

Regularización en métodos de regresión

PROYECTO FINAL, CIENCIA DE DATOS

Autores: Enrique Santibáñez

02 de Junio de 2021

Centro de Investigación en Matemáticas,
Maestría en Cómputo Estadístico.



Contenido

Introducción

Métodos de regularización o *shrinkage*

RIDGE

LASSO

Elastic Net

Extensión LASSO

Ejemplos numéricos.

Delitos

Contenido de grasa

Extensión de regularización para el caso multivariado.

Ejemplo. Datos sintéticos

Conclusiones

Introducción

Motivación

En problemas con muchas variables regresoras o explicativas potenciales que pueden estar en parte altamente correlacionadas entre sí, el enfoque clásico de regresión por mínimos cuadrados puede sufrir el hecho de que los coeficientes de regresión estimados pueden llegar a estar bastante mal determinados, es decir, tener una alta varianza incluso cuando la superficie de regresión ajustada puede estar bien determinada. (Boehmke & Greenwell, 2019)

Considerando el planteamiento de regresión lineal,

$$y_i = X\beta + \epsilon, \quad (1.1)$$

donde $\beta, x_i \in R^p$, y X es una matriz de tamaño $n \times p$ con renglones x_1, x_2, \dots, x_n .

Regresión lineal

El método más frecuente para ajustar (1.1) es utilizar mínimos cuadrados ordinarios, el cual consiste en identificar como mejor modelo el hiperplano que minimiza la suma de errores cuadrados, es decir,

$$\beta^{OLS} = \arg \min_{\beta} \left\{ \|y - X\beta\|^2 \right\}. \quad (1.2)$$

Este enfoque funciona bastante bien cuando nuestros datos cumplen una serie de suposiciones:

- Relación lineal.
- Hay más observaciones (n) que variables (p), ($n > p$).
- Poca o nula colinealidad.

Entonces, en problemas cuando tenemos $n > p$ y los datos presentan multicolinealidad, una alternativa a la regresión lineal usando mínimos cuadrados es utilizar la regresión regularizada (también conocida como modelos regularizados o métodos de shrinkage) para restringir el tamaño total de todas las estimaciones de coeficientes.

Observación

Las siguientes consideraciones y métodos que aquí se presentan se pueden aplicar tanto regresión lineal múltiple y multivariada, GLM (logística y poisson) e incluso para modelos de supervivencias. Sin embargo, nos centraremos en modelos de regresión lineal múltiple la mayoría de este trabajo.

Métodos de regularización o shrinkage

Los métodos de regularización son estrategias que incorporan penalizaciones en el ajuste por mínimos cuadrados ordinarios con el objetivo de evitar

- *overfitting*
- reducir varianza (a costa de tener estimadores insesgado)
- atenuar el efecto de la correlación entre predictores
- minimizar la influencia en el modelo de los predictores menos relevantes.

La función objetivo un modelo de regresión regularizado es similar al OLS, con

$$\arg \min_{\beta} \Omega(\beta) + P(\beta). \quad (2.1)$$

Donde $\Omega(\beta)$ es una función error y $P(\beta)$ es una función regularización. (Faraway, 2006)

Este concepto se puede generalizar a todos los modelos GLM (ejemplo regresión logística y poisson). En Tibshirani, 1997 dan un ejemplo de implementación regularización en un modelo de supervivencia: Modelo de riesgo proporcional (Cox).

Tres de los métodos de regularización más empleados son

- *Ridge*
- *Lasso*
- *Elastic net*

Dado que estos métodos de regularización actúan sobre la magnitud de los coeficientes del modelo, **todos deben de estar en la misma escala, por esta razón es necesario estandarizar los predictores antes de entrenar el modelo**

Ridge

Los coeficientes de ridge minimizan una suma cuadrada del residual penalizada,

$$\hat{\beta}^{ridge}(\beta) = \arg \min_{\beta} \left\{ \|y - X\beta\|^2 + \lambda \|\beta\|^2 \right\} \quad (2.2)$$

Aquí $\lambda \geq 0$ es un parámetro de complejidad que controla la cantidad de contracción (shrinkage): cuanto mayor es el valor de λ , mayor es la cantidad de contracción. Los estimadores de regresión de ridge están dados por

$$\hat{\beta}^{ridge} = (X^T X + \lambda I)^{-1} X^T y \quad (2.3)$$

donde I es una matriz identidad de tamaño $p \times p$ (Hoerl & Kennard, 1970).

