Facultad de Ciencias Económicas y Estadística

Universidad Nacional de Rosario

Anteproyecto de Tesina

Regresión Lineal Múltiple en Grandes Dimensiones

Carrera: Licenciatura en Estadística

Alumno:

Iván Ariel Millanes

Directora:

Dra. Marta Quaglino

Co-Directora:

Lic. María Belén Allasia

2017

Índice

Título de la Tesina

Regresión Lineal Múltiple en Grandes Dimensiones

1. Introducción

El análisis de regresión es una técnica estadística utilizada para investigar y modelar la relación entre variables. Esta técnica estudia la relación entre una variable respuesta o dependiente y una o más variables explicativas o predictores. Se puede usar con un fin descriptivo, es decir, para conocer la función que describe la relación entre las variables, detectando cuáles de las variables explicativas están relacionadas con la respuesta y explorando la forma e intensidad de esa relación, o bien, una vez conocida esta relación, con un fin predictivo, es decir, para conocer el valor probable de la respuesta a partir del valor conocido de los predictores.

La regresión lineal fue el primer tipo de análisis de regresión en ser estudiado con rigurosidad y utilizado ampliamente en aplicaciones prácticas. En este enfoque, las relaciones se modelan usando funciones lineales en los parámetros. Por lo general, estos modelos se ajustan usando el método de estimación denominado mínimos cuadrados. Este método es muy popular debido a su fácil aplicación y buenas propiedades (Montgomery et al., 2015).

Sin embargo, en la actualidad, los grandes avances tecnológicos y la capacidad de almacenamiento creciente de los medios informáticos permite disponer de grandes bases de datos que hacen más compleja la tarea de extraer información en forma comprensible para interpretar los fenómenos investigados (Nisbet et al., 2009) (Han et al., 2011) (Leskovec et al., 2014) (Larose and Larose, 2015). En este contexto, es común encontrarse con situaciones en las que el número de variables explicativas es mucho mayor que el número de observaciones. El análisis de regresión en este escenario recibe el nombre de regresión en "grandes dimensiones".

El método de mínimos cuadrados falla en "grandes dimensiones", porque los estimadores que se obtienen no son únicos. Esta no unicidad de los estimadores hace que la interpretación de las soluciones pierda sentido, ya que para una solución el coeficiente estimado para un predictor puede ser positivo, mientras que para otra, puede ser negativo, es decir, el efecto de ese predictor sobre la respuesta depende de la solución elegida (Friedman et al., 2001).

El presente anteproyecto plantea realizar un trabajo orientado al estudio de métodos de estimación en modelos de regresión no tradicionales que se adecúen al contexto de grandes dimensiones de datos, en particular al estudio de las regresiones *Ridge* y LASSO (*Least Absolute Shrinkage and Selection Operator*), haciendo estudios comparativos que evidencien sus propiedades en cuanto a bondad de predicción del modelo y características distribucionales de los estimadores de los parámetros.

2. Objetivos

2.1. Objetivo General

Profundizar en el estudio de propuestas metodológicas para estimar los parámetros de modelos de regresión múltiple en contextos de bases de datos de grandes dimensiones donde el número de variables explicativas exceda al número de observaciones, haciendo estudios que permitan verificar de forma empírica sus propiedades.

2.2. Objetivos Específicos

- Realizar una búsqueda bibliográfica actualizada sobre métodos de estimación en regresión múltiple adecuados para "grandes dimensiones", enfocando la atención en las regresiones *Ridge* y LASSO.
- Hacer una síntesis de los métodos y su implementación.
- Indagar sobre los programas disponibles para la obtención de los estimadores de los métodos estudiados.
- Realizar estudios por simulación que permitan verificar empíricamente las propiedades de los métodos en cuanto a su capacidad predictiva.
- Estudiar propiedades distribucionales de los estimadores a través de simulaciones.

