

Proyecciones lineales de alta dimensionalidad

Parte II

Fernando Arias-Rodríguez

Banco Central de Bolivia

29 de agosto de 2024



- 1 Procedimientos de selección automática de variables
- 2 Selección de modelos

1 Procedimientos de selección automática de variables

Introducción

Forward Selection Regressions (FWD)

Regresión Ridge

Least Angle Regressions (LARS)

Least Absolute Shrinkage Selection Operator (LASSO)

Elastic Net Estimator (NET)

2 Selección de modelos

1 Procedimientos de selección automática de variables

Introducción

Forward Selection Regressions (FWD)

Regresión Ridge

Least Angle Regressions (LARS)

Least Absolute Shrinkage Selection Operator (LASSO)

Elastic Net Estimator (NET)

2 Selección de modelos

- Se fundamenta en la necesidad de automatizar los procedimientos de selección del mejor modelo posible.
- Una técnica estándar es la metodología *general-to-specific* (GET), donde se encuentran procedimientos como *stepwise*.
- Existen dos alternativas al acercamiento GET, conocidas como reglas de umbrales duros y suaves.
- Una regla de umbrales dura se basa en seleccionar un regresor según la significancia de su coeficiente de correlación con la variable objetivo.
- Reglas duras tienden a seleccionar predictores altamente correlacionados, complicando la estimación.

- Una regla de umbrales suave ordena y selecciona N regresores con base en un problema de minimización, el cual toma la forma:

$$\min_{\beta} \Phi(RSS) + \lambda \Psi(\beta_1, \dots, \beta_j, \dots, \beta_N) \quad (1)$$

con RSS igual a la suma de residuos al cuadrado de la regresión entre la variable objetivo y los N regresores.

- El parámetro λ gobierna la compresión, es decir, un λ más grande implica una penalización más alta por incluir un regresor extra en el modelo; Φ y Ψ son funciones de RSS y los parámetros (β) asociados a los N regresores.
- La correlación cruzada entre los regresores se toma en cuenta cuando se minimiza esta función de pérdida.

Dependiendo de las formas funcionales para Φ y Ψ , diferentes reglas de umbrales se tendrán y, por ende, diferentes procedimientos de selección. En particular, se estudiarán las siguientes:

- *Forward Selection Regressions* (FWD).
- *Least Angle Regressions* (LARS).
- Least Absolute Shrinkage Selection Operator (LASSO).
- *Elastic Net Estimator* (NET).

1 Procedimientos de selección automática de variables

Introducción

Forward Selection Regressions (FWD)

Regresión Ridge

Least Angle Regressions (LARS)

Least Absolute Shrinkage Selection Operator (LASSO)

Elastic Net Estimator (NET)

2 Selección de modelos

- Suponga que se desea pronosticar una variable y a partir de un conjunto de regresores X .
- El primer paso sería identificar el regresor que muestre la mayor correlación con y , digamos x_1 . FWD arranca en este punto, al hacer $y \sim x_1$, extraer los residuales $\hat{\epsilon}_1$ y buscar el regresor con mayor correlación con este término.
- Sea x_2 el regresor con mayor correlación con $\hat{\epsilon}_1$. El siguiente paso es correr $\hat{\epsilon}_1 \sim x_2$, calcular $\hat{\epsilon}_2$ e identificar el regresor con mayor correlación.

- El procedimiento continúa hasta que todas las variables sean ranqueadas o hasta que se satisfaga algún criterio, por ejemplo que el R^2 ajustado en la regresión $y \sim x_1, \dots$ se encuentre por encima de algún umbral.
- La filosofía con respecto a una regla dura es la contraria: aquí se desea mantener a las variables que sean lo más ortogonal posible entre ellas.

1 Procedimientos de selección automática de variables

Introducción

Forward Selection Regressions (FWD)

Regresión Ridge

Least Angle Regressions (LARS)

Least Absolute Shrinkage Selection Operator (LASSO)

Elastic Net Estimator (NET)

2 Selección de modelos

- Objetivo: Mantener todas las variables que se tienen disponibles, pero penalizar sus coeficientes asociados si estos se encuentran demasiado lejos de cero.
- Se desea disminuir la complejidad del modelo sin renunciar a variables.
- Especificación del modelo:

$$L_{ridge}(\beta) = \sum_{i=1}^N ((y_i - x_i' \hat{\beta}))^2 + \lambda \sum_{j=1}^m \hat{\beta}_j^2 \quad (2)$$

- Nótese que las definiciones del sesgo, $-\lambda(X'X + \lambda I)^{-1}\beta$, y la varianza, $\sigma^2(X'X + \lambda I)^{-1}X'X(X'X + \lambda I)^{-1}$, se afectan por λ : mientras más grande sea, menor varianza pero más sesgo.

