PONTIFICIA
UNIVERSIDAD
CATÓLICA
DEL PERÚ

Selección de Variables y Regularización

Profesor: Alex de la Cruz H.

MODELOS LINEALES 1

EST631

Contenido

Criterios de comparación de modelos

En el contexto de regresión lineal, los criterios de comparación de modelos permiten evaluar cuál modelo se ajusta mejor a los datos observados y, al mismo tiempo, tiene una estructura adecuada (ni demasiado compleja ni demasiado simple).

Sea el modelo lineal general:

$$\boldsymbol{y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}, \quad \mathbb{E}[\boldsymbol{\varepsilon}] = \mathbf{0}, \quad \mathrm{Var}(\boldsymbol{\varepsilon}) = \boldsymbol{\Sigma}$$

donde $\boldsymbol{y} \in \mathbb{R}^n$, $\boldsymbol{X} \in \mathbb{R}^{n \times p}$, $\boldsymbol{\beta} \in \mathbb{R}^p$ y $\boldsymbol{\varepsilon} \sim \mathcal{N}_n(\mathbf{0}, \boldsymbol{\Sigma})$.

Coeficiente de determinación

Indica la proporción de variabilidad de la variable respuesta explicada por el modelo.

$$R^2 = 1 - \frac{\text{RSS}}{\text{TSS}}$$

- RSS (residuos): $\text{RSS} = (\boldsymbol{y} - \boldsymbol{X}\hat{\boldsymbol{\beta}})^\top \boldsymbol{W}(\boldsymbol{y} - \boldsymbol{X}\hat{\boldsymbol{\beta}})$, donde $\boldsymbol{W} = \boldsymbol{\Sigma}^{-1}$
- TSS (total): $\text{TSS} = (\boldsymbol{y} - \bar{y}\mathbf{1})^\top \boldsymbol{W}(\boldsymbol{y} - \bar{y}\mathbf{1})$

donde $\bar{y}$ es el promedio ponderado de $\boldsymbol{y}$, definido como:

$$\bar{y} = \frac{\mathbf{1}^\top \boldsymbol{W}\boldsymbol{y}}{\mathbf{1}^\top \boldsymbol{W}\mathbf{1}}$$

Coeficiente de determinación ajustado

Ajusta el R^2 considerando la cantidad de predictores en el modelo.

$$R^2_{\text{ajustado}} = 1 - \left(\frac{n-1}{n-p}\right)(1-R^2)$$

p es el número de coeficientes de regresión. Por lo que penaliza el número de predictores en el modelo. Más útil que R^2 cuando se comparan modelos con diferente número de variables.

AIC y BIC

Criterio de información de Akaike (AIC), penaliza la complejidad del modelo:

$$\text{AIC} = -2 \log L(\hat{\boldsymbol{\theta}}) + 2k$$

donde:

$$\ell(\hat{\boldsymbol{\theta}}) = \log L(\hat{\boldsymbol{\theta}}) = -\frac{n}{2} \log(2\pi) - \frac{1}{2} \log |\boldsymbol{\Sigma}| - \frac{1}{2} (\boldsymbol{y} - \boldsymbol{X}\hat{\boldsymbol{\beta}})^\top \boldsymbol{\Sigma}^{-1} (\boldsymbol{y} - \boldsymbol{X}\hat{\boldsymbol{\beta}})$$

y k es el número de parámetros del modelo.

AIC y BIC

De manera similar, el criterio de información bayesiano (BIC), penaliza la complejidad del modelo:

$$\text{BIC} = -2 \log L(\hat{\boldsymbol{\theta}}) + k \log(n)$$

Penaliza más fuertemente la complejidad del modelo que el AIC. Útil para comparar modelos con diferentes cantidades de predictores.

Criterio de Mallows (C_p)

$$C_p = \frac{1}{n}\left(RSS + 2p\hat{\sigma}^2\right)$$

- C_p es un estimador insesgado del MSE de prueba, si $\hat{\sigma}^2$ es un estimador insesgado de σ^2.
- Un C_p menor es mejor.

