Introducción a la Ciencia de Datos

Maestría en Probabilidad y Estadística

Dr. Marco Antonio Aquino López

Centro de Investigación en Matemáticas

Agosto–Diciembre 2025

Clasificación vs. Regresión

Problemas de Clasificación

- Variable respuesta es **categórica** (p. ej., spam / no spam).
- Objetivo: asignar cada observación a una clase.

Problemas de Regresión

- Variable respuesta es **continua** (p. ej., precio de una casa).
- Objetivo: predecir un valor numérico dado un conjunto de covariables.

¿Por qué estudiar regresión?

- La regresión es una de las herramientas más utilizadas en estadística aplicada.
- Permite no solo **predecir**, sino también **interpretar relaciones** entre variables.
- Es la base de muchos métodos más avanzados: desde modelos de series de tiempo hasta deep learning.
- Sirve como contraste natural con la clasificación:

  En lugar de decidir una etiqueta, cuantificamos un resultado.

¿Por qué regresión lineal?

Motivación

- Muchos problemas científicos y aplicados buscan entender la **relación entre variables**.
- Ejemplos:
  - ¿Cómo influye el tiempo de estudio en el desempeño académico?
  - ¿Cómo cambia el consumo eléctrico con la temperatura?
  - ¿Qué factores explican el precio de una vivienda?
- La regresión lineal es el **modelo más simple y ampliamente utilizado** para cuantificar estas relaciones.
- Sirve como **punto de partida** para modelos más sofisticados (generalizados, no lineales, bayesianos).

De la intuición al modelo

Idea básica

Aproximamos la relación entre X e Y con una **función lineal**:

$$Y \approx \beta_0 + \beta_1 X$$

- El parámetro β_0 representa el **intercepto**.
- El parámetro β_1 mide la **pendiente** o efecto de X sobre Y.
- La diferencia entre el valor observado y el predicho es el **error**.

Modelo de regresión lineal

Definición

Dados datos $\{(\mathbf{x}_i, y_i)\}_{i=1}^n$ con $y_i \in \mathbb{R}$, se asume

$$y_i = \mathbf{x}_i^\top \beta + \varepsilon_i,$$

donde $\beta \in \mathbb{R}^p$ es desconocido y ε_i son errores aleatorios.

Planteamiento del problema

- Queremos estimar el vector de parámetros β en el modelo:

$$y = X\beta + \varepsilon$$

- La idea central: encontrar $\hat{\beta}$ que haga que las predicciones

$$\hat{y} = X\hat{\beta}$$

  estén lo más cerca posible de los datos observados y.
- ¿Cómo medimos la cercanía? Usando la suma de cuadrados de los errores.

Función de pérdida: suma de cuadrados

Criterio de mínimos cuadrados

$$L(\beta) = \|y - X\beta\|^2 = (y - X\beta)^\top (y - X\beta)$$

- Buscamos $\hat{\beta}$ que minimice $L(\beta)$.
- Este es un problema de optimización cuadrática en β.
- La solución se obtiene derivando y resolviendo la ecuación normal.

Derivación del estimador OLS

$$L(\beta) = (y - X\beta)^\top (y - X\beta)$$

$$= y^\top y - 2\beta^\top X^\top y + \beta^\top X^\top X\beta$$

Derivada respecto a β

$$\frac{\partial L}{\partial \beta} = -2X^\top y + 2X^\top X\beta$$

$$\Rightarrow \quad X^\top X\hat{\beta} = X^\top y$$

Solución y geometría

Estimador de mínimos cuadrados

$$\hat{\beta} = (X^\top X)^{-1} X^\top y$$

- $\hat{y} = X\hat{\beta}$ es la proyección ortogonal de y sobre el subespacio columna de X.
- Los residuos $\hat{\varepsilon} = y - \hat{y}$ son ortogonales a todas las columnas de X.

$$X^\top \hat{\varepsilon} = 0$$

el subespacio columna de X.

- Se interpreta como **minimizar el error cuadrático** en el espacio de hipótesis lineales, lo que conecta regresión lineal con redes neuronales lineales.
- Nos permite construir inferencias, ya que la ortogonalidad garantiza la independencia entre estimadores y errores bajo supuestos clásicos.

¿Qué necesitamos para confiar en OLS?

- El estimador OLS siempre existe (si X tiene rango completo).
- Pero sus propiedades (insesgamiento, mínima varianza) dependen de ciertos supuestos.
- Estos supuestos se llaman **supuestos de Gauss–Markov**.
- Vamos a construirlos uno por uno.

