a) {
  ye <- (y[ns] - y[s] %*% g %*% x[s, ns]) %*% diag(1/sqrt(diag(a))),
      nrow = nrow(a), ncol = ncol(a))
  ye
}
select <- function(yx, xx, mse, df, alfa) {
  sel <- fpc(yx, xx, mse, df)
  ssub <- sigsub(sel$tee, sel$d, alfa, df)
}

```

```

cat("Las componentes principales significantes son: ",
    ssub, "\n")
if (length(ssub) == 0) {
  s <- -1
  return(s)
}
s <- minsse(sel$d, sel$v, ssub)
s
}
fpc <- function(yx, xx, mse, df) {
  dv <- eigen(xx, symmetric = T)
  d <- dv$values
  v <- dv$vectors
  gh <- fgh(d, v, yx)
  tee <- sqrt(d/mse) * gh
  sel <- list(d = d, v = v, gh = gh, tee = tee)
  sel
}
fgh <- function(d, v, yx) {
  yx = as.vector(yx)
  gh <- (t(v) %*% yx)/d
  gh
}
sigsub <- function(tee, d, alfa, df) {
  tee1 <- as.matrix(tee)
  critt <- qt(1 - alfa/2, df - 1)
  if (any(abs(tee1) > critt)) {
    ssub <- row(tee1)[abs(tee1) > critt]
  }
  else {
    ssub <- numeric(0)
  }
  ssub
}

```

```

minsse <- function(d, v, sub) {
  ns <- c(1:length(d))[-sub]
  sses <- fr(d, v, ns)
  cat("Suma cuadrática del error: ", "\n")
  print(as.vector(sses), 4)
  minsse <- row(sses)[round(sses, 9) == round(min(sses),
    9)]
  minsse
}
fr <- function(d, v, ns) {
  r <- as.matrix((v * v)[, ns]) %*% d[ns]
  r
}
verify <- function(x, y, new, svars, df, alfa) {
  g <- fg(x, new)
  a <- fa(x, new, svars, g)
  ee <- fee(a)
  ye <- fye(y, x, new, svars, g, a)
  if (length(svars) < 2) {
    vt <- ft(ee, ee, ye, ye, df)
    if (abs(vt) > qt(1 - alfa/2, df)) {
      drop <- 0
    }
    else {
      drop <- 1
    }
  }
  else {
    h <- c(new, svars)
    mse <- (1 - t(y[h]) %*% solve(x[h, h]) %*% y[h])[1,
      1]/(df - 1)
    sel <- fpc(ye, ee, mse, df)
    ssub <- sigsub(sel$tee, sel$d, alfa, df)
    if (length(ssub) < length(sel$tee)) {

```

```

        wns <- c(1:length(sel$tee))[-ssub]
        drop <- minsse(sel$d, sel$v, wns)
    }
    else {
        drop <- 0
    }
}
drop
}
ft <- function(d, e, gh, y, df) {
    tee <- sqrt(d/((1 - y %*% solve(e) %*% t(y))[1, 1]/df)) *
        gh
    tee
}
stop <- F
setup <- first(x, y, stdz, n)
xx <- setup$xx
yx <- setup$yx
n <- setup$n
p <- setup$p
np <- c(1:p)
nsvars <- np
dvars <- numeric(0)
svars <- numeric(0)
full <- np
mse <- fmse(yx, xx, n - p)
s <- select(yx, xx, mse, n - p, alfa)
if (any(s < 0)) {
    cat("No hay componentes principales significativas",
        "\n")
    stop("No hay variables en el modelo")
}
cat("1ª selección es ", s, "\n")
svars <- s

```

```

nsvars <- np[-svars]
sres <- fres(xx, yx, svars, np)
s <- select(sres$yres, sres$xres, mse, n - 1, alfa)
new <- np[nsvars][s]
cat("2ª selección es: ", new, "\n")
drop <- verify(xx, yx, new, svars, n - 1, alfa)
if (any(drop) == 0) {
  svars <- c(svars, new)
  cat("1ª selección se conserva", "\n")
}
else {
  cat("Eliminando variables ", svars[drop], "\n")
  dvars <- c(dvars, svars[drop])
  svars <- c(svars[-drop], new)
}
nsvars <- np[-c(svars, dvars)]
while (!stop) {
  full <- sort(c(svars, nsvars))
  mse <- fmse(yx[full], xx[full, full], n - length(full))
  sres <- fres(xx, yx, c(svars, dvars), np)
  s <- select(sres$yres, sres$xres, mse, n - length(svars),
    alfa)
  if (any(s < 0)) {
    cat("No hay componentes principales significativas",
      "\n")
    cat("Mejor subconjunto son las variables: ", "\n",
      svars, "\n")
    break
  }
  new <- np[nsvars][s]
  cat("Siguiente selección es la variable ", new, "\n")
  cat("\n")
  drop <- verify(xx, yx, new, svars, n - 1, alfa)
  if (any(drop == 0)) {

```