- La regularización aquí se define en términos de escoger el valor de λ .
- Existen dos formas de hacerlo:
 - ① Validación cruzada, es decir, con pronósticos usando una muestra de evaluación.
 - ② Criterios de información como AIC (*Akaike*) o BIC (*Bayesian Information Criterion*).
 - ③ Precaución: Para hallar los grados de libertad en AIC o BIC, se debe usar la matriz H modificada: $H_{ridge} = (X'X + \lambda I)^{-1}X$ y los grados de libertad serán $df_{ridge} = tr(H_{ridge})$, con $tr()$ igual a la traza de la matriz.

1 Procedimientos de selección automática de variables

Introducción

Forward Selection Regressions (FWD)

Regresión Ridge

Least Angle Regressions (LARS)

Least Absolute Shrinkage Selection Operator (LASSO)

Elastic Net Estimator (NET)

2 Selección de modelos

- Este algoritmo arranca igual que FWD: se identifica el regresor con la mayor correlación con y . Se extraen los residuales $\hat{\epsilon}_1$ y se busca un regresor que tenga la mayor correlación con este.
- En este paso, LARS se aparta de FWD: LARS procede *equiangularmente* entre x_1 y x_2 .
- Lo anterior significa que, a diferencia de FWD, LARS estima una regresión tal que los residuales resultantes tengan *la misma correlación* con x_1 y x_2 .
- Si se repiten los pasos anteriores k veces, se tendrán k regresores con los cuales implementar la regresión lineal. En este caso, el algoritmo termina y los coeficientes de los $N - k$ regresores faltantes son iguales a cero.
- En este caso, k resulta ser el parámetro que rige la regla de optimización del algoritmo.

1 Procedimientos de selección automática de variables

Introducción

Forward Selection Regressions (FWD)

Regresión Ridge

Least Angle Regressions (LARS)

Least Absolute Shrinkage Selection Operator (LASSO)

Elastic Net Estimator (NET)

2 Selección de modelos

- LASSO es un caso particular de LARS, en el que se impone en cada paso del algoritmo una restricción sobre el signo de la correlación entre el nuevo regresor candidato y la proyección hecha en el paso inmediatamente anterior.
- La intuición es: suponga que la correlación entre x_1 & y es positiva. Si se supone que en la búsqueda de x_2 , el signo de la correlación debe ser positivo, se está en una regresión LASSO. Si no importa el signo, se está en LARS.

- LASSO puede relacionarse con el estimador *Ridge*, el cual es una estimación restringida e implementada por M.C.O. que penaliza sobreajuste. Dados M regresores, los coeficientes *Ridge* se obtienen al solucionar el siguiente problema de minimización:

$$\min_{\beta} RSS + \lambda \sum_{j=1}^M \beta_j^2 \quad (3)$$

donde RSS es la suma de residuos al cuadrado. El multiplicador de Lagrange gobierna la contracción: un valor alto de λ significa una mayor penalización por tener un regresor extra en el modelo.

- LASSO introduce una pequeña pero importante modificación en la función expresada en la Ecuación 3:

$$\min_{\beta} RSS + \lambda \sum_{j=1}^M |\beta_j| \quad (4)$$

implicando que, en LASSO, algunos coeficientes de la regresión son exactamente iguales a cero.

- Esta particularidad resulta muy útil cuando se implementan aplicaciones con *big data*.

1 Procedimientos de selección automática de variables

Introducción

Forward Selection Regressions (FWD)

Regresión Ridge

Least Angle Regressions (LARS)

Least Absolute Shrinkage Selection Operator (LASSO)

Elastic Net Estimator (NET)

2 Selección de modelos

- Este estimador es un refinamiento de LASSO, y es el producto de la solución al siguiente problema de minimización:

$$\min_{\beta} RSS + \lambda_1 \sum_{j=1}^M |\beta_j| + \lambda_2 \sum_{j=1}^M \beta_j^2 \quad (5)$$

- En este caso, la contracción depende de dos parámetros, λ_1 y λ_2 . Sin embargo, es posible reformular este problema de tal forma que pueda solucionarse como un modelo LASSO, usando el algoritmo LARS.

- 1 Procedimientos de selección automática de variables
- 2 Selección de modelos

- Los parámetros λ , λ_1 y λ_2 , que controlan el proceso de contracción, se seleccionan mediante *validación cruzada*.
- En un primer paso, se usa una muestra de entrenamiento para:
 - ① estimar modelos para distintos valores de los parámetros;
 - ② computar la función de pérdida;
 - ③ escoger el valor de los parámetros que minimice la pérdida.
- En un segundo paso, se toma una nueva muestra para computar la función de pérdida para los valores de los parámetros que se seleccionaron en el paso anterior y se comprueba que también producen buenos resultados por fuera de la muestra de entrenamiento.

- Empíricamente, no hay un criterio para definir cuál de los métodos tiene un mejor desempeño en términos de pronóstico.
- En aplicaciones empíricas, es posible estimar varios de estos modelos con una muestra de entrenamiento, compararlos entre sí y el mejor es el que se usará para realizar pronóstico.
- Este acercamiento es otro ejemplo de validación cruzada.