# Regularización de modelos lineales y aditivos

## Ridge, Lasso y Redes Elásticas

Gabriel Martos Venturini
gmartos@utdt.edu

Universidad Torcuato Di Tella

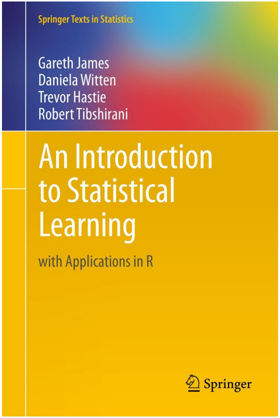
UNIVERSIDAD
TORCUATO DI TELLA

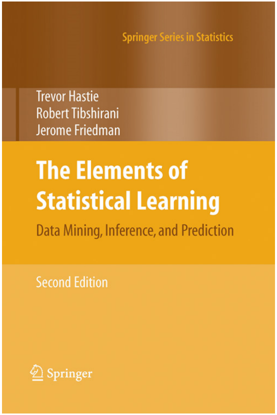
# Agenda

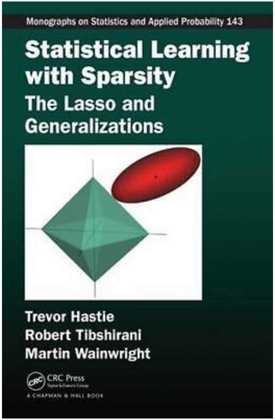
Regularización de modelos lineales y aditivos

Apéndice

# Bibliografía recomendada




**ISL**: 6 y 7.1–7.5.
**ESL**: 3.1–3.4 y 5.1–5.5.
**SLS**:1 y 2.

# Agenda

- Cuando $p \gg 0$ (típico contexto con modelos aditivos) los modelos de regresión suelen presentar mucha *variabilidad*[1].
  - Esto se traduce en que ante perturbaciones en los datos de entrenamiento se producen cambios considerables en los $p \gg 0$ parámetros estimados (aprendidos) del modelo en cuestión.
  - Error Esperado Modelo $=$ Bias$^2$ $+$ **Variance** $+$ cte.
- Para intentar controlar este fenómeno vamos a introducir una *restricción de presupuesto* sobre los parámetros del modelo de forma tal de controlar la variabilidad de forma explícita.
  - Reformular el problema de minimización del riesgo empírico.
- Será necesario estimar (aprender) un hiperparámetro adicional ($\lambda$) que controla el trade–off entre sesgo y variabilidad.
- Omitimos la referencia a los modelos aditivos para simplificar.

[1] *The curse of dimensionality* ver apéndice en § 26.

# Notación

- La discusión vale para modelos aditivos en general:

$$Y = \beta_0 + \sum_{b=1}^{B_1} \beta_{1b}\phi_b(X_1) + \sum_{b=1}^{B_2} \beta_{2b}\phi_b(X_2) + \cdots + \sum_{b=1}^{B_p} \beta_{pb}\phi_b(X_p) + \varepsilon$$

- Pero para simplificar la notación, siempre se plantea que:

$$Y = \beta_0 + \beta_1 X_1 + \cdots + \beta_p X_p + \varepsilon.$$

- Notar que se trata simplemente de renombrar los features:
  - $X_1 \equiv \phi_1(X_1)$ y $\beta_1 \equiv \beta_{11}$.
  - $X_2 \equiv \phi_2(X_1)$ y $\beta_2 \equiv \beta_{12}$.
  - ...
  - $X_p \equiv \phi_{B_p}(X_p)$ y $\beta_p \equiv \beta_{pB_p}$ (notar que generalmente $p \gg 0$).

# Ridge

- Asumimos que $Y = \underbrace{\beta_0 + \beta_1 X_1 + \cdots + \beta_p X_p}_{\text{Modelo para } f(X)} + \varepsilon$.

- Restricción que involucra a las pendientes: $\sum_{i=1}^{p} \beta_i^2 \leq \tau^2$.
  - Pedimos que $(\beta_1, \ldots, \beta_p) \in B(\mathbf{0}, \tau)$.

