RIDGE y LASSO

Jorge Gallego

Facultad de Economía, Universidad del Rosario

Junio 24 de 2017

Introducción

- El modelo lineal es bastante utilizado y tiene grandes ventajas
- Sin embargo, en términos de predicción, puede tener algunos problemas
- En ocasiones puede ser bastante impreciso
- Además, su interpretación puede ser difícil, en especial cuando hay muchos predictores
- Los modelos RIDGE y LASSO son extensiones del modelo lineal que atienden estos problemas

Bajo Poder Predictivo

- Cuando $n > p$, siendo n el número de obs. y p el de variables, las estimaciones OLS suelen tener baja varianza
- Pero si n no es mucho más grande que p, la varianza aumenta, lo cual reduce la precisión de las predicciones
- De hecho, si $n < p$, no hay solución única OLS y el método no sirve
- Por ende, restringir o reducir el número de coeficientes estimados ayuda a mejorar la precisión de las predicciones
- El costo es que puede aumentar un poco el sesgo

Interpretabilidad del Modelo

- Muchas veces algunas variables incluidas en modelo no tienen correlación con el resultado
- Todo lo cual dificulta su interpretación
- Si se remueven dichas variables se hace más fácil interpretar el modelo
- Pero es muy poco probable que OLS genere coeficientes exactamente iguales a 0

Regularización

- Los métodos de regularización sirven para resolver estos problemas
- Buscan "encoger" los coeficientes OLS hacia cero para las variables que lo merecen
- Dependiendo del método, los coeficientes se vuelven 0 exactamente, o se acercan a 0
- De esta forma, estos métodos también sirven para hacer selección de variables
- Y esto sirve para mejorar la predicción de los modelos

Regresión RIDGE

- Sabemos que el método OLS busca el vector β que minimice la suma residual de cuadrados:

$$RSS = \sum_{i=1}^{n} \left(y_i - \beta_0 - \sum_{j=1}^{p} \beta_j x_{ij} \right)^2$$

- La regresión RIDGE es muy similar. Pero su función objetivo es ligeramente diferente

- Se busca el vector $\hat{\beta}^R$ que minimice

$$\sum_{i=1}^{n} \left(y_i - \beta_0 - \sum_{j=1}^{p} \beta_j x_{ij} \right)^2 + \lambda \sum_{j=1}^{p} \beta_j^2 = RSS + \lambda \sum_{j=1}^{p} \beta_j^2$$