3. Metodología

Se considera una muestra aleatoria $(\boldsymbol{x}_i^T, y_i) \in \mathbb{R}^p \times \mathbb{R}$, $i = 1, ..., n$, donde $\boldsymbol{x}_i \in \mathbb{R}^p$ es un vector columna de p variables explicativas continuas del elemento i e $y_i \in \mathbb{R}$ es una variable respuesta, también continua, de ese mismo elemento. Las variables $\boldsymbol{x}_i$ e y_i se suponen relacionadas por el modelo

$$y_i = \boldsymbol{x}_i^T \boldsymbol{\beta_0} + \epsilon_i \qquad i = 1, ..., n \tag{1}$$

donde $\boldsymbol{\beta_0} \in \mathbb{R}^p$ es un vector desconocido de coeficientes y ϵ_i, $i = 1, ..., n$, son errores aleatorios con $E(\epsilon_i) = 0$. Sin pérdida de generalidad se ignora el término correspondiente a la ordenada al origen. Se asume que las variables explicativas están centradas. Escrito en forma matricial el modelo resulta

$$\boldsymbol{y} = \boldsymbol{X}\boldsymbol{\beta_0} + \boldsymbol{\epsilon} \tag{2}$$

donde $\boldsymbol{y} = (y_1, ..., y_n)^T \in \mathbb{R}^n$ es el vector de variables respuestas, $\boldsymbol{X} \in \mathbb{R}^{n \times p}$ es la matriz de variables explicativas centradas, con i-ésima fila $\boldsymbol{x}_i^T \in \mathbb{R}^p$, $i = 1, ..., n$, y j-ésima columna $\boldsymbol{X}_j \in \mathbb{R}^n$, $j = 1, ..., p$, y $\boldsymbol{\epsilon} = (\epsilon_1, ..., \epsilon_n)^T \in \mathbb{R}^n$ es el vector de errores aleatorios.

Regresión Mínimo Cuadrática

Los estimadores mínimo cuadráticos de los coeficientes del modelo (1) son aquellos coeficientes $\boldsymbol{\beta}$ que minimizan la Suma de Cuadrados del Error (SCE), es decir, se definen como la solución al siguiente problema de optimización

$$\min_{\boldsymbol{\beta} \in \mathbb{R}^p} SCE(\boldsymbol{\beta}) \iff \min_{\boldsymbol{\beta} \in \mathbb{R}^p} \sum_{i=1}^{n} (y_i - \boldsymbol{x}_i^T \boldsymbol{\beta})^2 \iff \min_{\boldsymbol{\beta} \in \mathbb{R}^p} ||\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta}||_2^2, \tag{3}$$

donde $||\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta}||_2^2$ es el cuadrado de la denominada norma 2 de $\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta}$.

Si el rango de la matriz $\boldsymbol{X}$ es igual a p, es decir, si los predictores $\boldsymbol{X}_1$, ..., $\boldsymbol{X}_p$ (columnas de la matriz $\boldsymbol{X}$, cada una de dimensión $n \times 1$) son linealmente independientes, entonces el problema de optimización mínimo cuadrático tiene solución única

$$\hat{\boldsymbol{\beta}}^{MC} = (\boldsymbol{X}^T \boldsymbol{X})^{-1} \boldsymbol{X}^T \boldsymbol{y}. \tag{4}$$

Problemas del Ajuste Mínimo Cuadrático en "Grandes Dimensiones"

Cuando el rango de la matriz $\boldsymbol{X}$ es menor que p, hecho que ocurre cuando $p > n$, existen infinitas soluciones al problema de optimización mínimo cuadrático (3). Dada una solución $\hat{\boldsymbol{\beta}}$, $\hat{\boldsymbol{\beta}} + \boldsymbol{\eta}$ también es solución para cualquier $\boldsymbol{\eta} \in null(\boldsymbol{X}) = \{\boldsymbol{\eta} \in \mathbb{R}^p : \boldsymbol{X}\boldsymbol{\eta} = \boldsymbol{0}\}$, donde $null(\boldsymbol{X})$ es el espacio nulo de la matriz $\boldsymbol{X}$, ya que