Error cuadrático medio (MSE)

En forma generalizada:

$$\text{MSE}_\Sigma = \frac{1}{n}(\boldsymbol{y} - \boldsymbol{X}\hat{\boldsymbol{\beta}})^\top \boldsymbol{\Sigma}^{-1}(\boldsymbol{y} - \boldsymbol{X}\hat{\boldsymbol{\beta}})$$

Validación cruzada (cross-validation)

- k-fold: Se divide la muestra en k partes, se entrena el modelo en $k-1$ y se evalúa en la restante.
- Evalúa capacidad de generalización del modelo.

Prueba F global

Evalúa si al menos uno de los predictores explica significativamente la variable respuesta.

$$H_0 : \boldsymbol{\beta}_1 = \boldsymbol{\beta}_2 = \cdots = \boldsymbol{\beta}_k = 0 \quad \text{vs} \quad H_1 : \text{al menos un } \beta_j \neq 0$$

El estadístico es:

$$F = \frac{(\text{TSS} - \text{RSS})/(p-1)}{\text{RSS}/(n-p)} \sim F_{k,n-p} \quad \text{bajo } H_0$$

Prueba t para coeficientes

Evalúan si un predictor específico tiene efecto significativo sobre la variable respuesta.

$$H_0 : \beta_j = 0 \quad \text{vs} \quad H_1 : \beta_j \neq 0$$

El estadístico es:

$$t_j = \frac{\hat{\beta}_j}{\sqrt{\text{Var}(\hat{\beta}_j)}} \sim t_{n-p} \quad \text{bajo } H_0$$

Donde

$$\text{Var}(\hat{\boldsymbol{\beta}}) = (\boldsymbol{X}^\top \boldsymbol{\Sigma}^{-1} \boldsymbol{X})^{-1}$$

Criterio	**¿Qué evalúa?**	**¿Preferencia?**
R^2, R^2_{ajustado}	Proporción explicada	Más alto mejor
AIC / BIC	Ajuste penalizado por complejidad	Más bajo mejor
MSE / Validación cruzada	Error de predicción	Más bajo mejor
Pruebas t y F	Significancia estadística	Valores p bajos

Selección de Variables y Regularización

En el análisis de regresión lineal, la relación entre el número de observaciones disponibles y la cantidad de variables predictoras juega un papel crucial en la calidad de los resultados.

Sea un modelo lineal clásico:

$$\boldsymbol{y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}, \quad \mathbb{E}[\boldsymbol{\varepsilon}] = \mathbf{0}, \quad \mathrm{Var}(\boldsymbol{\varepsilon}) = \boldsymbol{\sigma^2 I}$$

donde $\boldsymbol{y} \in \mathbb{R}^n$, $\boldsymbol{X} \in \mathbb{R}^{n \times p}$, $\boldsymbol{\beta} \in \mathbb{R}^p$ y $\boldsymbol{\varepsilon} \sim \mathcal{N}_n(\mathbf{0}, \boldsymbol{\sigma^2 I})$.

- MCO minimiza RSS (residuos): $\text{RSS} = (\boldsymbol{y} - \boldsymbol{X}\hat{\boldsymbol{\beta}})^\top(\boldsymbol{y} - \boldsymbol{X}\hat{\boldsymbol{\beta}})$
- En MCO o EMV se resuelve $\boldsymbol{X}^\top\boldsymbol{X}\hat{\boldsymbol{\beta}} = \boldsymbol{X}^\top\boldsymbol{y}$, esto es $\hat{\boldsymbol{\beta}} = \left(\boldsymbol{X}^\top\boldsymbol{X}\right)^{-1}\boldsymbol{X}^\top\boldsymbol{Y}$

Sin embargo, si $n < p$, el sistema de ecuaciones normales:

$$\boldsymbol{X}^\top\boldsymbol{X}\hat{\boldsymbol{\beta}} = \boldsymbol{X}^\top\boldsymbol{y}$$

no tiene solución única (el sistema es indeterminado), lo que provoca que las estimaciones tengan alta varianza o que ni siquiera se pueda estimar el modelo.