Supuesto 1: Modelo correctamente especificado

$$y = X\beta + \varepsilon$$

- Asumimos que la relación es **lineal en parámetros**.
- Si el modelo es incorrecto (por ejemplo, falta una variable relevante), OLS ya no es insesgado.
- Este es el "piso" sobre el cual construimos los demás supuestos.

Supuesto 2: Rango completo de X

No colinealidad

Ninguna columna de X es combinación lineal exacta de las demás.

- Garantiza que $(X^\top X)^{-1}$ existe.
- Sin este supuesto, el problema de OLS no tiene solución única.
- En la práctica: cuidado con multicolinealidad (aunque sea aproximada).

Supuestos 3–5: Errores

1. $E[\varepsilon_i] = 0$
   $\Rightarrow$ asegura que OLS es insesgado.
2. $Var(\varepsilon_i) = \sigma^2$ constante (homocedasticidad)
   $\Rightarrow$ permite calcular varianzas de $\hat{\beta}$ y compararlas.
3. $Cov(\varepsilon_i, \varepsilon_j) = 0$ para $i \neq j$
   $\Rightarrow$ evita correlación que inflaría o reduciría la varianza.

Idea: los errores deben ser "ruido puro", sin patrón sistemático.

Supuesto 6: Predictores independientes del error

- Se asume que X es fijo (diseño experimental) o que $X \perp \varepsilon$ (en muestreo).
- Así, la variabilidad de y se explica sólo por la parte lineal más el error.
- Sin este supuesto, los predictores podrían estar "contaminados" y OLS perdería insesgamiento.

Teorema de Gauss–Markov

Enunciado

Bajo los supuestos clásicos:

- Modelo lineal en parámetros: $y = X\beta + \varepsilon$.
- $E[\varepsilon] = 0$, $Var(\varepsilon) = \sigma^2 I$.
- X de rango completo.

Entonces, el estimador OLS

$$\hat{\beta} = (X^\top X)^{-1} X^\top y$$

es el **mejor estimador lineal insesgado (BLUE)**: cualquier otro estimador lineal insesgado de β tiene varianza mayor o igual.

Demostración: idea general

1. Mostrar que $\hat{\beta}$ es insesgado: $E[\hat{\beta}] = \beta$.
2. Calcular su varianza: $Var(\hat{\beta}) = \sigma^2(X^\top X)^{-1}$.
3. Considerar cualquier otro estimador lineal insesgado:

$$\tilde{\beta} = Cy, \quad CX = I_p.$$

4. Probar que $Var(\tilde{\beta}) - Var(\hat{\beta})$ es semidefinida positiva.

Paso 1: Insesgamiento de OLS

$$\hat{\beta} = (X^\top X)^{-1} X^\top y$$

Sustituyendo $y = X\beta + \varepsilon$:

$$\hat{\beta} = (X^\top X)^{-1} X^\top (X\beta + \varepsilon) = \beta + (X^\top X)^{-1} X^\top \varepsilon$$

Entonces:

$$E[\hat{\beta}] = \beta + (X^\top X)^{-1} X^\top E[\varepsilon] = \beta$$

Conclusión: OLS es insesgado.

Paso 2: Varianza de OLS

Como $\hat{\beta} = \beta + (X^\top X)^{-1} X^\top \varepsilon$:

$$Var(\hat{\beta}) = (X^\top X)^{-1} X^\top Var(\varepsilon) X (X^\top X)^{-1}$$

Dado que $Var(\varepsilon) = \sigma^2 I$:

$$Var(\hat{\beta}) = \sigma^2 (X^\top X)^{-1}$$

Modelo Bayesiano: Motivación

- La regresión lineal clásica asume parámetros fijos desconocidos.
- En el enfoque bayesiano, los parámetros son **variables aleatorias**.
- Esto permite:
  - Incorporar información previa (*priors*).
  - Obtener distribuciones posteriores que cuantifican la incertidumbre.

Modelo Lineal

Planteamiento

$$\mathbf{y} = X\boldsymbol{\beta} + \boldsymbol{\varepsilon}, \qquad \boldsymbol{\varepsilon} \sim \mathcal{N}(\mathbf{0}, \sigma^2 I_n)$$

- $\mathbf{y}$: vector $n \times 1$ de respuestas.
- X: matriz de diseño $n \times p$.
- $\boldsymbol{\beta}$: vector $p \times 1$ de coeficientes.
- σ^2: varianza común de los errores.

