Reflexiones acerca del modelo de Regresión Lineal

Cristóbal Lecaros C. *CIMT, Universidad de Chile*

El siguiente trabajo examina los métodos de estimación de los paramétros del modelo de regresión lineal y expone algunos instrumentos matemáticos que se implementan en él para permitir trabajar con residuos heterocedásticos y observaciones correlacionadas.

Keywords: Regresión lineal, Mínimos Cuadrados, Máxima Verosimilitud, Función varianza, Estructura de correlación

Introducción

El siguiente ensayo quiere presentar algunas consideraciones sobre aspectos teóricos del modelo de regresión lineal. Mi intención es desarrollar algunos conceptos que aparecen cuando se estudia el modelo y que sirven como base para la implementación de otro modelo estadístico que es el modelo lineal con efectos mixtos. La mayoría de la teoría de este trabajo proviene de las ideas expuestas en Gaecki y Burzykowski [1], aunque también hay ideas importantes que se encuentran en Strang [3, 4], en conjunto con lo visto en clases por Felipe Tobar (link). En la parte inicial se expondrá el modelo en su forma general y cómo la estimación de los coeficientes tiene varias metodologías (e.g., el método de mínimos cuadrados que toma la forma de una matriz de proyección; el método de máxima verosimilitud que asume un error ε que se distribuye normalmente), que bajo ciertas condiciones llegan al mismo resultado. Luego, se revisará cómo se introducen dos objetos matemáticos al modelo, que permiten darle flexibilidad para trabajar con datos heterocedásticos y con observaciones correlacionadas.

Definición

Como hemos visto en clases, el modelo se describe, para el nivel individual de cada dato u observación como:

$$y_i = x_i^{(1)}\beta_1 + \ldots + x_i^{(p)}\beta_p + \varepsilon_i \tag{1}$$

La fórmula para escribir todas las observaciones del modelo de una manera matricial es:

$$\mathbf{y} \equiv \begin{pmatrix} y_1 \\ \vdots \\ y_n \end{pmatrix}, \boldsymbol{\varepsilon} \equiv \begin{pmatrix} \varepsilon_1 \\ \vdots \\ \varepsilon_n \end{pmatrix}$$

$$X = \begin{pmatrix} x_1^{(1)} & x_1^{(2)} & \ldots & x_1^{(p)} \\ \vdots & \vdots & \ddots & \vdots \\ x_n^{(1)} & x_n^{(2)} & \ldots & x_n^{(p)} \end{pmatrix}$$

$$\mathbf{y} = X\boldsymbol{\beta} + \boldsymbol{\varepsilon} \tag{2}$$

Siendo $\varepsilon \sim \mathcal{N}(0, \sigma)$.
Nos interesa estimar $\boldsymbol{\beta}$ de la mejor forma posible.

Mínimos Cuadrados

El método de mínimos cuadrados puede entenderse como la estrategia para encontrar la mejor solución cuando $Ax = b$ *no tiene solución*, es decir, cuando las filas son más que las columnas de la matriz A ($m > n$), o cuando las columnas de A no son independientes. Por lo tanto, si se quiere estimar el mejor $\hat{x}$, tal que $A\hat{x} = b$

$$\begin{aligned} A\hat{x} &= b \quad /A^T \\ A^T A\hat{x} &= A^T b \quad /(A^T A)^{-1} \\ \hat{x} &= (A^T A)^{-1} A^T b \end{aligned}$$

De forma similar, se puede llegar a la interpretación geométrica si se entiende el problema como la proyección del vector b que se encuentra fuera del espacio vectorial de las columnas ($C(A)$), utilizando la matriz Proyección $P = A(A^T A)^{-1} A^T b$, que tiene las propiedades $P = P^T = P^2$. El problema de encontrar el mínimo error $e = b - Ax$ se logra mediante la derivación de la funcion $\|Ax - b\|_2^2$ en el punto en que esta derivada parcial (proveniente de una función cuadrada) es cero. Esta función costo tiene la forma

$$J = \frac{1}{2} \sum_{i=1}^{n} (y_i - a\mathbf{x}_i - b)^2$$

y su optimización para un modelo $y = ax + b$ es

$$a^*, b^* = \arg\min_{a,b} \frac{1}{2} \sum_{i=1}^{n} (y_i - a\mathbf{x}_i - b)^2$$

Puede verse entonces que la estimación de los coeficientes β de (2) tiene la siguiente forma, que es la forma mencionada anteriormente para $\hat{x}$:

$$\widehat{\boldsymbol{\beta}} \equiv \left(\sum_{i=1}^{n} \mathbf{x}_i \mathbf{x}_i^T \right)^{-1} \sum_{i=1}^{n} \mathbf{x}_i y_i = \left(\boldsymbol{X}^T \boldsymbol{X} \right)^{-1} \boldsymbol{X}^T \mathbf{y}$$