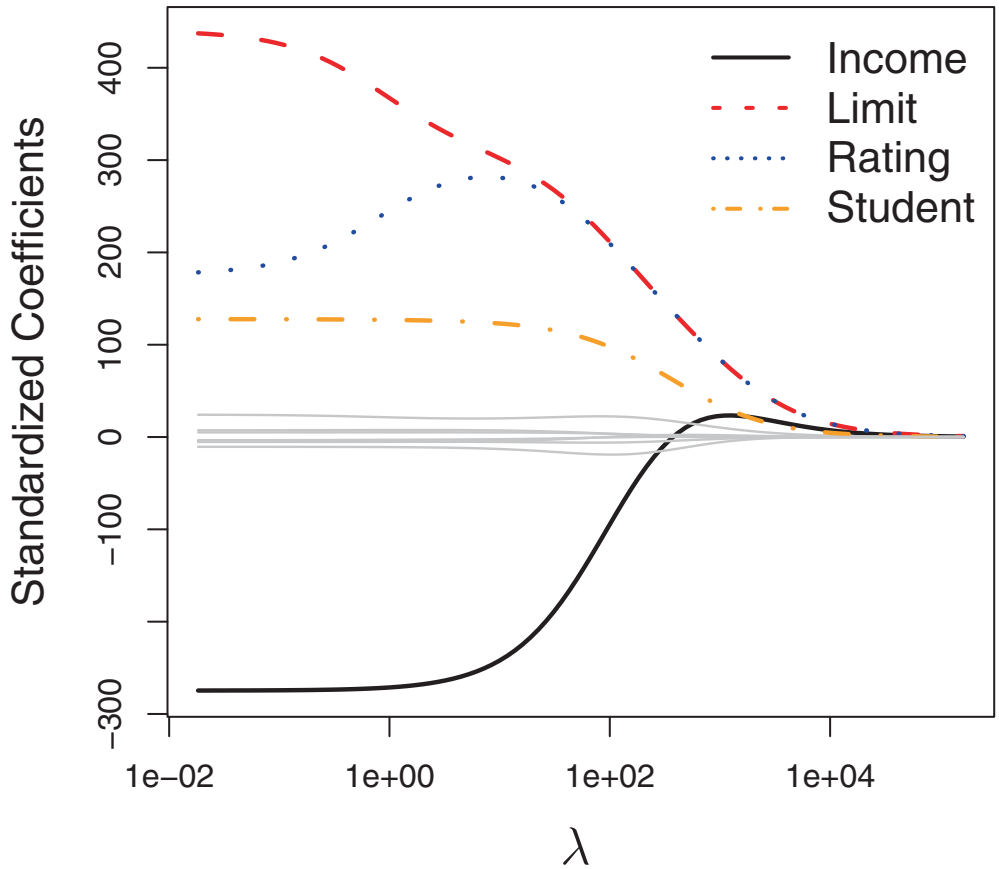
Regresión RIDGE

- Donde $\lambda \geq 0$ es un parámetro de sintonización que se determina separadamente
- Esta regresión cumple dos propósitos. También busca un buen ajuste para los datos al minimizar RSS
- Pero el término $\lambda \sum_{j=1}^{p} \beta_j^2$ es el *shrinkage penalty*. Hala los coeficientes hacia 0
- Si $\lambda = 0$, OLS y RIDGE coinciden. Si $\lambda \to \infty$, los coeficientes tienden a 0

Regresión RIDGE

- Los coeficientes RIDGE dependen de λ
- Para elegir este parámetro se usan métodos de *cross-validation*
- El punto es que cuánto más grande sea λ, más se encogen los coeficientes
- Veamos esto con un ejemplo de un modelo donde predecimos el nivel de endeudamiento en tarjetas de crédito

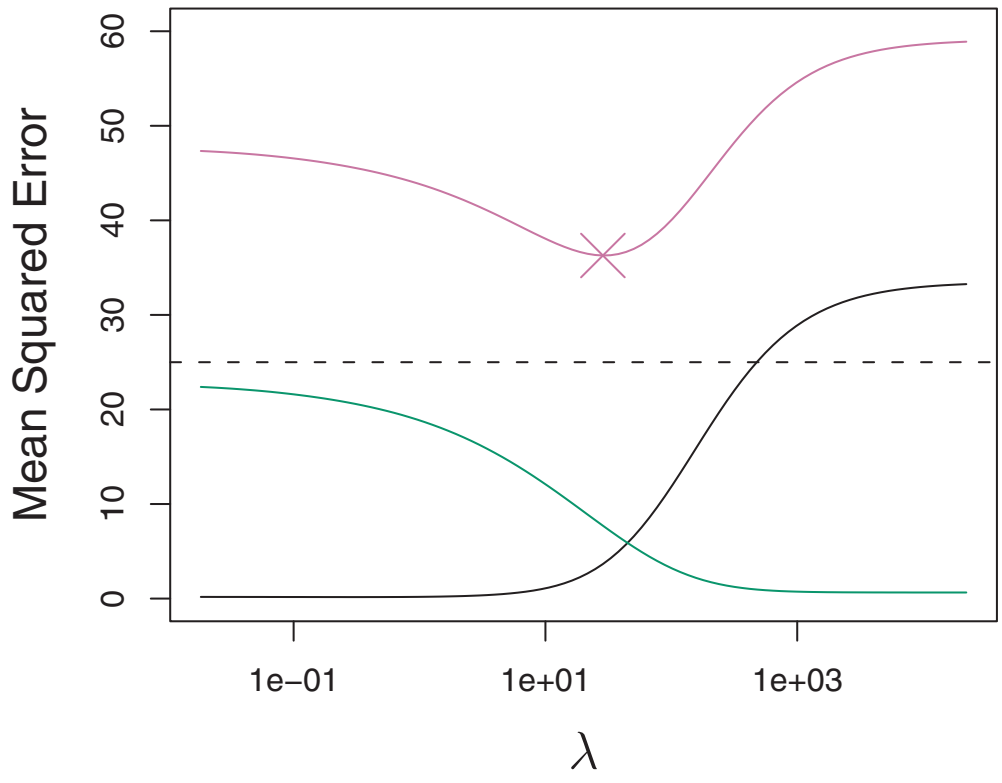
Regresión RIDGE



RIDGE vs. OLS

- ¿Por qué puede representar RIDGE una mejora respecto a OLS?
- La respuesta la da el *trade-off sesgo-varianza*
- Al aumentar λ se reduce la varianza de las predicciones pero aumenta el sesgo
- Cuando $\lambda = 0$ los coeficientes son insesgados pero la varianza es mayor
- Veamos esto gráficamente con datos simulados

RIDGE vs. OLS



RIDGE vs. OLS

- La curva verde es la varianza de las predicciones
- La curva negra es el sesgo de la estimaciones
- La morada el *mean squared error*: una función de la varianza más el sesgo al cuadrado
- El trade-off es claro. De hecho, hay un punto óptimo para el MSE

RIDGE vs. OLS

- En general, cuando la relación entre outcome y predictores tiende a ser lineal, OLS tiene bajo sesgo pero alta varianza
- Lo cual implica que cambios en los datos de entrenamiento conducen a resultados muy distintos
- Esto es grave cuando p es grande respecto a n
- RIDGE ayuda bastante en estos escenarios

LASSO

- Un problema de RIDGE es que los coeficientes no son exactamente iguales a 0
- Luego ninguna variables es descartada del todo a no ser que $\lambda = \infty$
- Esto puede no afectar la precisión predictiva. Pero si hace difícil interpretar el modelo cuando hay muchas variables
- LASSO es una alternativa a RIDGE que resuelve este problema
- Su especificación funcional permite descartar algunas variables que no son relevantes en el modelo