Suponiendo que la verdadera relación es aproximadamente lineal:

- $n \gg p$: varianza de las estimaciones de los coeficientes tienden a tener varianza pequeña.
- n no mucho mayor que p: varianza de las estimaciones de los coeficientes tienden a tener varianza grande.
- $n < p$: muchas soluciones posibles (modelo no identificado), varianza infinita, el modelo no puede ser utilizado.

Las técnicas de regularización y selección de variables, ayudan a mejorar el rendimiento del modelo en condiciones problemáticas como cuando n no es mucho mayor que p o $n < p$.

Por ello, se han desarrollado dos enfoques principales:

- Selección de subconjuntos
- Regularización o contracción

Selección de subconjuntos

Consiste en identificar un subconjunto de p predictores que creemos que están más relacionados con la respuesta.

Algunas alternativas son:

- Selección del mejor subconjunto
- Selección de modelos paso a paso (stepwise)

Selección de subconjuntos

Número total de modelos considerados:

$$\binom{p}{1} + \binom{p}{2} + \ldots + \binom{p}{p} = 2^p - 1$$

Incluyendo también el modelo nulo, se obtienen 2^p modelos posibles.

Selección del mejor subconjunto

1. Sea $\mathcal{M}_0$ el modelo nulo (sin predictor). Este modelo simplemente predice la media muestral para todas las observaciones.
2. Para $k = 1, 2, \ldots, p$:
   1) Ajustar todos los $\binom{p}{k}$ modelos que contienen exactamente k predictores.
   2) Elegir el mejor entre estos $\binom{p}{k}$ modelos, y llamarlo $\mathcal{M}_k$. Se considera "mejor" al que tenga el menor RSS, o equivalentemente, el mayor R^2.
3. Seleccionar un único modelo entre $\mathcal{M}_0, \mathcal{M}_1, \ldots, \mathcal{M}_p$, utilizando uno de los siguientes criterios:
   - Error de predicción estimado mediante validación cruzada, C_p, AIC, BIC o R^2 ajustado.

Selección del mejor subconjunto

- En general, el número de modelos a considerar se incrementa demasiado a medida que aumenta p
- $p = 10$ conduce a aprox. $2^{10} - 1 = 1024 - 1$ posibilidades
- $p = 20$ conduce a más de 1 millón de posibilidades
- Un gran espacio de búsqueda puede provocar un sobreajuste a los datos de entrenamiento.

Selección paso a paso hacia adelante (Forward Stepwise)

- Se inicia con un modelo sin predictores, $\mathcal{M}_0$.
- Se añaden predictores al modelo, uno a la vez, seleccionando en cada paso el predictor que mejora más el ajuste.
- El proceso continúa hasta incluir todos los predictores, generando la secuencia:

$$\mathcal{M}_0, \mathcal{M}_1, \mathcal{M}_2, \ldots, \mathcal{M}_p$$

- Finalmente, se selecciona el mejor modelo entre los $\mathcal{M}_k$, usando algún criterio de evaluación.

Selección paso a paso hacia adelante (Forward Stepwise)

Algoritmo 1: Forward Stepwise

1. Sea $\mathcal{M}_0$ el modelo nulo, que no contiene ningún predictor.
2. Para $k = 0, \ldots, p-1$:
   1) Considerar todos los $p - k$ modelos que se obtienen agregando un predictor adicional a los que ya contiene $\mathcal{M}_k$.
   2) Seleccionar el mejor entre estos modelos, y llamarlo $\mathcal{M}_{k+1}$. El "mejor´´ se define como el que tiene el menor RSS o el mayor R^2.
3. Seleccionar el mejor modelo entre $\mathcal{M}_0, \mathcal{M}_1, \ldots, \mathcal{M}_p$ utilizando validación cruzada, C_p, AIC, BIC o R^2 ajustado.

