Selección Comprehensiva de Modelos Lineales *

Luis F. Machado *Universidad del Norte, Barranquilla, Colombia*
Jorge I. Vélez *Universidad del Norte, Barranquilla, Colombia*

Resumen: Por lo general, la selección del *mejor* modelo se realiza a través de la minimización o maximización de algún criterio, C. Este procedimiento es sencillo cuando uno, dos o hasta tres criterios se utilizan simultáneamente (por ejemplo el AIC, el R^2 y el MSE). Sin embargo, cuando se tienen $C > 3$ criterios, dicha selección es mucho más compleja. En este documento se propone, implementa e ilustra una metodología para la selección de modelos lineales denominada *Selección Comprehensiva de Modelos Lineales*. Nuestra metodología utiliza el método de todas las regresiones posibles y calcula C estadísticos o criterios de interés para los $2^k - 1$ modelos de regresión. Posteriormente, emplea análisis de componentes principales y el algoritmo *k-means* para reducir la dimensión del problema inicial, facilitar la representación gráfica de los C estadísticos, identificar subgrupos de modelos y finalmente identificar aquel modelo que tienen un mejor desempeño cuando se consideran los C citerios simultáneamente. El resultado son modelos comprehensivamente mejores para explicar la variable respuesta Y a partir de p variables explicativas. En la práctica, este tipo de modelos son deseables puesto que permiten explicar, de manera más integral, la variable de interés. Nuestra implementación en R facilita el proceso de selección y permite realizar un trabajo inferencial con mayor objetividad.

Keywords: Regresión Lineal, Análisis de Componentes Principales, Selección de Modelos

Introducción

El *Método de Todas las Regresiones Posibles* permite, a partir de un conjunto de k variables independientes $X_1, X_2, \ldots, X_k$ que potencialmente podrían explicar una respuesta continua Y, seleccionar el *mejor* subconjunto de $p < k$ predictores tal que se minimice o se maximice algún criterio C de interés. Algunos de estos criterios incluyen el R^2, $\sqrt{MSE}$, AIC, BIC, PRESS y el índice de Mallows, también conocido como C_p. El método consiste en ajustar hasta $2^k - 1$ modelos de regresión y seleccionar *el mejor* utilizando uno o más de los criterios mencionados.

Sin embargo, cuando se tienen $C > 3$ criterios, dicha selección es mucho más compleja. En este documento se propone, implementa e ilustra en R una metodología para seleccionar *comprehensivamente*, y de manera automática, el *mejor* modelo. Dicha metodología, denominada *Selección Comphensiva de Modelos Lineales* (SCML), puesto que permite incluir tantos criterios como el Científico de Datos requiera y no uno, dos o tres criterios, anteriormente escogidos para facilitar la visualización de los resultados.

Esta metodología hace uso de:

1. El método de todas las regresiones posibles para ajustar $2^k - 1$ modelos de regresión y calcular los criterios de interés;
2. Análisis de componentes principales (PCA en inglés) para reducir la dimensionalidad del problema cuando el número de criterios de selección es $C > 3$ (Lever, Krzywinski and Altman, 2017; Ringnér, 2008);

* **Autor de contacto**: jvelezv@uninorte.edu.co.

3. *k*-means *clustering* para facilitar la representación gráfica de los C estadísticos e identificar subgrupos de modelos con comportamientos similares. Este paso es válido *sólo* cuando el número de componentes principales seleccionado es > 2.

Aplicación

Una compañía fabrica tapas plásticas para botella utilizando una inyectora automática en la que se controlan los parámetros $x_1, x_2, \ldots, x_{10}$. La variable respuesta Y de interés es el peso promedio por lote, en gramos. Los parámetros del proceso pueden modificarse como $x_j < 0$, que corresponde a una *disminución*; $x_j = 0$, que es la condición de operación *actual*; y $x_j > 0$, que implica un *aumento* de x_j, donde $-1 \leq x_j \leq 1$, $j = 1, 2, \ldots, 10$. Los datos se encuentran en https://www.dropbox.com/s/bz31vuakfeonfon/inyectora.txt?dl=1

Inicialmente debemos verificar si se encuentran disponibles los paquetes `paran` y `leaps`. Posteriormente estimamos $2^k - 1 = 1023$ modelos de regresión y calculamos los C criterios de interés.