El lasso es un método de contracción como *ridge*, con diferencias sutiles pero importantes. La estimación de lazo se define por (Tibshirani, 1996)

$$\hat{\beta}^{lasso}(\beta) = \arg \min_{\beta} \left\{ \|y - X\beta\|^2 + \lambda \|\beta\|_1 \right\} \quad (2.4)$$

Mientras que la penalización de la cresta restringe a las variables a aproximadamente pero no iguales a cero, la penalización del lasso en realidad restringe los coeficientes hasta cero.

Cuando un conjunto de datos tiene muchas variables, lasso se puede utilizar para identificar las variables más relevantes.

No existe una solución explícita para este problema como en el caso de regresión *ridge*, aunque se puede resolver de manera bastante eficiente.

- Regresión de ángulo mínimo (Efron y col., 2004).
- Gradiente Descendiente (Hastie y col., 2001).

Comparación: Ridge vs Lasso

Tanto lasso (2.4) y ridge (2.2) se pueden reescribir como un problema de optimización. Una forma equivalente de escribir el problema de ridge es

$$\hat{\beta}^{ridge}(t) = \arg \min_{\beta} \|y - X\beta\|^2 \quad (2.5)$$

$$s.a. \quad \|\beta\|^2 \leq t, \quad (2.6)$$

lo que hace explícita la restricción de tamaño de los parámetros. Y para lasso el problema de optimización es

$$\hat{\beta}^{lasso}(t) = \arg \min_{\beta} \|y - X\beta\|^2 \quad (2.7)$$

$$s.a. \quad \|\beta\|_1 \leq t, \quad (2.8)$$

La principal diferencia práctica entre *lasso* y *ridge*, es que en *lasso* es posible obtener coeficientes exactamente cero. Esto supone una ventaja notable de *lasso* en escenarios donde no todos los predictores son importantes para el modelo y se desea que los menos influyentes queden excluidos.

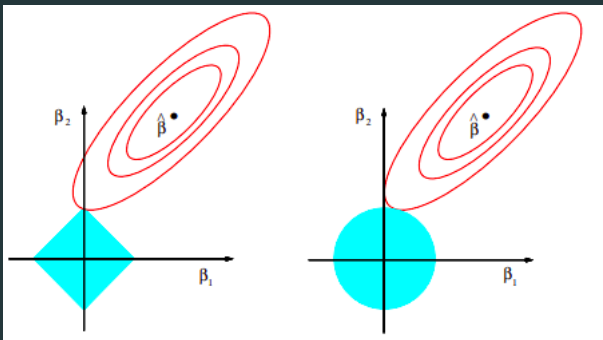


Figura 1: Regiones de los problemas de optimización, (Hastie y col., 2001)

- Cuando existen predictores altamente correlacionados (linealmente), ridge reduce la influencia de todos ellos a la vez y de forma proporcional, mientras que lasso tiende a seleccionar uno de ellos, dándole todo el peso y excluyendo al resto.
- En presencia de correlaciones, esta selección varía mucho con pequeñas perturbaciones (cambios en los datos de entrenamiento), por lo que, las soluciones de lasso, son muy inestables.

Para conseguir un equilibrio óptimo entre estas dos propiedades, se puede emplear lo que se conoce como penalización elastic net, que combina ambas estrategias.

Zou y Hastie, 2005 presentan por primera vez este enfoque de penalización, el cual es una generalización de las penalización *ridge* y *lasso*, llamado *elastic net*. La estimación de *elastic net* se define

$$\hat{\beta} = \arg \min_{\beta} \{ \|y - X\beta\|^2 + \lambda_2 \|\beta\|^2 + \lambda_1 \|\beta\| \} \quad (2.9)$$

Sea $\alpha = \lambda_2 / (\lambda_1 + \lambda_2)$, entonces resolver $\hat{\beta}$ en (2.9) es equivalente a el siguiente problema de optimización

$$\hat{\beta} = \arg \min_{\beta} \{ \|y - X\beta\|^2 \} \quad (2.10)$$

$$\text{s.a.} \quad (1 - \alpha) \|\beta\| + \alpha \|\beta\|^2 \leq t. \quad (2.11)$$

- Lasso adaptativo:

$$\arg \min_{\beta} \|y - \beta X\|^2 + \lambda \sum_{j=1}^p w_j |\beta_j|. \quad (2.12)$$

donde w es un vector de pesos conocido (Zou & Hastie, 2005).

- Lasso agrupado:

$$\arg \min_{\beta} \|y - \sum_{l=1}^L X_l \beta_l\|^2 + \lambda \sum_{l=1}^L \sqrt{p_l} \|\beta_l\|_2. \quad (2.13)$$

(Yuan & Lin, 2006)

Ejemplos numéricos.