- Dada una muestra de train $\{(\mathbf{x}_1, y_1), \ldots, (\mathbf{x}_n, y_n)\}$, los parámetros del modelo $\widehat{\boldsymbol{\beta}}_\tau^{\text{ridge}} \equiv (\widehat{\beta}_0, \widehat{\beta}_1, \ldots, \widehat{\beta}_p)_\tau^{\text{ridge}}$ se aprenden resolviendo:

$$\min_{b_0, b_1, \ldots, b_p} \underbrace{\sum_{i=1}^{n} \Big( y_i - (b_0 + b_1 x_{1,i} + \cdots + b_p x_{p,i}) \Big)^2}_{\text{RSS(Datos}, b_0, b_1, \ldots, b_p)} \quad \text{s.a.} \qquad \underbrace{\sum_{j=1}^{p} b_j^2 \leq \tau^2}_{\text{Restricción de presupuesto}}$$

- Veamos una ilustración (cuando $p = 2$) para tener intuición sobre las consecuencias de hacer variar $\tau$ sobre el fitting del modelo.

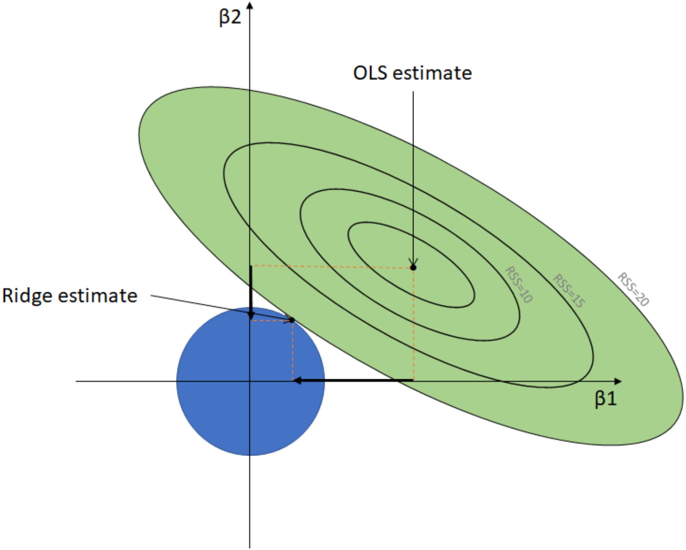



Figure: $\min_{b_1,b_2} \text{RSS}(\text{Datos}, b_1, b_2)$ s.a. $(b_1^2 + b_2^2) \leq \tau^2$.

# Recapitulemos

- $\downarrow \tau \rightarrow$ modelo con menos varianza (y más sesgo).
  - Parámetros *recortados* (shrink).
- Trabajamos con features *estandarizados*.
- De manera compacta (y equivalente*) definimos:

$$\widehat{\beta}_{\lambda}^{\text{ridge}} = \underset{b_1,\ldots,b_p}{\text{argmin}} \quad \underbrace{\text{RSS}(b_1,\ldots,b_p)}_{\text{Bias}} + \lambda \underbrace{||\mathbf{b}||_2^2}_{\text{Variance}}$$

donde $\mathbf{b} = (b_1, \ldots, b_p)$ y $||\mathbf{b}||_2 = \sqrt{\sum_{i=1}^{p} b_i^2}$ (norma $\ell_2$).

  - $||\mathbf{b}||_2$ mide la complejidad del modelo.
  - $\downarrow \tau$ equivale a $\uparrow \lambda$.
- $\lambda$ distribuye peso entre términos (**validación–cruzada**).

# *(Back Up slide)

- Para todo $\tau$ que tiene asociado:

$$\widehat{\boldsymbol{\beta}}_{\tau}^{\text{ridge}} = \underset{b_1,\ldots,b_p}{\text{argmin}} \ \text{RSS}(b_1,\ldots,b_p), \ \text{s.a.} \ ||\mathbf{b}||_2^2 \leq \tau^2,$$

- Existe un solo valor de $\lambda$ que verifica que $\widehat{\boldsymbol{\beta}}_{\tau}^{\text{ridge}} = \widehat{\boldsymbol{\beta}}_{\lambda}^{\text{ridge}}$, con:

$$\widehat{\boldsymbol{\beta}}_{\lambda}^{\text{ridge}} = \underset{b_1,\ldots,b_p}{\text{argmin}} \ \text{RSS}(b_1,\ldots,b_p) + \lambda||\mathbf{b}||_2^2.$$

- Dualidad de Lagrange

# Algunos detalles sobre el estimador Ridge

$$\min_{\mathbf{b}} \underbrace{(\mathbf{y}-\mathbf{X}\mathbf{b})^T(\mathbf{y}-\mathbf{X}\mathbf{b})}_{\text{RSS(Datos,}\mathbf{b})} + \lambda \underbrace{\mathbf{b}^T\mathbf{b}}_{\|\mathbf{b}\|_2^2}$$