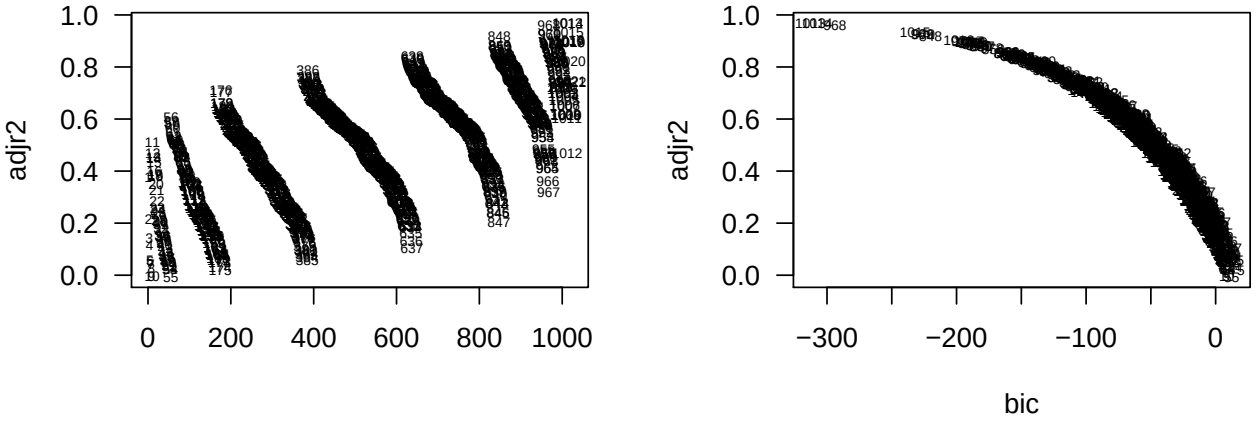

Figure 1: Desempeño de los modelos cuando $C = 1$ (izquierda) y $C = 2$ (derecha).

Para $C = 3$ criterios, tendríamos:

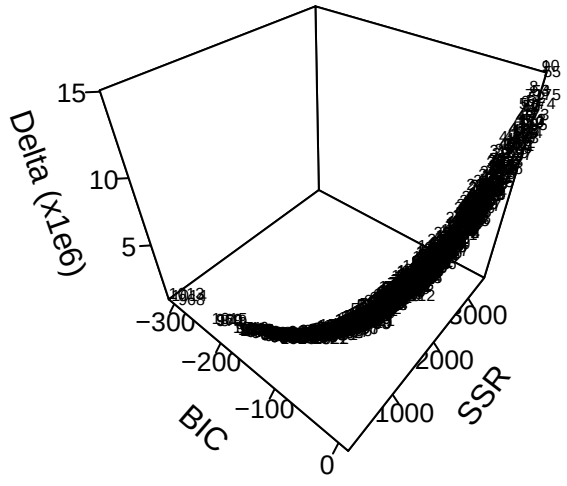

Figure 2: Desempeño de los modelos cuando $C = 3$.

Observe que el modelo `1013` resultaría seleccionado si usáramos el R^2 ajustado, el BIC o $\Delta = (C_p - p)^2$ como criterios, ya sea de manera univariada (Figura 1), bivariada (Figura 2) o simultánea (Figura 3). Qué pasa si se eligiera trabajar con *todos* los criterios disponibles? Inicialmente, la representación gráfica no es posible debido a la dimensionalidad del problema (i.e., el número de variables).

Reducción de dimensionalidad

La función `regsubsets` del paquete `leaps` permite calcular hasta cuatro criterios diferentes para seleccionar el *mejor* modelo. Otra alternativa es utilizar el paquete `olsrr` (Hebbali, 2018). Sobre estos resultados, se propone aplicar *Análisis de Componentes Principales*, seleccionar el número de componentes de manera automática haciendo uso de la función `paran` del paquete `paran` y posteriormente graficar los resutados:

```
## número de componentes que deben seleccionarse en PCA
set.seed(456)
mydata <- subset(out, select = rss:bic)
rownames(mydata) <- 1:NROW(mydata)
(k <- paran(mydata, quietly = TRUE, status = FALSE)$Retained)
```

```
##
## Using eigendecomposition of correlation matrix.
```

```
## [1] 1
```

Lo anterior indica que el número de componentes principales que deben retenerse es 1. Por lo tanto, el modelo *comprehensivamente mejor* que el resto será aquel donde **la primera componente principal sea máxima**[1]. El gráfico de la primera componente principal es:

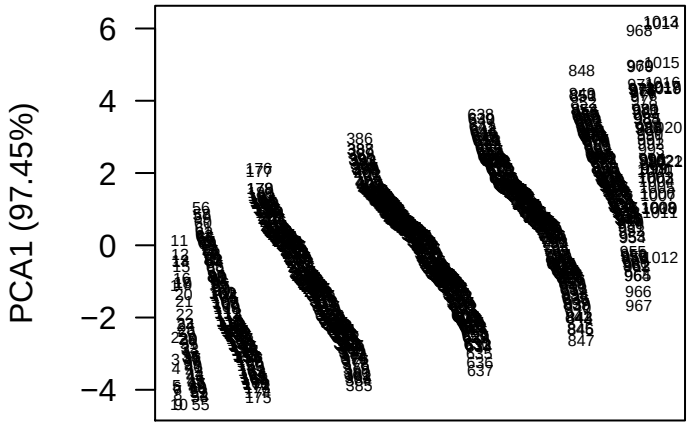

Figure 3: Primera componente principal para 4 criterios de selección.