Ejemplo: Delitos y Contenido de grasa.

A continuación veremos el efecto que tienen los diferentes tipos de regularización en dos problemas diferentes.

- Delitos: tuning y efecto del parametro λ .
- Contenido de grasa: efecto de estandarizar los datos.

Código

Ver `example_regularized_linear_regression.ipynb`

Extensión de regularización para el caso multivariado.

Regresión multivariada

La regresión multivariada es una generalización del modelo de regresión clásico pero considerando $q > 1$ variables respuestas. Es decir, sea X la matriz de las variables independientes $n \times p$, Y la matriz de las variables dependientes $n \times q$ y sea E la matriz de error aleatorio $n \times q$. Entonces el modelo de regresión multivariada es

$$Y = XB + E, \quad (4.1)$$

donde B es la matriz de coeficientes de regresión $p \times q$. Si $q = 1$ el modelo se simplifica al problema de regresión clásico donde B es el vector de coeficientes de regresión p -dimensional.

Función de verosimilitud.

La función de verosimilitud logarítmica negativa de (B, Ω) , donde $\Omega = \Sigma^{-1}$ se puede expresar como

$$g(B, \Omega) = \left[\frac{1}{n} (Y - XB)^T (Y - XB) \Omega \right] - \log(\det(\Omega)) \quad (4.2)$$

Es fácil ver (derivando con respecto a B e igualando a 0, y simplificando), que el estimador de máxima verosimilitud de B es

$$\hat{B}^{OLS} = (X^T X)^{-1} X^T Y. \quad (4.3)$$

Lo anterior es equivalente a realizar las estimaciones de B utilizando mínimos cuadrados ordinarios de forma separada para cada una de las q variables de respuestas y no este implica que no dependan de Ω .

De lo anterior podemos observar dos enfoques distintos cuando se considera una regresión multivariada. Lo primero es considerar que los datos no están correlacionados y el otro enfoque es considerar la matriz de covarianzas de los errores.

- REMMAP. El problema de minimización con restricciones propuesto por (Peng y col., 2010), consiste en optimizar

$$L(\hat{B}, X, Y) = \underset{B}{\operatorname{argmin}} \left\{ \frac{1}{2} \sum_{k=1}^q (y_k - xB_k)^2 \right\} + \lambda_1 \sum_{k=1}^q |B_k|$$

- MRCE. Rothman y col., 2010 plantea un procedimiento para construir un estimador de una matriz de coeficientes de regresión multivariada,

$$(\hat{B}, \hat{\Omega}) = \underset{B, \Omega}{\operatorname{arg\,mín}} \left\{ g(B, \Omega) + \lambda_1 \sum_{j' \neq j} |\omega_{j'j}| + \lambda_2 \sum_{j=1}^p \sum_{k=1}^q |b_{jk}| \right\} \quad (4.4)$$

Conjunto de datos

- El conjunto de datos sintéticos fue generado con la función *make_regression()* de la librería de Scikit-learn. Consideramos diferentes parámetros de la función anterior:
 $n_samples(n) = [100, 20]$, $n_features(p) = [20, 100]$, y $n_targets(q) = [2, 5]$.

Conjunto de datos

- El conjunto de datos sintéticos fue generado con la función *make_regression()* de la librería de Scikit-learn. Consideramos diferentes parámetros de la función anterior:
 $n_samples(n) = [100, 20]$, $n_features(p) = [20, 100]$, y $n_targets(q) = [2, 5]$.
- Consideramos partir el conjunto de datos original, en dos conjuntos uno de prueba y otro de entrenamiento.

Conjunto de datos

- El conjunto de datos sintéticos fue generado con la función *make_regression()* de la librería de Scikit-learn. Consideramos diferentes parámetros de la función anterior:
 $n_samples(n) = [100, 20]$, $n_features(p) = [20, 100]$, y $n_targets(q) = [2, 5]$.
- Consideramos partir el conjunto de datos original, en dos conjuntos uno de prueba y otro de entrenamiento.
- Además de que nuestro conjunto de datos, consideramos una estandarización debido a los supuestos que se tienen en los modelos.

Primeros conjunto de datos

Observando la **Figura 4**, podemos notar que cuando se consideran tamaños de $n < p$ notamos que los mejores predictores son ocupando la metodología de REMMAP.