Máxima verosimilitud

Una forma interesante de estimar los coeficientes es mediante un enfoque probabilístico. Si se asumen observaciones *independientes* identicamente distribuidas, la estimacion de β puede entenderse como un productorio de múltiples probabilidades. La función likelihood sobre el conjunto de datos se denota como:

$$L\left(\beta, \sigma^2; \mathbf{y}\right) \equiv \left(2\pi\sigma^2\right)^{-n/2} \prod_{i=1}^{n} \exp\left[-\frac{\left(y_i - \mathbf{x}_i^T \boldsymbol{\beta}\right)^2}{2\sigma^2} \right]$$

Como vimos en clase, $\max p(Y|X, \mathbf{w}, \sigma) = \min - \log p(Y|X, \mathbf{w}, \sigma)$, por lo que la estimación puede denotarse como:

$$\ell\left(\beta, \sigma^2; \mathbf{y}\right) \equiv -\frac{n}{2} \log\left(\sigma^2\right) - \frac{1}{2\sigma^2} \sum_{i=1}^{n} \left(y_i - \mathbf{x}_i^T \boldsymbol{\beta}\right)^2$$

Que tiene la misma solución que el método ordinario de mínimos cuadrados.

Función varianza

Definición

Para permitir al modelo trabajar sin la restricción de homocedasticidad, se puede entender el problema mediante una abstracción ingeniosa en que se asume que σ es un escalar e introducir una función varianza $\lambda(\cdot)$, que permita modificar la varianza entre las observaciones.

De ese modo, el modelo queda definido como:

$$y_i = x_i^{(1)}\beta_1 + \cdots + x_i^{(p)}\beta_p + \varepsilon_i \equiv \mathbf{x}_i'\boldsymbol{\beta} + \varepsilon_i$$

siendo

$$\varepsilon_i \sim \mathcal{N}\left(0, \sigma_i^2\right)$$
$$\sigma_i^2 = \sigma^2\lambda_i^2$$

y por lo tanto:

$$\text{Var}\left(\varepsilon_i\right) = \sigma^2\lambda^2\left(\boldsymbol{\delta}, \mu_i; \mathbf{v}_i\right)$$

$\lambda(\cdot)$ tiene tres parámetros: δ, μ y $\mathbf{v}$; asume valores positivos, es continua y diferenciable con respecto al parámetro δ. Es importante notar que el modelo definido así tiene $n + p$ parámetros, incluyendo n parámetros σ_i y p parámetros β. Esto es más que n, el número de observaciones. Esto hace que el modelo se vuelva no identificable[1]. Para solucionar esto, se deben imponer restricciones adicionales a la función varianza que operen sobre los residuos.

Tipos de función varianza

La función varianza $\lambda(\cdot)$ puede ser clasificada en cuatro grupos:

1. Pesos conocidos, $\lambda(\cdot) = \lambda(\mathbf{v})$
2. Funciones varianza que dependen en δ pero no en μ, $\lambda(\cdot) = \lambda(\delta; \mathbf{v})$
3. Funciones varianza que dependen en δ y μ, $\lambda(\cdot) = \lambda(\delta; \mu; \mathbf{v})$
4. Funciones varianza que dependen en μ pero no en δ, $\lambda(\cdot) = \lambda(\mu; \mathbf{v})$

La clasificación tiene implicancias importantes en términos de los métodos de estimación y sus resultados. Por ejemplo, para funciones varianza que no dependen de μ, la distribución del estadístico proveniente del F-test es solo aproximada.

Fórmula general

Si uno quisiera escribir la función varianza para todas las observaciones, el modelo debe ser especificado de la siguiente manera:

$$\boldsymbol{R} \equiv \boldsymbol{\Lambda}\boldsymbol{\Lambda}$$
$$\boldsymbol{\Lambda} = \text{diag}\left(\lambda_1, \ldots, \lambda_n\right)$$

$$\mathbf{y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}; \quad \boldsymbol{\varepsilon} \sim \mathcal{N}(\mathbf{0}, \mathcal{R}), \quad \mathcal{R} = \sigma^2\boldsymbol{R}$$

[1] Una definción de identificabilidad que se encuentra en wikipedia [2] es la siguiente: Sea $\mathcal{P} = \{P_\theta : \theta \in \Theta\}$ un modelo estadístico con el espacio de parámetros Θ. $\mathcal{P}$ es identificable si el mapeo $\theta \mapsto P_\theta$ es uno-a-uno: $P_{\theta_1} = P_{\theta_2} \quad \Rightarrow \theta_1 = \theta_2 \quad$ para todo $\quad \theta_1, \theta_2 \in \Theta$.

