# EST-25134: Aprendizaje Estadístico

**Profesor**: Alfredo Garbuno Iñigo — Primavera, 2022 — Métodos de selección.
**Objetivo**: Detalles en selección de variables. *Shrinkage* y penalización. Reducción de dimensiones.
**Lectura recomendada**: Referencia.

## 1. INTRODUCCIÓN

Como hemos visto es natural `extender` el modelo lineal

$$Y = \beta_0 + \beta_1 X_1 + \cdots + \beta_p X_p + \epsilon\,. \tag{1}$$

Veremos (mas adelante) la idea de incorporar relaciones `no lineales` manteniendo el supuesto de `aditividad`.

Incluso aunque el modelo lineal es sencillo, tiene sus ventajas pues nos ayuda a tener un modelo `interpretable` y al mismo tiempo con buena `capacidad predictiva`.

El libro de Kuhn and Johnson [2] tiene una buena discusión sobre las ventajas algorítmicas de un modelo lineal. Usualmente en la práctica queremos tener nuestro modelo en un ambiente productivo. Lo cual necesita que las predicciones sean fácilmente calculables. ¿Qué pasaría si en la plataforma de Netflix o Amazon se tarda mucho en aparecer las sugerencias? Los modelos lineales son fácilmente calculables en prácticamente cualquier ambiente productivo.

Estudiaremos estrategias para mejorar modelos lineales a través de procedimientos alternativos de ajuste.

### 1.1. Opciones para ajustar modelos

- Basados en `precisión de ajuste`, ideal cuando $p > n$ con el objetivo de *reducir* varianza.
- Basados en `interpretabilidad`. Por ejemplo, eliminar variables que no tengan capacidad predictiva.

## 2. ESTRATEGIAS DE SELECCIÓN DE VARIABLES

- Selección por subconjuntos.
- Reducción de coeficientes (regularización).
- Reducción de dimensiones.

### 2.1. Selección por subconjuntos

1. Utilizar el modelo nulo $\mathcal{M}_0$ (sin predictores).
2. Para $k = 1, \ldots, p$:
   *a*) Ajustar todos modelos posibles con $k$ predictores.
   *b*) Elegir el *mejor* de esa colección de modelos, le pondremos $\mathcal{M}_k$.
3. Elegir el mejor modelo dentro de la colección $\mathcal{M}_0, \ldots, \mathcal{M}_p$ utilizando un criterio de comparación de modelos.

#### *2.1.1. Para pensar:* ¿Por qué no puedes utilizar el criterio de `RSS`?

### 2.2. En el contexto de clasificación

La `devianza` –el negativo de dos veces la log-verosimilitud– se utiliza como una métrica de bondad de ajuste (como el `RSS`) para una clase mas amplia de modelos.

### 2.3. Selección iterativa

Podemos elegir empezar con el modelo mas sencillo e ir incorporando una variable a la vez mas predictores. En cada paso podemos evaluar la `mejora adicional` de haber incorporado estas nuevas características.

### 2.4. Pseudo-código (selección hacia adelante)

1. Denotamos por $\mathcal{M}_0$ el modelo `nulo`.
2. Para $k = 1, \ldots, p-1$:
   a) Considera todos los $p-k$ modelos que aumentan el modelo en la iteración anterior $\mathcal{M}_k$ con un predictor adicional.
   b) Escoge el `mejor` de estos $p-k$ modelos y llámale $\mathcal{M}_{k+1}$ .
3. Escoge el mejor de los modelos entre $\mathcal{M}_0, \ldots, \mathcal{M}_p$ utilizando un criterio de comparación de modelos.

### 2.5. Aplicación: créditos

```
   Income Limit Rating Cards Age Education Gender Student Married Balance
1      23  3476    257     2  50        15 Female      No      No     209
2      46  6637    491     4  42        14   Male      No     Yes    1046
3      21  1448    145     2  58        13 Female      No     Yes       0
4     181 11966    832     2  58         8 Female      No     Yes    1405
5     122  7818    584     4  50         6   Male      No     Yes     701
```

LISTING 1. *Muestra de datos del connjunto* `Credit`.