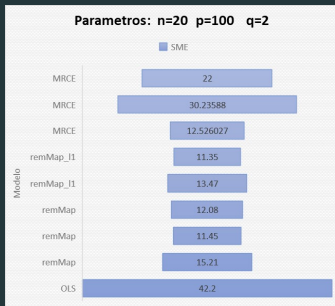


Figura 2

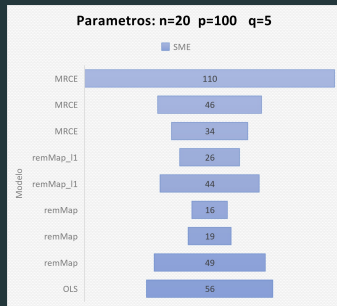


Figura 3

Figura 4: MSE considerando distintos modelos, con $n < p$.

$$n > p$$

Cuando $n > p$, podemos notar que los estimadores OLS tienen buen rendimiento. Además MRCE tiene rendimientos similares.

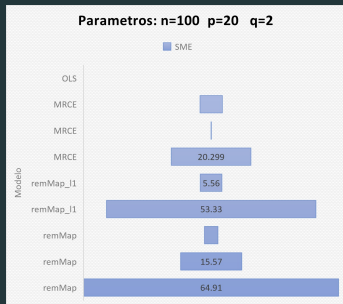


Figura 5

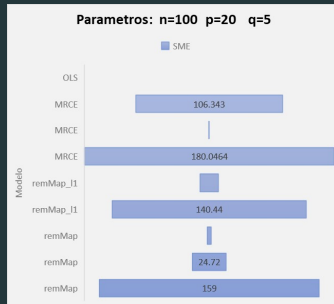


Figura 6



Figura 7: MSE considerando distintos modelos, con $n > p$.

Conclusiones

- Presentamos una alternativa importante cuando los datos incumple los supuestos de mínimos cuadrados. Estas alternativas actualmente son muy usadas en problemas de distintas disciplinas.
- Además este enfoque se puede trasladar a diferentes métodos o modelos, por ejemplo, Tibshirani, 1997 presenta una modificación al modelo de Cox utilizando regularización.

- Presentamos una alternativa importante cuando los datos incumple los supuestos de mínimos cuadrados. Estas alternativas actualmente son muy usadas en problemas de distintas disciplinas.
- Además este enfoque se puede trasladar a diferentes métodos o modelos, por ejemplo, Tibshirani, 1997 presenta una modificación al modelo de Cox utilizando regularización.
- Resaltamos las diferencias y ventajas que tiene los diferentes tipos de regularización, es decir, como regresión ridge funciona mejor en casos cuando existe multicolinealidad y regresión lasso tiene mejor rendimiento cuando existe un grupo de variables representativas en el modelo.

Referencias

-  Boehmke, B. C. & Greenwell, B. M. (2019). Hands-On Machine Learning with R.
-  Efron, B., Hastie, T., Johnstone, I. & Tibshirani, R. (2004). Least Angle Regression. *The Annals of Statistics*, 32(2), 407-451.
-  Faraway, J. (2006). *Extending the Linear Model with R: Generalized Linear, Mixed Effects and Nonparametric Regression Models*. CRC Press.
-  Hastie, T., Tibshirani, R. & Friedman, J. (2001). *The Elements of Statistical Learning*. Springer New York Inc.
-  Hoerl, A. E. & Kennard, R. W. (1970). Ridge Regression: Biased Estimation for Nonorthogonal Problems. *Technometrics*, 12, 55-67.
-  Peng, J., Zhu, J., Bergamaschi, A., Han, W., Noh, D.-Y., Pollack, J. R. & Wang, P. (2010). Regularized multivariate regression for identifying master predictors with application

to integrative genomics study of breast cancer. *The Annals of Applied Statistics*, 4(1), 53-77.

<https://doi.org/10.1214/09-AOAS271>



Rothman, A. J., Levina, E. & Zhu, J. (2010). Sparse Multivariate Regression With Covariance Estimation [PMID: 24963268]. *Journal of Computational and Graphical Statistics*, 19(4), 947-962. <https://doi.org/10.1198/jcgs.2010.09188>



Tibshirani, R. (1996). Regression Shrinkage and Selection via the Lasso. *Journal of the Royal Statistical Society (Series B)*, 58, 267-288.



Tibshirani, R. (1997). The lasso method for variable selection in the cox model. *Statistics in Medicine*, 385-395.



Yuan, M. & Lin, Y. (2006). Model selection and estimation in regression with grouped variables. *JOURNAL OF THE ROYAL STATISTICAL SOCIETY, SERIES B*, 68, 49-67.



Zou, H. & Hastie, T. (2005). Regularization and variable selection via the Elastic Net. *Journal of the Royal Statistical Society, Series B*, 67, 301-320.