- Estimador Ridge:

$$\widehat{\boldsymbol{\beta}}_\lambda^{\text{ridge}} = (\mathbf{X}^T\mathbf{X} + \lambda\mathbf{I}_p)^{-1}\mathbf{X}^T\mathbf{y}$$

- Programación cuadrática (escalable con $n$ y $p$).
- En el caso de diseños ortonormales: $\widehat{\boldsymbol{\beta}}_\lambda^{\text{ridge}} = \widehat{\boldsymbol{\beta}}^{\text{ols}}/(1+\lambda)$.
  - Por tanto: $\widehat{\boldsymbol{\beta}}_\lambda^{\text{ridge}} \xrightarrow{\lambda\to 0} \widehat{\boldsymbol{\beta}}^{\text{ols}}$ y $\widehat{\boldsymbol{\beta}}_\lambda^{\text{ridge}} \xrightarrow{\lambda\to\infty} \mathbf{0}$.
- Cuando $\lambda \to \infty$, la solución Rigde restringe todos los parámetros estimados a cero (salvo la ordenada al origen).

# Implementación en R

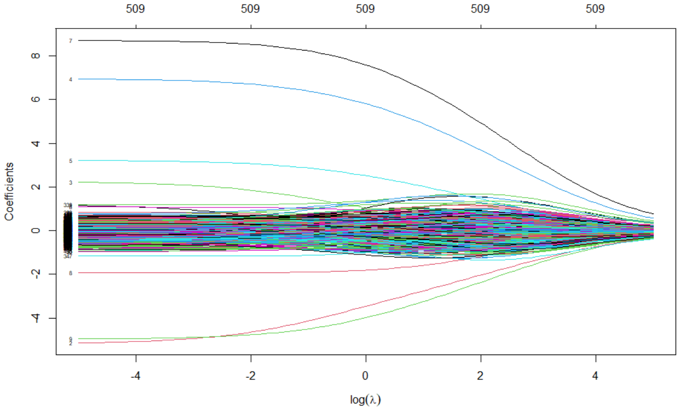



Figure: `Simulaciones`. Las curvas representan los valores estimaciones de los parámetros del modelo Ridge para distintos valores de $\log(\lambda)$.

# Comentarios

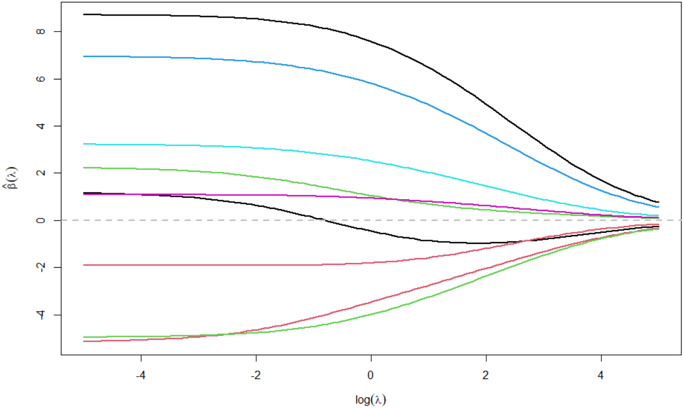



Figure: Sólo las variables relevantes en el modelo. Algunos parámetros estimados cambian de signo al variar $\lambda$ (inconsistencia).

- Con $\lambda \gg 0$ algunos $\widehat{\beta} \approx 0$ (no hacemos selección de modelo).
- Los parámetros estimados cambian de signo :(

# Lasso (Least Absolute Shrinkage and Selection Operator)

- Dada $\{(\mathbf{x}_1, y_1), \ldots, (\mathbf{x}_n, y_n)\}$ y $\tau \geq 0$, Lasso resuelve:

$$(\widehat{\beta}_1, \ldots, \widehat{\beta}_p)^{\text{lasso}}_{\tau} = \underset{b_1,\ldots,b_p}{\text{argmin}} \quad \sum_{i=1}^{n} \Big( y_i - (b_1 x_{1,i} + \cdots + b_p x_{p,i}) \Big)^2$$

$$\text{subject to} \quad \sum_{j=1}^{p} |b_j| \leq \tau$$

- Variables *estandarizadas*.