Selección paso a paso hacia adelante (Forward Stepwise)

Algunas observaciones importantes:

- Este procedimiento reduce el número de modelos ajustados de 2^p a:

$$1 + \sum_{k=0}^{p-1}(p-k) = 1 + \frac{p(p+1)}{2}$$

- En cada paso se elige el predictor más prometedor.
- No se garantiza encontrar el mejor modelo posible con un subconjunto de los p predictores.

Selección paso a paso hacia adelante (Forward Stepwise)

- A diferencia de la selección por subconjuntos, la selección hacia adelante puede aplicarse incluso cuando $n < p$ (alta dimensión).
- Si se limita el algoritmo a considerar solo los modelos con hasta $n-1$ predictores, se tiene la secuencia:

$$\mathcal{M}_0, \mathcal{M}_1, \ldots, \mathcal{M}_{n-1}$$

Selección paso a paso hacia atrás (Backward Stepwise)

Algoritmo 2: Backward Stepwise

1. Sea $\mathcal{M}_p$ el modelo completo, que contiene todos los p predictores.
2. Para $k = p, p-1, \ldots, 1$:
    1) Considere todos los modelos con k predictores que contienen todos menos uno de los predictores en $\mathcal{M}_k$, para un total de $k-1$ predictores.
    2) Elija el *mejor* entre estos k modelos y llámelo $\mathcal{M}_{k-1}$. Aquí, el *mejor* se define como el que tiene el menor RSS o el mayor R^2.
3. Seleccione un único mejor modelo entre $\mathcal{M}_0, \ldots, \mathcal{M}_p$ usando el error de predicción validado por validación cruzada, C_p, AIC, BIC o R^2 ajustado.

Selección paso a paso hacia atrás (Backward Stepwise)

Propiedades similares al algoritmo de selección hacia adelante (Forward)

- Buscar $1 + \frac{p(p+1)}{2}$ en lugar de 2^p modelos.
- Es una búsqueda guiada, no elegimos $1 + \frac{p(p+1)}{2}$ modelos para considerarlos al azar.
- No se garantiza que produzca el mejor modelo que contenga un subconjunto de p predictores.
- Sin embargo, la selección hacia atrás requiere que el número de observaciones n sea mayor que el número de variables p Para que se pueda ajustar el modelo completo.

Ejemplo de selección de subconjuntos

'

Aplicación

Para ilustrar, usaremos el conjunto de datos `Hitters` que proviene del paquete `ISLR` en R y contiene información sobre jugadores profesionales de béisbol de las Grandes Ligas (MLB).

- Muestra: 322 jugadores (tras eliminar valores perdidos: 263)
- Variables: 20 variables (1 variable respuesta y otros 19 predictores)
- Variable respuesta: salario del jugador en miles de dólares (`Salary`).

Contexto: Desempeño y características de jugadores en una temporada reciente, con su salario como variable de interés.

Desarrollo en clase

Regularización

Introducción

En lugar de seleccionar un subconjunto, los métodos de contracción mantienen todos los predictores pero **reducen (contraen)** sus coeficientes hacia cero, lo cual reduce la varianza de las estimaciones.