El objetivo es predecir `Saldo` en términos utilizando las demás características. El ejemplo de [1] ha implementado la búsqueda por subconjuntos y la búsqueda iterativa hacia adelante. Estos son los mejores modelos encontrados.

| # Variables | Best subset | Forward stepwise |
|---|---|---|
| One | `rating` | `rating` |
| Two | `rating`, `income` | `rating`, `income` |
| Three | `rating`, `income`, `student` | `rating`, `income`, `student` |
| Four | `cards`, `income` `student`, `limit` | `rating`, `income`, `student`, `limit` |

Nota que el mecanismo iterativo no tiene garantía de encontrar el mejor modelo dentro de las $\binom{p}{k}$ posibilidades.

```
   estrategia sigma r.squared adj.r.squared  AIC deviance
1 subconjunto   100      0.95          0.95 4823  3915058
2    adelante   101      0.95          0.95 4835  4032502
```

LISTING 2. *Métricas de bondad de ajuste para los datos de* `Credit`.

*2.5.1. Para pensar:* ¿Cuántos modelos en total se ajustan con el procedimiento de búsqueda iterativa hacia adelante? Considera $p = 20$.

### 2.6. Selección iterativa hacia atrás

Empezamos con el modelo completo que contenga los $p$ predictores. Eliminando variables, una a la vez, cuando un predictor no sea tan útil. La única restricción que necesitamos es que $n > p$ .

## 3. MÉTRICAS DE DESEMPEÑO

Si utilizáramos el `RSS` para comparar entre $\mathcal{M}_0, \ldots, \mathcal{M}_k$ tendríamos un problema pues eliminar predictores siempre perjudicaría la capacidad predictiva del modelo. Necesitamos `compensar` por el sesgo de sobre-ajuste. Es decir, considerar una métrica que pueda estimar el error de `generalización`.

### 3.1. $C_p$ de Mallow

Es un criterio de bondad de ajuste (`menor mejor`) definida como

$$C_p(\mathcal{M}_d) = \frac{1}{n}\left(\mathsf{RSS}(d) + 2d\hat{\sigma}^2\right) . \tag{2}$$

> Tenemos una penalización a la suma de residuales al cuadrado (`RSS`) que considera un aumento en predictores utilizados.

### 3.2. El criterio de información de Akaike (AIC)

Se utiliza para evaluar modelos ajustados por máxima verosimilitud (`menor mejor`)

$$\mathsf{AIC}(\mathcal{M}_d) = -2\log L + 2d\,. \tag{3}$$

*3.2.1. Ejercicio:* Prueba que en el caso del modelo lineal con errores Gaussianos el criterio de mínimos cuadrados y máxima verosimilitud es el mismo. Además los criterios $C_p$ y $\mathsf{AIC}$ son lo mismo.

### 3.3. $R^2$ ajustada

Se calcula como

$$R^2_A(\mathcal{M}_d) = 1 - \frac{\mathsf{RSS}/(n-d-1)}{\mathsf{TSS}/(n-1)}\,. \tag{4}$$

Es una métrica de correlación entre predicción ($\hat{y}$) y respuesta ($y$) (`mayor mejor`). Al contrario de la $R^2$ tradicional esta métrica si se afecta por la inclusión de variables inecesarias/redundantes.

### 3.4. Validación cruzada

Cada uno de los procedimientos de selección de variables regresa una secuencia de modelos $\mathcal{M}_k$. Lo que queremos es escoger la $k^\star$ de acuerdo al error de generalización. El error de generalización estimado tiene la ventaja de no hacer la estimación de $\sigma^2$.

*3.4.1. Selección de modelo: Datos de crédito* El objetivo es predecir el `Saldo` en términos de los demás predictores.

Escogemos el modelo con el error mas pequeño. Sin embargo, validación cruzada nos puede dar una métrica de incertidumbre (¿cuál?). ¿Y si el problema de decisión lo planteamos como una prueba de hipótesis?