De manera equivalente y compacta:

$$\widehat{\boldsymbol{\beta}}^{\text{lasso}}_{\lambda} = \underset{b_1,\ldots,b_p}{\text{argmin}} \quad \text{RSS}(b_1, \ldots, b_p) + \lambda ||\mathbf{b}||_1$$

donde $\mathbf{b} = (b_1, \ldots, b_p)$ y $||\mathbf{b}||_1 = \sum_{i=1}^{p} |b_i|$ (norma $\ell_1$).

# Geometría de los problemas de optimización

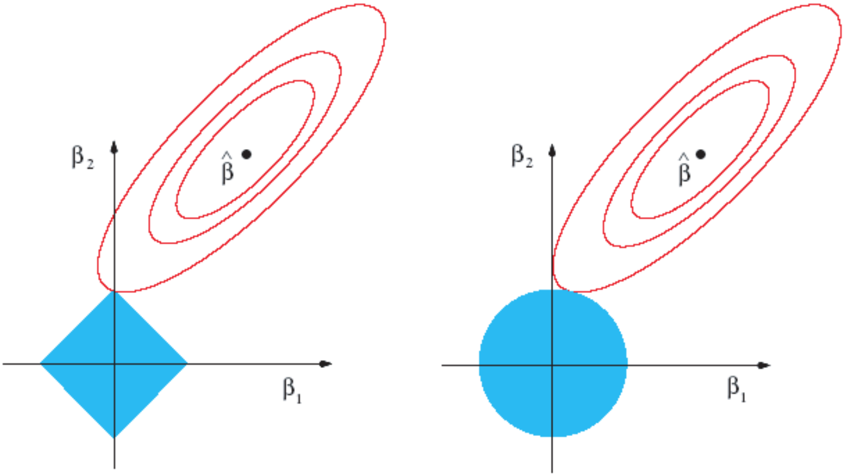



Figure: Lasso ($|\beta_1| + |\beta_2| \leq \tau$) vs Ridge ($\beta_1^2 + \beta_2^2 \leq \tau^2$). Cuando $p \gg 0$, la soluciones de esquina con Lasso son más frecuentes. ((run–giff))

- Problema de optimización convexo escalable en $n$ y $p$.

- Relación con OLS en diseños ortonormales:

$$\widehat{\beta}_{\lambda,j}^{\text{lasso}} = \text{sign}(\widehat{\beta}_j^{\text{ols}})\max(|\widehat{\beta}_j^{\text{ols}}| - \lambda, 0) \text{ para } j = 1, \ldots, p.$$

- Cuando $\lambda \to 0$, Lasso se aproxima a OLS.

- Lasso hace **selección de modelos**: Cuando $|\widehat{\beta}_j^{\text{ols}}| \leq \lambda$ *desaparece* el feature/covariable $j$-ésimo del modelo.

- A medida que $\lambda \to \infty$, la solución de Lasso consiste en descartar todos las features (solo ordenada al origen).

- Coherencia: Lasso provee estimaciones que no cambian de signo ante cambios en $\lambda$ (Ridge no cumple siempre esta prop).

# Conexiones con los modelos Bayesianos (simplificado)

- T. Park and G. Casella: *The Bayesian Lasso*. JASA, 2012.

$$\begin{aligned} \mathbf{y}|\mathbf{X}\boldsymbol{\beta}, \lambda, \sigma^2 &\sim N(\mathbf{X}\boldsymbol{\beta}, \sigma^2\mathbf{I}_n) \\ \boldsymbol{\beta}|\lambda, \sigma^2 &\sim \Pi_{j=1}^{p} \frac{\lambda}{2\sigma} e^{-\frac{\lambda}{\sigma}|\beta_j|} \text{ (iid Laplace).} \end{aligned}$$

- Llamamos $\pi(\boldsymbol{\beta}|\pi(\boldsymbol{\beta}|\text{Datos}, \sigma, \lambda))$ a la posterior del modelo:

$$-\ln(\pi(\boldsymbol{\beta}|\text{Datos}, \sigma, \lambda)) \propto \sum_{i=1}^{n} \frac{1}{2\sigma^2}\Big(y_i - (\beta_1 X_{1i} + \cdots + \beta_p X_{pi})\Big)^2 + \frac{\lambda}{\sigma}||\boldsymbol{\beta}||_1.$$

- $\widehat{\boldsymbol{\beta}}_{\frac{\lambda}{\sigma}}^{\text{lasso}}$ es la moda a posteriori del modelo Bayesiano.