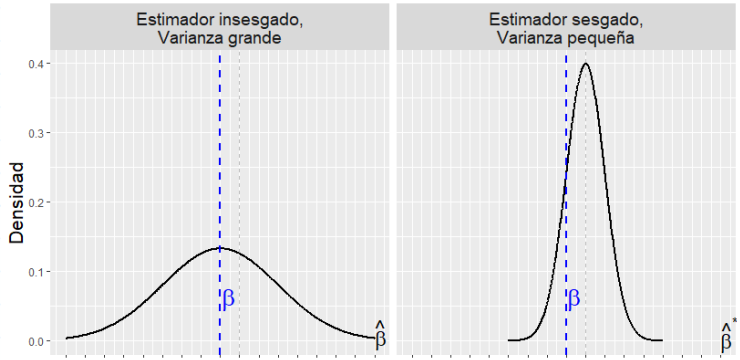
Características:

- Se penalizan los coeficientes grandes, forzándolos a ser más pequeños que los estimados por mínimos cuadrados.
- El objetivo es mejorar la generalización del modelo.

Introducción

- El problema con el MCO es el requisito de que $E(\hat{\beta}) = \beta$.
- La propiedad de Gauss-Markov asegura que el estimador de mínimos cuadrados tiene varianza mínima dentro de la clase de los estimadores lineales insesgados, pero no hay garantía de que esa varianza sea pequeña.
- Si la varianza de $\hat{\beta}$ es grande:
  - los intervalos de confianza para β serán amplios, y
  - el estimador puntual de β será muy inestable.

Introducción



Supóngase que se puede determinar un estimador sesgado de β, $\hat{\beta}^{\star}$, que tenga menor varianza que el estimador insesgado $\hat{\beta}$.

Introducción

Definición 1.

Sea $\hat{\boldsymbol{\theta}}$ un estimador de un parámetro vectorial $\boldsymbol{\theta} \in \mathbb{R}^p$. El Error Cuadrático Medio (ECM) de $\hat{\boldsymbol{\theta}}$ se define como

$$\text{ECM}(\hat{\boldsymbol{\theta}}) = \mathbb{E}\left[\|\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}\|^2\right]$$

En el caso univariado, cuando $\theta \in \mathbb{R}$, esta definición se reduce a:

$$\text{ECM}(\hat{\theta}) = \mathbb{E}\left[(\hat{\theta} - \theta)^2\right]$$

$$\text{ECM}(\hat{\theta}) = \text{Var}(\hat{\theta}) + \left(\mathbb{E}[\hat{\theta}] - \theta\right)^2$$

Introducción

- Si se permite un pequeño sesgo en $\hat{\beta}^{\star}$, es posible reducir su varianza, lo que puede dar lugar a un ECM menor que el del estimador insesgado $\hat{\beta}$.
- Esto se traduce en estimadores más estables y en intervalos de confianza más angostos.

Regularización

Consiste en ajustar un modelo que contenga todos los predictores p

- utilizando una técnica que restringe (o regulariza) las estimaciones de coeficientes
- o de manera equivalente, que reduce las estimaciones de coeficientes a cero.

Reducir el número de parámetros efectivos

- Manteniendo la capacidad de captar los aspectos más interesantes del problema.

Regularización

Las dos técnicas más conocidas para reducir los coeficientes de regresión a cero son:

- la regresión Ridge.
- la regresión LASSO.

Uno de los procedimientos para obtener estimadores sesgados de coeficientes de regresión es la *regresión ridge (o de cresta)*, propuesta originalmente por Hoerl y Kennard (1970).

Definición 2.

Sea el modelo lineal $y_i = \boldsymbol{x}_i^\top \boldsymbol{\beta} + \varepsilon_i$. El estimador de ridge se define como la solución de

$$\hat{\boldsymbol{\beta}}^{\text{ridge}} = \arg\min_{\boldsymbol{\beta}} \left\{ \sum_{i=1}^{n} (y_i - \boldsymbol{x}_i^\top \boldsymbol{\beta})^2 + \lambda \sum_{j=1}^{p} \beta_j^2 \right\}$$

que en notación matricial equivale a:

$$\hat{\boldsymbol{\beta}}^{\text{ridge}} = \arg\min_{\boldsymbol{\beta}} \left\{ (\boldsymbol{y} - X\boldsymbol{\beta})^\top (\boldsymbol{y} - X\boldsymbol{\beta}) + \lambda \boldsymbol{\beta}^\top \boldsymbol{\beta} \right\}$$

donde $\lambda \geq 0$ es conocido como el parámetro de regularización o penalización.