[1]En esta aplicación en particular, este es el caso. Sin embargo, durante la fase de prueba de la metodología SCML hemos encontrado que el mejor modelo también puede estar determinado por una primera componente principal mínima.

En este caso, la primera componente explica $\sim 98\%$ de la variabilidad total:

```
##                             PC1      PC2        PC3          PC4
## Standard deviation     1.974418 0.318184 0.02082114 0.0008205804
## Proportion of Variance 0.974580 0.025310 0.00011000 0.0000000000
## Cumulative Proportion  0.974580 0.999890 1.00000000 1.0000000000
```

Modelo estimado

Este modelo *comprehensivamente mejor* tiene los siguientes indicadores:

```
##      modelo      rss     adjr2       cp      bic
## 1013   1013 98.03084 0.9696159 30.52284 -312.863
```

Comparación con el método `stepwise`

El modelo seleccionado excluye la variable `x2`. Comparativamente, los métodos de selección `backward`, `forward` y `both`, implementados en la función `?step` de R no excluyen dicha variable `x2`, como sí lo hace la metodología SCML.

Conclusiones

En aplicaciones en las que modelos parsimoniosos son considerablemente mejor percibidos, nuestra propuesta denominada *Selección Comprehensiva de Modelos Lineales* (SCML) proporciona una alternativa plausible, de fácil implementación, que puede extenderse para otro tipo de modelos y qu ofrece la posibilidad de incluir otros criterios diferentes a los ofrecidos por el paquete `leaps`.

Trabajos futuros incluyen

1. La creación de un paquete en R que permita la SCML;
2. ofrecer la posibilidad de incluir otros criterios de selección;
3. estudiar, a través de simulación estadística, el comportamiento del método SCML y compararlo con otros métodos similares (i.e., regresión *stepwise* tipo `backward`, `forward` o `both`) para el caso de modelos de regresión lineal múltiple;
4. extender la metodología SCML para *Generalized Lineal Models* y *Generalized Additive Models*.

Aspectos computacionales

La SCML está basada en los paquetes

- `leaps` (Thomas Lumley based on Fortran code by Alan Miller, 2017) para el método de todas las regresiones posibles;
- `plot3D` (Soetaert, 2017) para construir la Figura 3;
- `car` (Fox and Weisberg, 2019) para realizar la prueba de independencia y varianza constante de los residuales del modelo ajustado;
- `plot3D` (Xie, 2015) para la generación de este reporte; y
- `paran` (Dinno, 2018) para seleccionar el número de componentes principales.

References

Dinno, Alexis. 2018. *paran: Horn's Test of Principal Components/Factors*. R package version 1.5.2.
URL: *https://CRAN.R-project.org/package=paran*

Fox, John and Sanford Weisberg. 2019. *An R Companion to Applied Regression*. Third ed. Thousand Oaks CA: Sage.
URL: *https://socialsciences.mcmaster.ca/jfox/Books/Companion/*

Hebbali, Aravind. 2018. *olsrr: Tools for Building OLS Regression Models*. R package version 0.5.2.
URL: *https://CRAN.R-project.org/package=olsrr*

Lever, Jake, Martin Krzywinski and Naomi Altman. 2017. "Principal component analysis." *Nature Methods* 14:641 EP –.

Ringnér, Markus. 2008. "What is principal component analysis?" *Nature Biotechnology* 26(3):303–304.

Soetaert, Karline. 2017. *plot3D: Plotting Multi-Dimensional Data*. R package version 1.1.1.
URL: *https://CRAN.R-project.org/package=plot3D*

Thomas Lumley based on Fortran code by Alan Miller. 2017. *leaps: Regression Subset Selection*. R package version 3.0.
URL: *https://CRAN.R-project.org/package=leaps*

Xie, Yihui. 2015. *Dynamic Documents with R and knitr*. 2nd ed. Boca Raton, Florida: Chapman and Hall/CRC. ISBN 978-1498716963.
URL: *https://yihui.name/knitr/*