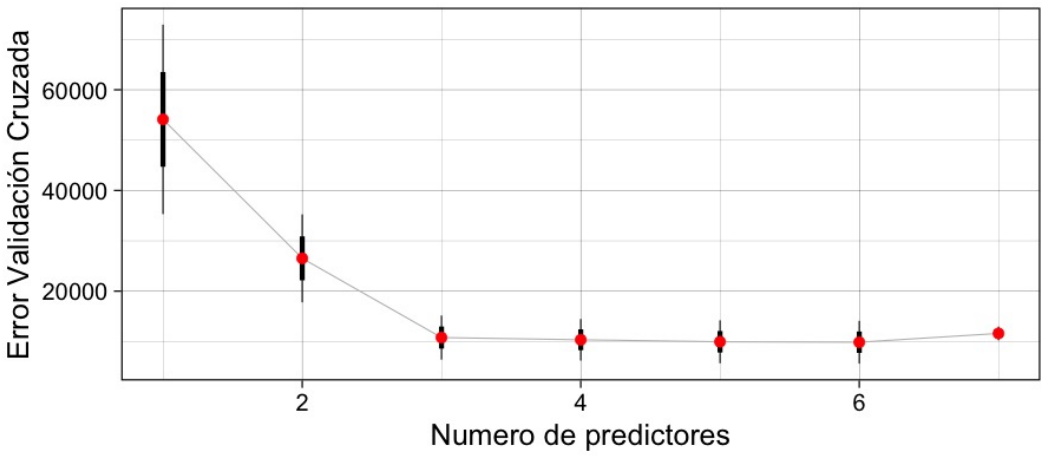


FIGURA 1. *Error de generalización estimado por validación cruzada con $K = 10$. Para los datos de* `Credit`.

## 4. REGULARIZACIÓN

Los procedimientos selección de variables discretos/iterativos pueden generar una varianza muy alta en las estimaciones del error y podría no reducir el error de predicción del modelo completo. Estudiaremos dos métodos de regularización, `Ridge` y `LASSO`, donde ajustamos un modelo con todas las características *penalizando* de alguna manera la complejidad del modelo.

### 4.1. Regresión Ridge

Nuestra formulación anterior consideraba encontrar $\beta_0, \beta_1, \ldots, \beta_n$ minimizando

$$\text{RSS} = \sum_{i=1}^{n} \left( y_i - \beta_0 - \sum_{j=1}^{p} \beta_j x_j \right)^2 . \tag{5}$$

Lo que haremos ahora será incorporar un `término de penalización` en la función objetivo

$$\text{RSS} + \lambda \sum_{j=1}^{p} \beta_j^2 , \tag{6}$$

donde $\lambda \geq 0$ es un `hiper-parámetro`.

El objetivo sigue siendo el mismo, ajustar el modelo lo mejor posible. El término adicional favorece soluciones con $\beta_1, \ldots, \beta_p$ pequeños. El parámetro $\lambda$ controla qué tanto penalizamos el *tamaño* de los coeficientes.

*4.1.1. Para pensar:* Un valor muy pequeño para $\lambda$ implica una penalización `pequeña`, por lo tanto la solución tenderá a ser un modelo `altamente flexible`. Por otro lado un valor de $\lambda$ grande implica una penalización `fuerte`. Esto se traduce en un solución `poco flexible`.

### 4.2. Ridge: datos de crédito

Al penalizar sobre los coeficientes necesitamos que todos *platiquen* en el mismo idioma. Es por esto que tenemos que estandarizar los predictores. Si queremos estimar el error de generalización métodos de separación de muestras, ¿en qué momento lo hacemos? Es decir, ¿antes de separar los datos o en cada paso del proceso de ajuste?