λ equilibra el compromiso entre:

- el ajuste al modelo (primer término): $(\boldsymbol{y} - X\boldsymbol{\beta})^\top(\boldsymbol{y} - X\boldsymbol{\beta})$, que mide el error cuadrático.
- la complejidad del modelo (segundo término): $\boldsymbol{\beta}^\top\boldsymbol{\beta} = \|\boldsymbol{\beta}\|_2^2$, que penaliza los coeficientes grandes.

Interpretación según el valor de λ:

- Si $\lambda = 0$: se recupera método de mínimos cuadrados ordinarios (MCO).
- Si $\lambda \to \infty$: los coeficientes $\boldsymbol{\beta}$ tienden a cero, es decir, el modelo se vuelve muy simple (alta regularización).
- Para $\lambda > 0$: se introduce sesgo a cambio de reducir la varianza del estimador, lo que puede mejorar la capacidad de generalización del modelo.

Estimación

La solución del estimador de *ridge* se obtiene a partir del siguiente problema de optimización:

$$\frac{\partial}{\partial \boldsymbol{\beta}} \left[(\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta})^\top (\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta}) + \lambda \boldsymbol{\beta}^\top \boldsymbol{\beta} \right] = 0$$

Luego:

$$\hat{\boldsymbol{\beta}}^{\text{ridge}} = (\boldsymbol{X}^\top \boldsymbol{X} + \lambda \boldsymbol{I})^{-1} \boldsymbol{X}^\top \boldsymbol{y}$$

Propiedades

- **Esperanza**: el estimador ridge es sesgado, pues

$$\mathbb{E}(\hat{\boldsymbol{\beta}}^{\text{ridge}}) = (\boldsymbol{X}^\top X + \lambda \boldsymbol{I})^{-1} \boldsymbol{X}^\top \boldsymbol{X} \boldsymbol{\beta}$$

  y el sesgo del estimador es: $\text{Bias} = \left[(\boldsymbol{X}^\top \boldsymbol{X} + \lambda \boldsymbol{I})^{-1} \boldsymbol{X}^\top \boldsymbol{X} - \boldsymbol{I}\right] \boldsymbol{\beta}$

- **Varianza**

$$\text{Var}(\hat{\boldsymbol{\beta}}^{\text{ridge}}) = \sigma^2 (\boldsymbol{X}^\top \boldsymbol{X} + \lambda I)^{-1} \boldsymbol{X}^\top \boldsymbol{X} (\boldsymbol{X}^\top \boldsymbol{X} + \lambda \boldsymbol{I})^{-1}$$

Ejercicio: demostrar

Propiedades

Error cuadrático medio (ECM)

$$\begin{aligned}\text{ECM}(\hat{\boldsymbol{\beta}}^{\text{ridge}}) &= \text{tr}\left[\sigma^2(\boldsymbol{X}^\top\boldsymbol{X}+\lambda\boldsymbol{I})^{-1}\boldsymbol{X}^\top\boldsymbol{X}(\boldsymbol{X}^\top\boldsymbol{X}+\lambda\boldsymbol{I})^{-1}\right] \\ &\quad + \left\|\left[(\boldsymbol{X}^\top\boldsymbol{X}+\lambda\boldsymbol{I})^{-1}\boldsymbol{X}^\top\boldsymbol{X}-\boldsymbol{I}\right]\boldsymbol{\beta}\right\|^2 \\ &= \sigma^2\sum_{j=1}^{p}\frac{\lambda_j}{(\lambda_j+\lambda)^2} + \lambda^2\boldsymbol{\beta}^\top(\boldsymbol{X}^\top\boldsymbol{X}+\lambda\boldsymbol{I})^{-2}\boldsymbol{\beta}\end{aligned}$$

en donde $\lambda_1, \lambda_2, \ldots, \lambda_p$ son los autovalores de $\boldsymbol{X}^\top\boldsymbol{X}$.