- Selección de modelo e inferencia vía $\pi(\boldsymbol{\beta}|\text{Datos}, \sigma, \lambda)$.
  - Ver § 3 de Park and G. Casella.

# Lasso vs Ridge: Ejercicio realizado en clase con R

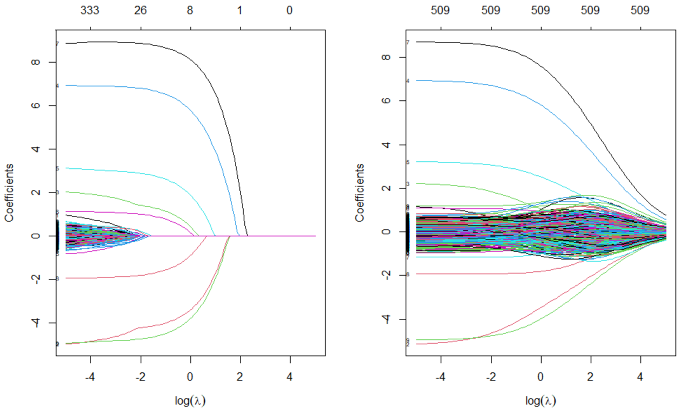



Figure: Comparativa Lasso (izquierda) vs Ridge (derecha).

- Lasso selecciona variables (descarta lo irrelevante).
- Lasso evita que los parámetros estimados alternen signo.

# Comparativa

- LASSO $\succ$ RIDGE: Muchas pendientes son cero.
- RIDGE $\succ$ LASSO: Covariables altamente correladas.

Simulación en R[2]:

| Método | $\alpha$ | $\lambda^*$ | Parámetros | VC Train (se) | Test |
|---|---|---|---|---|---|
| MCO | 0 | 0 | 509 | 0.62 (–) | 7.59 |
| RIDGE | 0 | 0.12 | 509 | 14.67 (0.74) | 7.44 |
| LASSO | 1 | 0.13 | **17** | 2.71 (0.09) | **2.87** |

Table: Estimaciones del ECM (SE entre paréntesis)

- Con lasso estimamos modelos parsimónicos.
- En escenarios ralos, lasso tiene propiedades asintóticas (ver simulación en R para discutir *sparsistencia*).

[2]El escenario de sparcidad hace a lasso el método óptimo.

# Generalización del método

$$\widehat{\beta}_{\lambda,q} = \underset{b_1,\dots,b_p}{\text{argmin}} \quad \sum_{i=1}^{n} \Big(y_i - (b_1 X_{1,i} + \cdots + b_p X_{p,i})\Big)^2 + \lambda ||\mathbf{b}||_q^2,$$

donde $\mathbf{b} = (b_1, \dots, b_p)$ y $||\mathbf{b}||_q = (\sum_{i=1}^{p} |b_i|^q)^{1/q}$ (norma $\ell_q$).

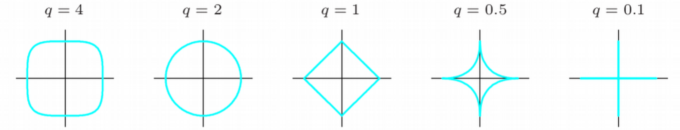



**FIGURE 3.12.** *Contours of constant value of $\sum_j |\beta_j|^q$ for given values of $q$.*

- Lasso como una relajación convexa cuando $q < 1$.
- No se gana demasiado aprendiendo "el mejor $q$".
  - Redes elásticas.

# Elastic Nets (Redes Elásticas)

$$\widehat{\beta}^{\text{EN}}_{\lambda,\alpha} = \underset{b_1,\ldots,b_p}{\operatorname{argmin}} \sum_{i=1}^{n} \Big(y_i - (b_1 X_{1,i} + \cdots + b_p X_{p,i})\Big)^2 + \lambda\big((1-\alpha)||\mathbf{b}||_2^2 + \alpha||\mathbf{b}||_1\big)$$

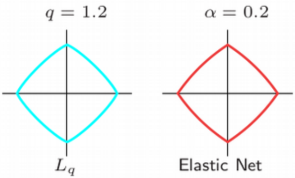



**FIGURE 3.13.** *Contours of constant value of* $\sum_j |\beta_j|^q$ *for* $q = 1.2$ *(left plot), and the elastic-net penalty* $\sum_j (\alpha\beta_j^2 + (1-\alpha)|\beta_j|)$ *for* $\alpha = 0.2$ *(right plot). Although visually very similar, the elastic-net has sharp (non-differentiable) corners, while the* $q = 1.2$ *penalty does not.*

- Ridge y Lasso en un solo método.
- Elastic net selecciona variables.
- Aprendemos $(\lambda, \alpha)$ por validación cruzada.