Estructura de correlación

El supuesto fundamental del modelo de regresión lineal es que las observaciones son independientes entre sí. Este supuesto es restrictivo cuando se realizan experimentos que entregan datos correlacionados. Por ejemplo, datos que provienen de series de tiempo o en el que existen grupos o clusters dados por condiciones espaciales. Los modelos que relajan el supuesto de independencia se conocen como modelos de efectos fijos y errores residuales correlacionados para datos agrupados. Para conseguir esto, se debe introducir un objeto matemático sobre los modelos presentados anteriormente, y que se conoce como estructura de correlación.
Para datos con diferentes niveles de agrupamiento, tenemos N grupos indexados por i ($i = 1, ..., N$) y n_i observaciones por grupo indexadas por j ($j = 1, ..., n_i$).

Sea el modelo

$$\mathbf{y}_i = \boldsymbol{X}_i\boldsymbol{\beta} + \boldsymbol{\varepsilon}_i$$

donde

$$\mathbf{y}_i \equiv \begin{pmatrix} y_{i1} \\ \vdots \\ y_{ij} \\ \vdots \\ y_{in_i} \end{pmatrix}, \quad \boldsymbol{\varepsilon}_i \equiv \begin{pmatrix} \varepsilon_{i1} \\ \vdots \\ \varepsilon_{ij} \\ \vdots \\ \varepsilon_{in_i} \end{pmatrix},$$

$$\boldsymbol{X}_i \equiv \begin{pmatrix} x_{i1}^{(1)} & x_{i1}^{(2)} & \dots & x_{i1}^{(p)} \\ \vdots & \vdots & \ddots & \vdots \\ x_{in_i}^{(1)} & x_{in_i}^{(2)} & \dots & x_{in_i}^{(p)} \end{pmatrix}$$

$$\varepsilon_i \sim \mathcal{N}_{n_i}\left(\mathbf{0}, \mathcal{R}_i\right)$$

$$\mathcal{R}_i = \sigma^2 \boldsymbol{R}_i.$$

La estructura de correlación se introduce sobre la varianza de los residuos como una matriz de correlación $\boldsymbol{C_i}$, que permite la correlación entre las observaciones dentro del grupo i.

$$\boldsymbol{R}_i = \boldsymbol{\Lambda}_i \boldsymbol{C}_i \boldsymbol{\Lambda}_i$$

La estructura de correlación se especifica asumiendo que los coeficientes de correlación entre dos errores residuales, ε_{ij} y $\varepsilon_{ij'}$, que corresponden a dos observaciones del mismo grupo i, está dada por

$$\text{Corr}\left(\varepsilon_{ij}, \varepsilon_{ij'}\right) = h\left[d\left(\mathbf{t}_{ij}, \mathbf{t}_{ij'}\right), \boldsymbol{\varrho}\right]$$

donde $\boldsymbol{\varrho}$ es un vector de parámetros de correlación, $d\left(\mathbf{t}_{ij}, \mathbf{t}_{ij}\right)$ es una función distancia de los vectores de posición $\mathbf{t}_{ij}$ y $\mathbf{t}_{ij'}$ que corresponden con ε_{ij} y $\varepsilon_{ij'}$, respectivamente, y $h(\cdot, \cdot)$ es una función continua con respecto a ϱ, que toma valores entre -1 y 1, y donde $h(0, \varrho) \equiv 1$. Utilizando diferentes funciones de distancia y de correlación, se pueden obtener diferentes estructuras de correlación, que se clasifican generalmente en dos grupos: seriales y espaciales.

Referencias

[1] Andrzej Gaecki y Tomasz Burzykowski. *Linear Mixed-Effects Models Using R: A Step-by-Step Approach*. Springer Texts in Statistics. New York: Springer-Verlag, 2013. ISBN: 978-1-4614-3899-1. URL: https://www.springer.com/gp/book/9781461438991 (visitado 06-07-2019).

[2] *Identifiability*. En: *Wikipedia*. Page Version ID: 891510783. 8 de abr. de 2019. URL: https://en.wikipedia.org/w/index.php?title=Identifiability&oldid=891510783 (visitado 06-07-2019).

[3] Gilbert Strang. *Introduction to linear algebra*. Fifth edition. OCLC: 964067316. Wellesley, MA: Wellesley-Cambridge Press, 2016. 574 págs. ISBN: 978-0-9802327-7-6.

[4] Gilbert Strang. *Matrix Methods in Data Analysis, Signal Processing, and Machine Learning*. MIT OpenCourseWare. 2018. URL: https://ocw.mit.edu/courses/mathematics/18-065-matrix-methods-in-data-analysis-signal-processing-and-machine-learning-spring-2018/ (visitado 06-07-2019).