Observaciones

- Si $\lambda \to \infty$, los estimadores de los coeficientes se aproximan a cero sin que estos desaparezcan del modelo.
- La penalización no se aplica al intercepto, β_0.
- En general, *ridge* produce predicciones más precisas que los modelos obtenidos por MCO + selección "clásica" de variables, a menos que el verdadero modelo sea ralo o "esparso" (mayoría de coeficientes nulos).

Observaciones

- Aunque una mayor penalización contrae los coeficientes estimados hacia cero, ninguno de ellos llega a ser exactamente cero; por tanto, no se produce selección de variables. Todas las variables originales permanecen en el modelo final.
- Algunos autores, Hoerl, Kennard y Baldwin (1975), sugieren que una elección adecuada de λ es

$$\lambda = \frac{p\hat{\sigma}^2}{\hat{\boldsymbol{\beta}}^\top \hat{\boldsymbol{\beta}}}$$

donde $\hat{\boldsymbol{\beta}}$ y $\hat{\sigma}^2$ es de MCO y p número de covariables.

Elección por validación cruzada

Sea la predicción que hacemos de la observación y_i cuando empleamos el estimador ridge de parámetros λ obtenido con una muestra de la que excluimos la observación i-ésima. Definamos

$$CV(\lambda) = \sum_{i=1}^{N} (y_i - \hat{y}_{(i),\lambda})^2$$

Entonces,

$$\lambda_{CV} = \arg\min_{k} CV(\lambda)$$

Elección por validación cruzada

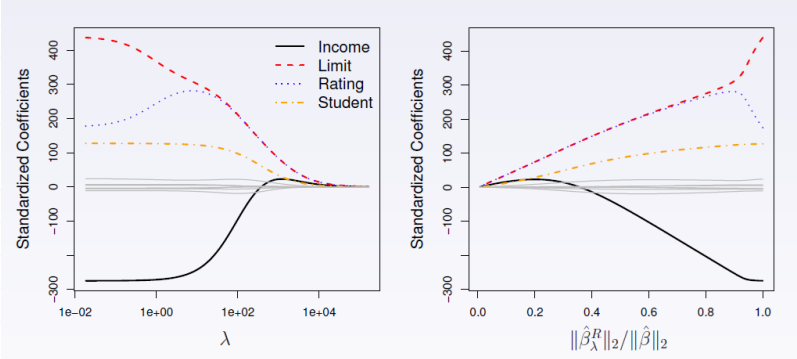


Figura 1: Ejemplo *Credit*: Los coeficientes de regresión Ridge estandarizados se muestran para el conjunto de datos de crédito.

Eficacia: ¿Por qué funciona?

- A medida que aumenta λ, la flexibilidad del ajuste disminuye.
- Esto conduce a una disminución de la varianza, pero a un mayor sesgo.
- El *ECM* es una función tanto de la varianza como del sesgo al cuadrado: Es necesario encontrar un punto óptimo.
- Las estimaciones por mínimos cuadrados presentan alta varianza (muchos p en comparación con n).

Eficacia: ¿Por qué funciona?

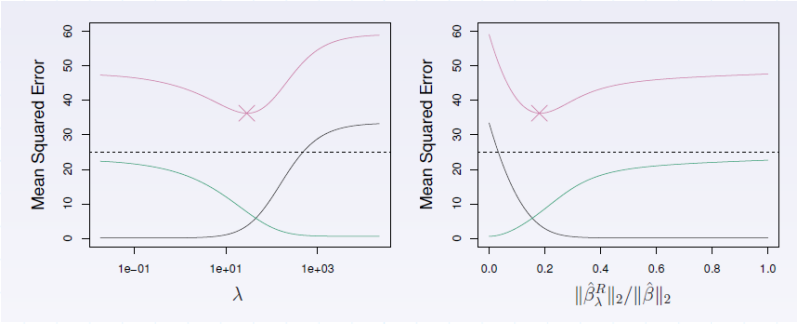


Figura 2: Regresión Ridge (MSE).