```
    svars <- c(svars, new)
    cat("No se ha eliminado ninguna selección", "\n")
  }
  else {
    cat("Eliminando variable ", svars[drop], "\n")
    dvars <- c(dvars, svars[drop])
    svars <- c(svars[-drop], new)
  }
  nsvars <- np[-c(svars, dvars)]
  if (length(nsvars) == 0) {
    cat("Todas las variables han sido seleccionadas o eliminadas",
        "\n")
    cat("Las variables seleccionadas son: ", "\n", svars,
        "\n")
    break
  }
}
return(svars)
}
```

Apéndice B

Corridas en los conjuntos de datos

A continuación se muestra los resultados de las corridas realizadas sobre los conjuntos de datos:

1. Bodyfat

Las componentes principales significantes son: 1 2 3 5 6 7 10 12

Suma cuadrática del error:

[1] 0.019715 0.027482 0.007704 0.122898 0.025605 0.015293 0.059561 0.037672

[9] 0.059241 0.176536 0.259512 0.331160 0.070197

1ª selección es 3

Las componentes principales significantes son: 1 2 3 5 6 9 10 11

Suma cuadrática del error:

[1] 0.022013 0.021701 0.128766 0.025749 0.010913 0.013271 0.004642 0.055457

[9] 0.413344 0.263115 0.181534 0.062861

2ª selección es: 8

1ª selección se conserva

Las componentes principales significantes son: 1 2 3 6 7 10

Suma cuadrática del error:

[1] 0.43434 0.13817 0.25505 0.13712 0.00679 0.30716 0.39782 0.14336 0.02950

[10] 0.11699 0.32849

Siguiente selección es la variable 6

Suma cuadrática del error:

[1] 0.5634 0.5634

Eliminando variable 3 8

Las componentes principales significantes son: 1 2 4 5 9

Suma cuadrática del error:

[1] 0.6814 0.1858 0.2337 0.5930 0.1747 0.4563 0.1146 0.4091 0.2862 0.2328

Siguiente selección es la variable 10

No se ha eliminado ninguna selección

Las componentes principales significantes son: 1 2 4

Suma cuadrática del error:

[1] 0.5590 0.1685 0.3139 0.5949 0.3185 0.6202 0.5140 0.3828 0.3975

Siguiente selección es la variable 2

No se ha eliminado ninguna selección

Las componentes principales significantes son: 3 4 6 8

Suma cuadrática del error:

[1] 0.6661 0.7015 0.5942 0.7550 0.5980 0.4376 0.4755 0.7275

Siguiente selección es la variable 11

No se ha eliminado ninguna selección

Las componentes principales significantes son: 3 7

Suma cuadrática del error:

[1] 0.7720 0.8673 0.9658 0.9813 0.7139 0.4708 0.7676

Siguiente selección es la variable 12

No se ha eliminado ninguna selección

Las componentes principales significantes son: 1 3 4

Suma cuadrática del error:

[1] 0.4289 0.3993 0.3543 0.3395 0.6534 0.3185

Siguiente selección es la variable 13

Suma cuadrática del error:

[1] 0.9978 0.3771 0.9871 0.9424 0.7743

Eliminando variable 10

Las componentes principales significantes son:

No hay componentes principales significativas

Mejor subconjunto son las variables:

6 2 11 12 13

[1] 6 2 11 12 13

2. Pollution

Las componentes principales significantes son: 1 3 6 7 9 12

Suma cuadrática del error:

[1] 0.3249 0.8439 0.4014 0.7773 0.5476 0.1826 0.2213 0.3521 0.4205 0.4199

[11] 0.4028 0.4485 0.4296 0.1509 0.4227

1ª selección es 14

Las componentes principales significantes son: 1 6 7 8 11

Suma cuadrática del error:

[1] 0.2331 0.8712 0.4236 0.7909 0.5034 0.4609 0.1108 0.4967 0.4084 0.6715

[11] 0.4131 0.6126 0.6391 0.7758

2ª selección es: 7

1ª selección se conserva

Las componentes principales significantes son: 1 2 6 7 10 11