LASSO

- Los coeficientes LASSO, $\hat{\beta}^L$, minimizan la función

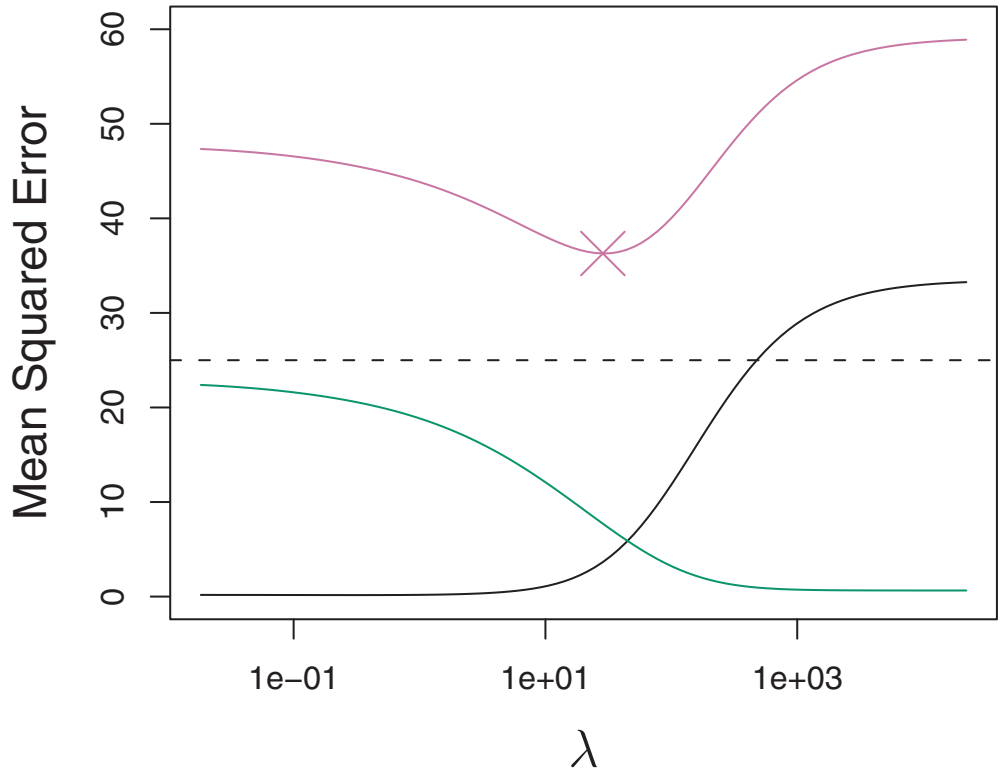
$$\sum_{i=1}^{n}\left(y_i - \beta_0 - \sum_{j=1}^{p}\beta_j x_{ij}\right)^2 + \lambda\sum_{j=1}^{p} \mid \beta_j \mid = RSS + \lambda\sum_{j=1}^{p} \mid \beta_j \mid$$

- La diferencia con RIDGE está en que se penaliza con $\mid \beta_j \mid$ en lugar de β_j^2
- Al igual que RIDGE, hala los coeficientes hacia 0
- Pero a diferencia de RIDGE, algunos de estos *sí* pueden ser 0

LASSO

- De esta forma, los modelos LASSO son más fáciles de interpretar, es especial en presencia de muchas variables
- Es un modelo de selección de variables
- Como en RIDGE, la selección de λ es crucial
- Se hace con técnicas de *cross-validation*

LASSO



LASSO

- Cuando $\lambda = 0$, los coeficientes LASSO son los mismos OLS
- A medida que crece, van desapareciendo algunos coeficientes
- A diferencia de RIDGE, para ciertos valores de λ algunas variables no entran en el modelo

¿LASSO o RIDGE?

- ¿Cuál de los dos modelos tiene mayor poder predictivo?
- En principio, ninguno domina al otro
- En general, si el predictor depende de un número reducido de variables, LASSO es mejor que RIDGE
- En cambio si el predictor depende de muchas características, RIDGE será mejor que LASSO
- La técnica de *cross-validation* sirve para determinar qué modelo es mejor de usar para ciertos datos

Selección de λ (abrebocas)

- La técnica para seleccionar λ la veremos la próxima clase
- Pero la intuición es simple.
- Se selecciona un conjunto de valores de λ para poner a prueba
- Para cada uno se calcula el error de predicción asociado al modelo generado por dicho λ
- Se escoge el λ que genere el mejor error
- La clave está en cómo se construye dicho error. Lo veremos la próxima clase

Extensiones de RIDGE y LASSO

- Los modelos RIDGE y LASSO pueden extenderse a modelos no-lineales
- Por ejemplo a Logit, Probit, Multinomial Logit, Poisson, etc.
- La penalidad se introduce en la función de log-verosimilitud que se optimiza en cada caso
- El efecto es el mismo: encoger, e incluso anular, algunos de los coeficientes