Desventajas de la regresión Ridge

A diferencia de los métodos anteriores, la regresión ridge incluirá todos los predictores p en el modelo final.

- La penalización λ reducirá todos los coeficientes hacia cero.
- Pero no establecerá ninguno de ellos exactamente en cero (a menos que $\lambda = \infty$).

Esto puede no ser un problema para la precisión de la predicción, pero dificulta la interpretación del modelo para p grande.

- También motivado por el objetivo de encontrar una técnica de regresión lineal que fuera estable pero que realizara selección de variables, Tibshirani (1996) propuso Lasso (Least Absolute Shrinkage and Selection Operator).
- El método LASSO aproxima los estimadores de los parámetros a cero, en ocasiones haciéndolos exactamente igual a cero (cosa que no ocurre en regresión ridge), lo que es equivalente a excluir el regresor correspondiente del modelo.

Definición 3.

Sea el modelo lineal $y_i = \boldsymbol{x}_i^\top \boldsymbol{\beta} + \varepsilon_i$. El estimador de Lasso (Least Absolute Shrinkage and Selection Operator) se define como la solución de

$$\hat{\boldsymbol{\beta}}^{\text{lasso}} = \arg\min_{\boldsymbol{\beta}} \left\{ \sum_{i=1}^{n} (y_i - \boldsymbol{x}_i^\top \boldsymbol{\beta})^2 + \lambda \sum_{j=1}^{p} |\beta_j| \right\}$$

donde λ es un parámetro de precisión. Si $\lambda = 0$, regresión lineal tradicional ($\hat{\boldsymbol{\beta}}^{lasso} = \hat{\boldsymbol{\beta}}$).

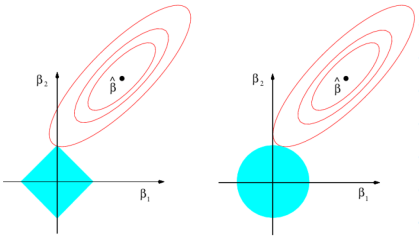
Formulaciones alternativas

- **Ridge**

$$\min_{\boldsymbol{\beta}} \left\{ \sum_{i=1}^{n} \left(y_i - \beta_0 - \sum_{j=1}^{p} \beta_j x_{ij} \right)^2 \right\} \quad \text{sujeto a} \quad \sum_{j=1}^{p} \beta_j^2 \leq s$$

- **Lasso**

$$\min_{\boldsymbol{\beta}} \left\{ \sum_{i=1}^{n} \left(y_i - \beta_0 - \sum_{j=1}^{p} \beta_j x_{ij} \right)^2 \right\} \quad \text{sujeto a} \quad \sum_{j=1}^{p} |\beta_j| \leq s$$



Las elipses rojas son los contornos de la RSS.

Las zonas azules sólidas son las regiones con restricciones, $|\beta_1|+|\beta_2| \leq s$ y $\beta_1^2+\beta_2^2 \leq s$.

La explicación es válida para $p > 2$, pero difícil de visualizar.

Ejemplo de aplicación de regularización

Aplicación

Para ilustrar, usaremos el mismo conjunto de datos `Hitters` del `ISLR` en R y contiene información sobre jugadores profesionales de béisbol de las Grandes Ligas (MLB).

- Muestra: 322 jugadores (tras eliminar valores perdidos: 263)
- Variables: 20 variables (1 variable respuesta y otros 19 predictores)
- Variable respuesta: salario del jugador en miles de dólares (`Salary`).

Contexto: Desempeño y características de jugadores en una temporada reciente, con su salario como variable de interés.

Desarrollo en clase