Suma cuadrática del error:

[1] 0.31142 0.17583 0.25588 0.37546 0.60028 0.62872 0.38319 0.04704 0.72303

[10] 0.25919 0.21049 0.21071 0.82108

Siguiente selección es la variable 9

Suma cuadrática del error:

[1] 0.3982 0.3982

Eliminando variable 14 7

Las componentes principales significantes son: 1 3

Suma cuadrática del error:

[1] 0.6861 0.5758 0.3358 0.8241 0.6506 0.6010 0.9399 0.5634 0.8677 0.1775

[11] 0.1833 0.7547

Siguiente selección es la variable 12

No se ha eliminado ninguna selección

Las componentes principales significantes son: 2 3

Suma cuadrática del error:

[1] 0.8644 0.7193 0.4700 0.5484 0.5242 0.4149 0.9356 0.2863 0.8726 0.9420
 [11] 0.6156
 Siguiete selección es la variable 10

No se ha eliminado ninguna selección

Las componentes principales significantes son: 2 3

Suma cuadrática del error:

[1] 0.8918 0.6016 0.4577 0.5358 0.9102 0.7448 0.5764 0.7174 0.9054 0.3850

Siguiete selección es la variable 15

Suma cuadrática del error:

[1] 0.9911 0.6046 0.6120

Eliminando variable 12

Las componentes principales significantes son: 2

Suma cuadrática del error:

[1] 0.8660 0.6324 0.6901 0.5346 0.9791 0.9647 0.8183 0.7937 0.9464

Siguiete selección es la variable 4

No se ha eliminado ninguna selección

Las componentes principales significantes son: 4

Suma cuadrática del error:

[1] 0.8983 0.9920 0.9819 0.9804 0.9882 0.6656 0.9166 0.6838

Siguiete selección es la variable 8

No se ha eliminado ninguna selección

Las componentes principales significantes son: 6

Suma cuadrática del error:

[1] 0.9970 0.9994 0.9582 0.8263 0.8840 0.9086 0.9542

Siguiete selección es la variable 5

Suma cuadrática del error:

[1] 0.4253 0.5464 0.9906 0.2325 0.9908

Eliminando variable 4

Las componentes principales significantes son: 4

Suma cuadrática del error:

[1] 0.8510 0.9830 0.6061 0.9281 0.9900 0.9528

Siguiente selección es la variable 3

Suma cuadrática del error:

[1] 0.77794 0.51266 0.04544 0.69501 0.94628

Eliminando variable 15

Las componentes principales significantes son: 3

Suma cuadrática del error:

[1] 0.1998 0.9983 0.9999 0.8337 0.9991

Siguiente selección es la variable 1

Suma cuadrática del error:

[1] 0.9463 0.2359 0.9741 0.6317 0.8520

Eliminando variable 10

Las componentes principales significantes son:

No hay componentes principales significativas

Mejor subconjunto son las variables:

9 8 5 3 1

3. Longley

Las componentes principales significantes son: 1 2 3 5

Suma cuadrática del error:

[1] 9.392e-03 4.732e-04 5.222e-06 8.922e-05 5.229e-03 1.172e-04

1ª selección es 3

Las componentes principales significantes son: 1 2 4

Suma cuadrática del error:

[1] 0.0159164 0.0006445 0.0002747 0.0091720 0.0001324

2ª selección es: 6

Eliminando variables 3

Las componentes principales significantes son: 3

Suma cuadrática del error:

[1] 0.9973 0.9754 0.5769 0.8581

Siguiente selección es la variable 4

No se ha eliminado ninguna selección

Las componentes principales significantes son:

No hay componentes principales significativas

Mejor subconjunto son las variables:

6 4

[1] 6 4

Apéndice C

Corridas de la simulación

El código presentado a continuación corresponde a la codificación de dos funciones en *R* para probar el programa `step.pc` utilizando simulación.