$$\boldsymbol{X}(\hat{\boldsymbol{\beta}} + \boldsymbol{\eta}) = \boldsymbol{X}\hat{\boldsymbol{\beta}} + \boldsymbol{X}\boldsymbol{\eta} = \boldsymbol{X}\hat{\boldsymbol{\beta}}, \tag{5}$$

es decir, con $\hat{\boldsymbol{\beta}}$ y $\hat{\boldsymbol{\beta}} + \boldsymbol{\eta}$ se alcanza el mismo mínimo en (3). Al no haber una solución única, la interpretación de las soluciones pierde sentido, ya que para al menos un $j \in \{1, ..., p\}$ se tendrá que $\hat{\beta}_j > 0$ para una solución $\hat{\boldsymbol{\beta}}$, pero $\tilde{\beta}_j < 0$ para otra solución $\tilde{\boldsymbol{\beta}}$, es decir, una variable explicativa influiría en forma directa e inversa sobre la respuesta simultáneamente. Este hecho también invalida las predicciones sobre nuevas observaciones (Hastie et al., 2015).

Incluso cuando el rango de la matriz $\boldsymbol{X}$ es igual a p, situación en la cual existe una solución única al problema de optimización mínimo cuadrático, puede que no sea conveniente utilizar mínimos cuadrados si p es muy cercano a n, debido a que el error de predicción dentro de la muestra del estimador mínimo cuadrático puede ser alto. Este error es igual a $\sigma^2 \frac{p}{n}$, el cual crece a medida que p se aproxima a n (siendo $n > p$).

Para tratar estos problemas, se pueden utilizar métodos de regresión penalizada, también denominados métodos de regularización. Al utilizar estos métodos, el estimador mínimo cuadrático se modifica en una de dos formas:

- Forma restringida:

$$\min_{\boldsymbol{\beta} \in \mathbb{R}^p} ||\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta}||_2^2 \quad \text{sujeto a} \quad \boldsymbol{\beta} \in C, \tag{6}$$

donde C es algún conjunto usualmente convexo,

- Forma penalizada:

$$\min_{\boldsymbol{\beta} \in \mathbb{R}^p} ||\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta}||_2^2 + P(\boldsymbol{\beta}), \tag{7}$$

donde $P(\cdot)$ es alguna función de penalización usualmente convexa.

Los métodos de regularización buscan una reducción importante en la variancia de

las estimaciones a costa de la introducción de algo de sesgo, lo que mejora los resultados globalmente (Bühlmann and Van De Geer, 2011).

Las regresiones *Ridge* y LASSO son métodos de regularización utilizados en el contexto de grandes dimensiones.

Regresión *Ridge*

La regresión *Ridge* fue presentada por Hoerl and Kennard (1970) como una alternativa a los estimadores mínimo cuadráticos en presencia de multicolinealidad. Se trata de un proceso continuo que contrae los coeficientes y por lo tanto es estable. Sin embargo, no fija los coeficientes de variables muy poco asociadas con la respuesta exactamente en cero, razón por la cual no provee modelos fácilmente interpretables en presencia de muchas variables explicativas.

El estimador *Ridge* se define como

$$\hat{\boldsymbol{\beta}}^{Ridge} = \underset{\boldsymbol{\beta}\in\mathbb{R}^p}{\operatorname{argmin}} \left\{ \frac{1}{2}\sum_{i=1}^{n}(y_i - \boldsymbol{x}_i^T\boldsymbol{\beta})^2 + \lambda\sum_{j=1}^{p}\beta_j^2 \right\}, \tag{8}$$

donde $\lambda > 0$ es una constante denominada parámetro de suavizado.

Regresión LASSO

En 1996, Tibshirani, intentando retener lo mejor de la selección de un subconjunto de variables y de la regresión *Ridge*, propuso la técnica denominada LASSO, la cual contrae algunos coeficientes y fija en cero a otros (Tibshirani, 1996) (Tibshirani et al., 2013).