Enfoque Bayesiano

- Ahora tratamos $\boldsymbol{\beta}$ y σ^2 como aleatorios.
- Teorema de Bayes:

$$p(\boldsymbol{\beta}, \sigma^2 \mid \mathbf{y}, X) \propto p(\mathbf{y} \mid X, \boldsymbol{\beta}, \sigma^2)\, p(\boldsymbol{\beta}, \sigma^2)$$

Prior Conjugado

Normal–Inversa Gamma

$$\boldsymbol{\beta} \mid \sigma^2 \sim \mathcal{N}(\boldsymbol{\beta}_0, \sigma^2 V_0), \qquad \sigma^2 \sim \text{Inv-Gamma}(a_0, b_0)$$

- Familia conjugada que garantiza posterior con la misma forma.

Posterior Conjugada

Distribuciones posteriores

$$\boldsymbol{\beta} \mid \sigma^2, \mathbf{y} \sim \mathcal{N}(\boldsymbol{\beta}_n, \sigma^2 V_n)$$

$$\sigma^2 \mid \mathbf{y} \sim \text{Inv-Gamma}(a_n, b_n)$$

Posterior Conjugada

Distribuciones posteriores

$$\boldsymbol{\beta} \mid \sigma^2, \mathbf{y} \sim \mathcal{N}(\boldsymbol{\beta}_n, \sigma^2 V_n)$$

$$\sigma^2 \mid \mathbf{y} \sim \text{Inv-Gamma}(a_n, b_n)$$

Parámetros actualizados

$$V_n = \left(V_0^{-1} + X^\top X\right)^{-1}, \quad \boldsymbol{\beta}_n = V_n \left(V_0^{-1}\boldsymbol{\beta}_0 + X^\top \mathbf{y}\right)$$

$$a_n = a_0 + \tfrac{n}{2}, \quad b_n = b_0 + \tfrac{1}{2}\left(\mathbf{y}^\top \mathbf{y} + \boldsymbol{\beta}_0^\top V_0^{-1}\boldsymbol{\beta}_0 - \boldsymbol{\beta}_n^\top V_n^{-1}\boldsymbol{\beta}_n\right)$$

Propiedades de la Posterior

- **Distribución conjunta**: $(\boldsymbol{\beta}, \sigma^2)$ sigue una forma conjugada tractable.
- **Marginal de** $\boldsymbol{\beta}$: una distribución t multivariada.
- **Marginal de** σ^2: una Inversa-Gamma.
- Refleja la combinación de evidencia previa y datos observados.

¿Por qué una Distribución Predictiva?

- La posterior nos dice lo que creemos de los **parámetros**.
- Pero en la práctica, muchas veces lo que queremos es:
  - Predecir una nueva observación y_*.
  - Evaluar el ajuste del modelo frente a datos no observados.
  - Construir intervalos de predicción con interpretación probabilística.

¿Por qué una Distribución Predictiva?

- La posterior nos dice lo que creemos de los **parámetros**.
- Pero en la práctica, muchas veces lo que queremos es:
  - Predecir una nueva observación y_*.
  - Evaluar el ajuste del modelo frente a datos no observados.
  - Construir intervalos de predicción con interpretación probabilística.
- La **distribución predictiva** combina:
  - La incertidumbre en los parámetros $(\boldsymbol{\beta}, \sigma^2)$.
  - La variabilidad inherente de los errores.

¿Por qué una Distribución Predictiva?

- La posterior nos dice lo que creemos de los **parámetros**.
- Pero en la práctica, muchas veces lo que queremos es:
  - Predecir una nueva observación y_*.
  - Evaluar el ajuste del modelo frente a datos no observados.
  - Construir intervalos de predicción con interpretación probabilística.
- La **distribución predictiva** combina:
  - La incertidumbre en los parámetros $(\boldsymbol{\beta}, \sigma^2)$.
  - La variabilidad inherente de los errores.
- Resultado: una distribución completa para y_* que refleja toda la incertidumbre disponible.

Distribución Predictiva

- Definición general:

$$p(y_* \mid \mathbf{y}, X, \mathbf{x}_*) = \int \int p(y_* \mid \boldsymbol{\beta}, \sigma^2, \mathbf{x}_*)\ p(\boldsymbol{\beta}, \sigma^2 \mid \mathbf{y}, X)\ d\boldsymbol{\beta}\, d\sigma^2$$

- Dos integrales:
  1. Primero sobre $\boldsymbol{\beta}$.
  2. Luego sobre σ^2.