```
equi=function(n=50,ro=0.8,v1=1,v2=6){
  require(step.pc)
  require(MASS)
  equixxt=matrix(c(1,ro,ro,ro,rep(0,4),ro,1,ro,ro,rep(0,4),
  ro,ro,1,ro,rep(0,4),ro,ro,ro,1,rep(0,4),rep(0,4),1,rep(0,3),
  rep(0,5),1,rep(0,2),rep(0,6),1,0,rep(0,7),1),nrow=8,ncol=8)
  equix=mvrnorm(n,mu=rep(0,8),Sigma=equixxt)
  equiy=0.43*equix[,v1]+0.62*equix[,v2]+rnorm(n,0,1)
  equixx=cor(equix)
  equiyx=cor(equix,equiy)
  modstep.pc=step.pc(equixx,equiyx,stdz=T,n=n)
  equiunido=cbind(equix,equiy)
  mod1=lm(equiy~.,data=as.data.frame(equiunido))
  modstep=step(mod1)
  return(list(modstep.pc,modstep))
}
```

```
markovian=function(n=50,ro=0.8,v1=1,v2=6){
  require(step.pc)
  require(MASS)
  equixxt=matrix(c(1,ro,ro^2,ro^3,rep(0,4),ro,1,ro,ro^2,
```

```

ep(0,4),ro^2,ro,1,ro,rep(0,4),ro^3,ro^2,ro,1,rep(0,4),
rep(0,4),1,rep(0,3),rep(0,5),1,rep(0,2),rep(0,6),1,0,
rep(0,7),1),nrow=8,ncol=8)
equix=mvrnorm(n,mu=rep(0,8),Sigma=equixxt)
equiy=0.43*equix[,v1]+0.62*equix[,v2]+rnorm(n,0,1)
equixx=cor(equix)
equiyx=cor(equix,equiy)
modstep.pc=step.pc(equixx,equiyx,stdz=T,n=n)
equiunido=cbind(equix,equiy)
mod1=lm(equiy~.,data=as.data.frame(equiunido))
modstep=step(mod1)
return(list(modstep.pc,modstep))
}

```

También se muestra los resultados de las cuatro corridas realizadas:

1. Equi-correlación, variables significativas 1 y 4

Las componentes principales significantes son: 1 3 8

Suma cuadrática del error:

```
[1] 0.17480 0.03845 0.19760 0.22291 0.99189 0.13423 0.99880 0.80782
```

1ª selección es 2

Las componentes principales significantes son: 1 3 7

Suma cuadrática del error:

```
[1] 0.001479 0.195891 0.446660 0.788638 0.950365 0.980355 0.584117
```

2ª selección es: 1

Eliminando variables 2

Las componentes principales significantes son: 2

Suma cuadrática del error:

```
[1] 0.4546 0.6703 0.9230 0.9524 0.9991 0.8366
```

Siguiente selección es la variable 3

No se ha eliminado ninguna selección

Las componentes principales significantes son: 3

Suma cuadrática del error:

```
[1] 0.9651 0.9177 0.2124 0.9873 0.9214
```

Siguiente selección es la variable 6

Suma cuadrática del error:

[1] 0.8887 0.8887

Eliminando variable 1 3

Las componentes principales significantes son: 2

Suma cuadrática del error:

[1] 0.3756 0.6772 0.9757 0.9400

Siguiente selección es la variable 4

Eliminando variable 6

Las componentes principales significantes son:

No hay componentes principales significativas

Mejor subconjunto son las variables:

4

Start: AIC= 10.25

equiy ~ V1 + V2 + V3 + V4 + V5 + V6 + V7 + V8

	Df	Sum of Sq	RSS	AIC
- V3	1	0.272	43.090	8.563
- V5	1	0.735	43.554	9.098
- V7	1	1.592	44.410	10.072
- V8	1	1.733	44.551	10.231
<none>			42.818	10.247
- V2	1	2.191	45.009	10.742
- V6	1	4.906	47.724	13.671
- V4	1	4.976	47.794	13.744
- V1	1	11.610	54.429	20.243

Step: AIC= 8.56

equiy ~ V1 + V2 + V4 + V5 + V6 + V7 + V8

	Df	Sum of Sq	RSS	AIC
- V5	1	0.718	43.808	7.389

- V7	1	1.590	44.680	8.375
<none>			43.090	8.563
- V8	1	1.860	44.949	8.676
- V2	1	1.959	45.049	8.787
- V6	1	5.072	48.162	12.128
- V4	1	5.410	48.500	12.477
- V1	1	15.581	58.671	21.996

Step: AIC= 7.39

equiy ~ V1 + V2 + V4 + V6 + V7 + V8

	Df	Sum of Sq	RSS	AIC
- V7	1	1.325	45.132	6.879
<none>			43.808	7.389
- V8	1	2.143	45.950	7.777
- V2	1	2.231	46.038	7.873
- V6	1	5.017	48.825	10.811
- V4	1	5.592	49.399	11.396
- V1	1	16.856	60.663	21.666

Step: AIC= 6.88

equiy ~ V1 + V2 + V4 + V6 + V8

	Df	Sum of Sq	RSS	AIC
- V8	1	1.541	46.673	6.557
<none>			45.132	6.879
- V2	1	2.614	47.746	7.694
- V6	1	5.526	50.658	10.654
- V4	1	6.928	52.061	12.019
- V1	1	16.162	61.294	20.183

Step: AIC= 6.56

equiy ~ V1 + V2 + V4 + V6

	Df	Sum of Sq	RSS	AIC
<none>			46.673	6.557
- V2	1	2.229	48.902	6.890
- V6	1	5.502	52.175	10.129
- V4	1	6.699	53.372	11.263
- V1	1	14.683	61.356	18.233

2. Equi-correlación, variables significativas 1 y 6

Las componentes principales significantes son: 1 2 5

Suma cuadrática del error:

[1] 0.2100 0.1614 0.1679 0.1204 0.3511 0.1243 0.6513 0.6629

1ª selección es 4

Las componentes principales significantes son: 1 2 5 7

Suma cuadrática del error:

[1] 0.3309 0.2991 0.4292 0.4324 0.0408 0.7006 0.3790

2ª selección es: 6

1ª selección se conserva

Las componentes principales significantes son: 5

Suma cuadrática del error:

[1] 0.8829 0.9953 0.9995 0.7680 0.6704 0.9529

Siguiente selección es la variable 7

Suma cuadrática del error:

[1] 0.4393 0.4393

Eliminando variable 4 6

Las componentes principales significantes son: 4

Suma cuadrática del error:

[1] 0.5414 0.9976 0.7978 0.9944 0.8189

Siguiente selección es la variable 1

Eliminando variable 7

Las componentes principales significantes son:

No hay componentes principales significativas

Mejor subconjunto son las variables:

1

Start: AIC= -15.65

equiy ~ V1 + V2 + V3 + V4 + V5 + V6 + V7 + V8

	Df	Sum of Sq	RSS	AIC
- V4	1	0.011	25.522	-17.625
- V3	1	0.031	25.542	-17.586
- V2	1	0.080	25.590	-17.491
- V7	1	0.557	26.067	-16.567
- V5	1	0.731	26.242	-16.234
<none>			25.510	-15.647
- V1	1	4.123	29.633	-10.156
- V8	1	4.250	29.761	-9.942
- V6	1	10.438	35.948	-0.497

Step: AIC= -17.63

equiy ~ V1 + V2 + V3 + V5 + V6 + V7 + V8

	Df	Sum of Sq	RSS	AIC
- V3	1	0.071	25.592	-19.487
- V2	1	0.101	25.623	-19.427
- V7	1	0.550	26.072	-18.558
- V5	1	0.725	26.247	-18.224
<none>			25.522	-17.625
- V8	1	4.364	29.885	-11.733
- V1	1	5.182	30.703	-10.383
- V6	1	10.430	35.952	-2.492

Step: AIC= -19.49

equiy ~ V1 + V2 + V5 + V6 + V7 + V8

	Df	Sum of Sq	RSS	AIC
- V2	1	0.317	25.909	-20.871
- V7	1	0.529	26.121	-20.464