# Regularización con glmnet

```
install.packages('glmnet')
library(glmnet)

glmnet (x ,y , alpha = 0 , lambda )

x,y   =  en formato MATRIZ.

alpha = 0 (Ridge).
alpha = 1 (LASSO).
alpha en (0,1) (ENet).

lambda = penalización del regularizador.

Output:
Coeficientes estimados para cada valor de lambda.
```

# Preguntas / implementación en `R`

- Analiza los resultados del código `Simulación Lasso (sparsistencia)`. Investiga por tu cuenta a que hace referencia esta propiedad de los modelos estimados mediante lasso denominada **sparsistencia**.

- Siguiendo el ejercicio planteado en clase para ilustrar Ridge y Lasso, completa la Tabla en la slide 18 utilizando ENets.

- Utiliza los datos `wines` para implementar Elastic Nets (tendrás que construir un pequeño bucle para poder optimizar el valor del hiperparámetro $\alpha$).

- Conceptuales: Resuelve los puntos 2,3 y 4 planteados en § 6.8 de ISLR (pp. 259–a–261).

# Agenda

# Propiedades estadísticas de lasso (simplificado)

- Modelo de regresión: $\mathbf{y} = \mathbf{X}\boldsymbol{\beta}_0 + \varepsilon$, con $\boldsymbol{\beta}_0 \in \mathbb{R}^p$.
  - Active set: $\mathcal{S} = \{i | \beta_i \neq 0 \text{ con } \beta_i \in \boldsymbol{\beta}_0\}$.
  - Modelo ralo: $|\mathcal{S}| \ll p$ (formalmente $|\mathcal{S}| = o(\sqrt{n/\log(p)})$).
- A medida que $n \to \infty$ con $p/n \to 0$ (y bajo condiciones adecuadas sobre como elegir asintóticamente el valor de $\lambda$):
  - Consistencia: $\widehat{\beta}_i^{\text{lasso}} \to_P \beta_i$.
  - Identificamos el modelo correcto: $\widehat{\mathcal{S}}_n \to_P \mathcal{S}$.
  - El estimador lasso $\widehat{\boldsymbol{\beta}}^{\text{lasso}}$ es "esparsistente".
- La norma $\|\mathbf{X}(\boldsymbol{\beta}_0 - \widehat{\boldsymbol{\beta}}^{\text{lasso}})\|_2^2/n$ asociada al estimador lasso decrece a una tasa de al menos $\sqrt{n}$ (cond. grales. en $\mathbf{X}$).

# ECM del modelo de regresión con $p \gg 0$ volver a § 5

$$\mathbf{y}_{1\times n} = \mathbf{X}_{n\times p}\boldsymbol{\beta}_{p\times 1} + \boldsymbol{\varepsilon}_{1\times n}.$$

- Consideremos $\mathbf{X}$ fija y por lo tanto $\mathrm{Cov}(\mathbf{y}) = \sigma^2 \mathbf{I}_{n\times n}$.
- El ECM de predicción asociado a $\widehat{\beta}^{\text{ols}} = (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{y}$ es:

$$\frac{1}{n}E\Big\{\underbrace{\|\mathbf{X}\widehat{\beta}^{\text{ols}} - \mathbf{y}\|^2}_{\sum_{i=1}^n(\widehat{y}_i - y_i)^2}\Big\} = \frac{1}{n}E\Big\{\|\underbrace{\mathbf{X}(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T}_{P_{\mathbf{X}}}\mathbf{y} - \mathbf{X}\beta - \varepsilon\|^2\Big\}.$$

- Recordá que $\|P_{\mathbf{X}}\| = \mathrm{tr}(P_{\mathbf{X}}^T P_{\mathbf{X}}) = \mathrm{tr}(P_{\mathbf{X}}) = p$, luego:

$$\frac{1}{n}E\Big\{\|\mathbf{X}\widehat{\beta}^{\text{ols}} - \mathbf{y}\|^2\Big\} = \frac{\sigma^2}{n}\big(\mathrm{tr}(P_{\mathbf{X}}) + \mathrm{tr}(\mathbf{I})\big) = \frac{\sigma^2}{n}p + \sigma^2.$$

- Con $n$ fijo, el ECM de predicción crece de forma lineal con $p$.