El estimador LASSO se define como

$$\hat{\boldsymbol{\beta}}^{LASSO} = \underset{\boldsymbol{\beta}\in\mathbb{R}^p}{\operatorname{argmin}} \left\{ \frac{1}{2}\sum_{i=1}^{n}(y_i - \boldsymbol{x}_i^T\boldsymbol{\beta})^2 + \lambda\sum_{j=1}^{p}|\beta_j| \right\}, \tag{9}$$

donde $\lambda > 0$ es una constante denominada parámetro de suavizado.

En ambos métodos, la solución depende de la elección de λ, dando lugar a un camino de soluciones (Osborne et al., 2000) (Efron et al., 2004) (Tibshirani, 2011). De todas las soluciones calculadas para algunas variantes de λ previamente establecidas, se elige la

mejor de acuerdo a algún criterio (como por ejemplo, minimización del Error Cuadrático Medio) utilizando validación cruzada.

4. Aplicación

4.1. Simulaciones

Se propone estudiar el comportamiento de los estimadores, en principio de lo métodos *Ridge* y LASSO, mediante la simulación de escenarios donde el número de variables explicativas es mayor que el número de observaciones, con el objetivo de comparar sus propiedades, así como la bondad de predicción de los modelos.

Se plantea la inclusión de un escenario en el cual los estimadores mínimo cuadráticos puedan ser calculados en forma única ($n = p$).

Las comparaciones se realizarán teniendo en cuenta medidas globales de capacidad predictiva del modelo y propiedades distribucionales de los estimadores de cada parámetro del modelo, enfocando en especial sesgo y precisión.

5. Cronograma

Marzo 2017 - Abril 2017

- Investigación bibliográfica.
- Adquisición de conocimientos y experiencia con los métodos de estimación de parámetros en modelos de regresión lineal múltiple en el contexto de grandes dimensiones de datos.

Mayo 2017 - Junio 2017

- Estudio de las regresiones Ridge y LASSO y de sus propiedades.
- Revisión de las implementaciones en el software R

Julio 2017 - Septiembre 2017

- Estudio por simulación para evaluar el comportamiento de los estimadores.

- Redacción de la tesina.

Octubre 2017

- Revisión general del trabajo.

Noviembre 2017

- Presentación de la tesina.

6. Bibliografía

Bühlmann, P. and Van De Geer, S. (2011). *Statistics for high-dimensional data: methods, theory and applications*. Springer Science & Business Media.

Efron, B., Hastie, T., Johnstone, I., Tibshirani, R., et al. (2004). Least angle regression. *The Annals of statistics*, 32(2):407–499.

Friedman, J., Hastie, T., and Tibshirani, R. (2001). *The elements of statistical learning*, volume 1. Springer series in statistics Springer, Berlin.

Han, J., Pei, J., and Kamber, M. (2011). *Data mining: concepts and techniques*. Elsevier.

Hastie, T., Tibshirani, R., and Wainwright, M. (2015). *Statistical learning with sparsity: the lasso and generalizations*. CRC press.

Hoerl, A. E. and Kennard, R. W. (1970). Ridge regression: Biased estimation for nonorthogonal problems. *Technometrics*, 12(1):55–67.

Larose, D. T. and Larose, C. D. (2015). *Data mining and predictive analytics*. John Wiley & Sons.

Leskovec, J., Rajaraman, A., and Ullman, J. D. (2014). *Mining of massive datasets*. Cambridge University Press.

Montgomery, D. C., Peck, E. A., and Vining, G. G. (2015). *Introduction to linear regression analysis*. John Wiley & Sons.

Nisbet, R., Miner, G., and Elder IV, J. (2009). *Handbook of statistical analysis and data mining applications*. Academic Press.

Osborne, M. R., Presnell, B., and Turlach, B. A. (2000). A new approach to variable selection in least squares problems. *IMA journal of numerical analysis*, 20(3):389–403.

Tibshirani, R. (1996). Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society. Series B (Methodological)*, pages 267–288.

Tibshirani, R. J. (2011). *The solution path of the generalized lasso*. Stanford University.

Tibshirani, R. J. et al. (2013). The lasso problem and uniqueness. *Electronic Journal of Statistics*, 7:1456–1490.