```

- V5    1    0.706  26.298 -20.127
<none>                25.592 -19.487
- V8    1    4.373  29.965 -13.599
- V1    1    6.024  31.617 -10.917
- V6    1   10.729  36.321  -3.981

```

Step: AIC= -20.87

equiy ~ V1 + V5 + V6 + V7 + V8

	Df	Sum of Sq	RSS	AIC
- V7	1	0.409	26.319	-22.087
- V5	1	0.569	26.479	-21.784
<none>			25.909	-20.871
- V8	1	4.563	30.472	-14.761
- V6	1	10.713	36.622	-5.568
- V1	1	15.120	41.029	0.113

Step: AIC= -22.09

equiy ~ V1 + V5 + V6 + V8

	Df	Sum of Sq	RSS	AIC
- V5	1	0.418	26.736	-23.300
<none>			26.319	-22.087
- V8	1	4.412	30.731	-16.338
- V6	1	10.305	36.624	-7.566
- V1	1	15.640	41.958	-0.767

Step: AIC= -23.3

equiy ~ V1 + V6 + V8

	Df	Sum of Sq	RSS	AIC
<none>			26.736	-23.300
- V8	1	4.378	31.114	-17.718
- V6	1	10.161	36.897	-9.194

- V1 1 15.228 41.965 -2.759

3. Correlación markoviana, variables significativas 1 y 4

Las componentes principales significantes son: 1 5 6 7

Suma cuadrática del error:

[1] 0.2674 0.1587 0.1468 0.3304 0.8311 0.6438 0.7536 0.7838

1ª selección es 3

Las componentes principales significantes son: 1 4

Suma cuadrática del error:

[1] 0.3735 0.3977 0.5362 0.7814 0.6060 0.6232 0.8854

2ª selección es: 1

Eliminando variables 3

Las componentes principales significantes son: 4

Suma cuadrática del error:

[1] 0.9927 0.4304 0.9817 0.8640 0.9433 0.9409

Siguiente selección es la variable 4

No se ha eliminado ninguna selección

Las componentes principales significantes son:

No hay componentes principales significativas

Mejor subconjunto son las variables:

1 4

Start: AIC= 13.97

equiy ~ V1 + V2 + V3 + V4 + V5 + V6 + V7 + V8

	Df	Sum of Sq	RSS	AIC
- V2	1	0.046	46.171	12.016
- V6	1	0.098	46.222	12.072
- V5	1	0.182	46.307	12.164
- V7	1	1.241	47.366	13.294
- V3	1	1.468	47.593	13.533
<none>			46.125	13.966
- V8	1	5.922	52.047	18.006
- V1	1	7.144	53.269	19.167

- V4 1 20.548 66.673 30.389

Step: AIC= 12.02

equiy ~ V1 + V3 + V4 + V5 + V6 + V7 + V8

	Df	Sum of Sq	RSS	AIC
- V6	1	0.114	46.285	10.140
- V5	1	0.214	46.385	10.248
- V7	1	1.195	47.366	11.294
<none>			46.171	12.016
- V3	1	2.208	48.379	12.353
- V8	1	5.909	52.080	16.038
- V1	1	10.147	56.318	19.949
- V4	1	21.040	67.211	28.790

Step: AIC= 10.14

equiy ~ V1 + V3 + V4 + V5 + V7 + V8

	Df	Sum of Sq	RSS	AIC
- V5	1	0.135	46.420	8.285
- V7	1	1.291	47.577	9.516
<none>			46.285	10.140
- V3	1	2.718	49.003	10.993
- V8	1	6.080	52.365	14.311
- V1	1	10.613	56.898	18.462
- V4	1	21.592	67.877	27.284

Step: AIC= 8.29

equiy ~ V1 + V3 + V4 + V7 + V8

	Df	Sum of Sq	RSS	AIC
- V7	1	1.327	47.747	7.695
<none>			46.420	8.285
- V3	1	2.596	49.016	9.006

```

- V8    1    5.961 52.381 12.326
- V1    1   10.574 56.994 16.546
- V4    1   21.659 68.079 25.432

```

Step: AIC= 7.7

```
equiy ~ V1 + V3 + V4 + V8
```

	Df	Sum of Sq	RSS	AIC
<none>			47.747	7.695
- V3	1	3.309	51.057	9.046
- V8	1	5.028	52.775	10.701
- V1	1	12.982	60.729	17.720
- V4	1	22.247	69.994	24.820

4. Correlación markoviana, variables significativas 1 y 6

Las componentes principales significantes son: 1 3 4

Suma cuadrática del error:

```
[1] 0.3239 0.2487 0.1311 0.3759 0.6266 0.0929 0.6197 0.5900
```

1ª selección es 6

Las componentes principales significantes son: 1 6

Suma cuadrática del error:

```
[1] 0.2132 0.1850 0.1301 0.4200 0.9890 0.9942 0.8556
```

2ª selección es: 3

1ª selección se conserva

Las componentes principales significantes son: 3

Suma cuadrática del error:

```
[1] 0.8401 0.9721 0.5048 0.8932 0.8709 1.0000
```

Siguiente selección es la variable 4

No se ha eliminado ninguna selección

Las componentes principales significantes son: 3

Suma cuadrática del error:

```
[1] 0.8057 0.9968 0.3870 0.9422 0.9976
```

Siguiente selección es la variable 5

Suma cuadrática del error:

[1] 0.9968 0.8617 0.8599

Eliminando variable 4

Las componentes principales significantes son: 1

Suma cuadrática del error:

[1] 0.2517 0.2710 0.9807 0.9919

Siguiente selección es la variable 1

Suma cuadrática del error:

[1] 0.8574 0.7420 0.6658

Eliminando variable 5

Las componentes principales significantes son:

No hay componentes principales significativas

Mejor subconjunto son las variables:

6 3 1

Start: AIC= 19.88

equiy ~ V1 + V2 + V3 + V4 + V5 + V6 + V7 + V8

	Df	Sum of Sq	RSS	AIC
- V2	1	0.135	52.054	18.013
- V8	1	0.168	52.086	18.043
- V7	1	0.522	52.440	18.383
- V3	1	0.691	52.609	18.543
- V4	1	0.911	52.829	18.752
- V5	1	2.072	53.990	19.839
<none>			51.918	19.882
- V1	1	10.622	62.540	27.189
- V6	1	17.059	68.977	32.088

Step: AIC= 18.01

equiy ~ V1 + V3 + V4 + V5 + V6 + V7 + V8

	Df	Sum of Sq	RSS	AIC
--	----	-----------	-----	-----

- V8	1	0.157	52.211	16.163
- V7	1	0.546	52.599	16.534
- V4	1	0.978	53.031	16.943
- V3	1	1.261	53.315	17.210
<none>			52.054	18.013
- V5	1	2.577	54.630	18.428
- V1	1	12.496	64.549	26.770
- V6	1	16.961	69.014	30.115

Step: AIC= 16.16

equiy ~ V1 + V3 + V4 + V5 + V6 + V7

	Df	Sum of Sq	RSS	AIC
- V7	1	0.439	52.650	14.582
- V4	1	1.036	53.247	15.146
- V3	1	1.189	53.399	15.289
<none>			52.211	16.163
- V5	1	2.625	54.835	16.615
- V1	1	12.707	64.917	25.055
- V6	1	17.097	69.308	28.327

Step: AIC= 14.58

equiy ~ V1 + V3 + V4 + V5 + V6

	Df	Sum of Sq	RSS	AIC
- V4	1	0.979	53.629	13.503
- V3	1	1.162	53.811	13.673
<none>			52.650	14.582
- V5	1	2.479	55.129	14.883
- V1	1	13.326	65.975	23.863
- V6	1	17.606	70.255	27.006

Step: AIC= 13.5

equiy ~ V1 + V3 + V5 + V6

	Df	Sum of Sq	RSS	AIC
- V3	1	0.328	53.956	11.808
<none>			53.629	13.503
- V5	1	3.125	56.754	14.335
- V1	1	12.637	66.266	22.082
- V6	1	19.377	73.005	26.926

Step: AIC= 11.81

equiy ~ V1 + V5 + V6

	Df	Sum of Sq	RSS	AIC
<none>			53.956	11.808
- V5	1	2.839	56.795	12.371
- V1	1	18.055	72.011	24.240
- V6	1	19.870	73.826	25.484

Apéndice D

Instalación del paquete `step.pc`

El algoritmo fue implementado en un paquete de R, al cual se lo denominó `step.pc`, el cual se adjunta a este trabajo en formato zip que es un paquete binario para Windows construido en la versión 2.1.1 del R.

El paquete se instala eligiendo la opción *Install package from local zip file* y eligiendo el archivo `.zip` entregado.

Este paquete tiene como única función la función `step.pc`, la cual se ejecuta de la siguiente manera:

```
step.pc(x,y,stdz=F)
```

Los parámetros de entrada son la matriz de predictores y el vector de la respuesta. En este caso, la bandera `stdz=F` indica que se están dando los datos originales y no la información de correlación.

```
step.pc(xx,yx,stdz=T,n=130)
```

Los parámetros de entrada son la matriz de correlación de los predictores y el vector de correlación entre los predictores y la respuesta. En este caso, la bandera `stdz=T` indica que se está dando la información de correlación, en cuyo caso también hay que dar el número de casos.

Para mayor información digite `?step.pc` una vez instalado el paquete.

El paquete también incluye la información de correlación del conjunto de datos *pitprop* para probar la función.