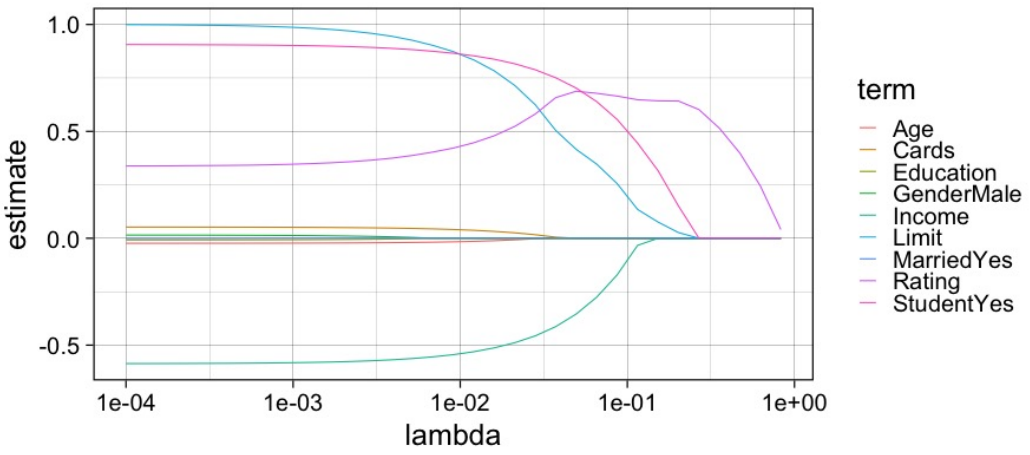


FIGURA 2. *Trayectorias de los coeficientes al aumentar la penalización* $\lambda$.

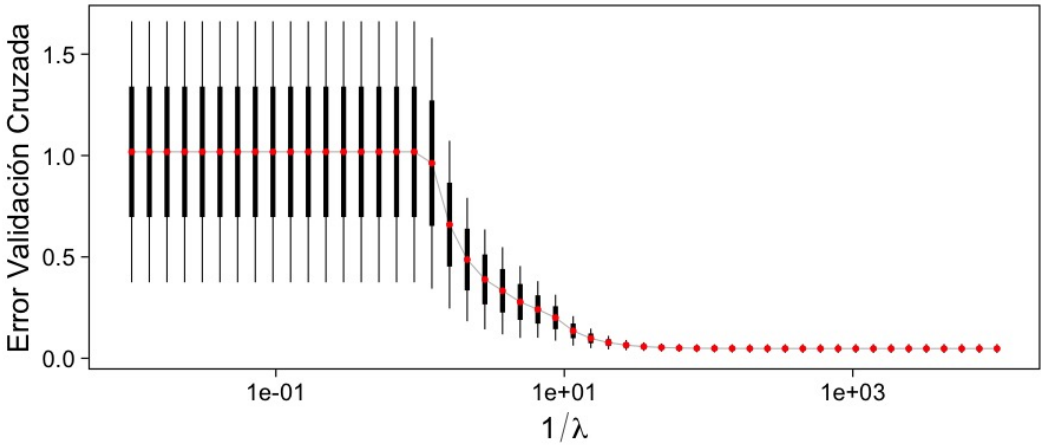


FIGURA 3. *Error de validación calculada con* $K = 10$. *Nota que graficamos contra* $1/\lambda$.

## 4.3. Regresión LASSO

## 4.4. Least angle regression

Ruta completa de LASSO en un pasada. En cada paso identifica la mejor variable para incluir y después ajusta la solución de mínimos cuadrados. Pero sólo deja entrar una fracción de esa variable. Encuentra la variable con mayor correlación con la respuesta. Cambia el coeficiente de la variable continuamente hasta que otra variable *alcance* una correlación con los residuales.

## 4.5. Partial least squares

## REFERENCIAS

[1] G. James, D. Witten, T. Hastie, and R. Tibshirani. *An Introduction to Statistical Learning: With Applications in R.* Springer Texts in Statistics. Springer US, New York, NY, 2021. ISBN 978-1-07-161417-4 978-1-07-161418-1. . 2

[2] M. Kuhn and K. Johnson. *Applied Predictive Modeling.* Springer New York, New York, NY, 2013. ISBN 978-1-4614-6848-6 978-1-4614-6849-3. . 1