Paso 1: Integración sobre $\boldsymbol{\beta}$

$$p(y_* \mid \sigma^2, \mathbf{y}, X, \mathbf{x}_*) = \int \int p(y_* \mid \boldsymbol{\beta}, \sigma^2, \mathbf{x}_*)\ p(\boldsymbol{\beta} \mid \sigma^2, \mathbf{y}, X)\ d\boldsymbol{\beta}$$

Paso 1: Integración sobre $\boldsymbol{\beta}$

$$p(y_* \mid \sigma^2, \mathbf{y}, X, \mathbf{x}_*) = \int \int p(y_* \mid \boldsymbol{\beta}, \sigma^2, \mathbf{x}_*)\; p(\boldsymbol{\beta} \mid \sigma^2, \mathbf{y}, X)\; d\boldsymbol{\beta}$$

Resultado

$$y_* \mid \sigma^2, \mathbf{y}, X, \mathbf{x}_* \sim \mathcal{N}\Big(\mathbf{x}_*^\top \boldsymbol{\beta}_n,\; \sigma^2(1 + \mathbf{x}_*^\top V_n \mathbf{x}_*)\Big)$$

Paso 2: Integración sobre σ^2

$$p(y_* \mid \mathbf{y}, X, \mathbf{x}_*) = \int p(y_* \mid \sigma^2, \mathbf{y}, X, \mathbf{x}_*)\ p(\sigma^2 \mid \mathbf{y}, X)\ d\sigma^2$$

Paso 2: Integración sobre σ^2

$$p(y_* \mid \mathbf{y}, X, \mathbf{x}_*) = \int p(y_* \mid \sigma^2, \mathbf{y}, X, \mathbf{x}_*)\; p(\sigma^2 \mid \mathbf{y}, X)\; d\sigma^2$$

- Sabemos que:
$$\sigma^2 \mid \mathbf{y}, X \sim \text{Inv-Gamma}(a_n, b_n)$$
- La mezcla Normal–Inv-Gamma es conocida: produce una distribución t.

Resultado Final

Distribución predictiva

$$y_* \mid \mathbf{y}, X, \mathbf{x}_* \sim t_{2a_n}\left(\mathbf{x}_*^\top \boldsymbol{\beta}_n,\ \frac{b_n}{a_n}\left(1 + \mathbf{x}_*^\top V_n \mathbf{x}_*\right)\right)$$

- Grados de libertad: $2a_n$.
- Media: $\mathbf{x}_*^\top \boldsymbol{\beta}_n$.
- Varianza escalada: $\frac{b_n}{a_n}\left(1 + \mathbf{x}_*^\top V_n \mathbf{x}_*\right)$.

Predicción Bayesiana

Nueva observación $\mathbf{x}_*$

$$y_* \mid \mathbf{y}, X, \mathbf{x}_* \sim t_{2a_n}\left(\mathbf{x}_*^\top \boldsymbol{\beta}_n,\ \frac{b_n}{a_n}\left(1 + \mathbf{x}_*^\top V_n \mathbf{x}_*\right)\right)$$

- La distribución predictiva es una t de Student.
- Incorpora tanto la incertidumbre de parámetros como la varianza residual.

Supuestos de la Regresión Lineal Bayesiana

- **Linealidad**: la relación entre covariables y respuesta se modela como

$$\mathbf{y} = X\boldsymbol{\beta} + \boldsymbol{\varepsilon}.$$

- **Normalidad de los errores**:

$$\boldsymbol{\varepsilon} \sim \mathcal{N}(\mathbf{0},\, \sigma^2 I_n).$$

- **Independencia**: los errores son independientes entre observaciones.
- **Varianza constante**: homocedasticidad, la misma σ^2 para todos los y_i.
- **Especificación del prior**: se asume una distribución previa para $(\boldsymbol{\beta}, \sigma^2)$, típicamente Normal–Inversa Gamma.
- **Matriz de diseño X conocida y fija**: no aleatoria en el análisis.

Clásica vs. Bayesiana (I)

Aspecto	**Regresión Clásica**	**Regresión Bayesiana**
Tratamiento de parámetros	Fijos pero desconocidos	Variables aleatorias con distribución a priori
Resultado principal	Estimadores puntuales e IC basados en asintótica	Distribuciones posteriores e intervalos creíbles con interpretación probabilística
Uso de información previa	No incluye (salvo restricciones)	Incorpora priors que reflejan conocimiento experto

Clásica vs. Bayesiana (II)

Aspecto	**Regresión Clásica**	**Regresión Bayesiana**
Predicción	Basada en estimadores puntuales, IC aproximados	Predictiva completa que incluye incertidumbre de parámetros y ruido
Interpretación	Simple, estándar en estadística clásica	Más rica, pero requiere mayor formación conceptual
Costo computacional	Bajo (soluciones analíticas cerradas)	Puede ser alto (inferencia numérica/MCMC si no hay conjugación)
Flexibilidad	Limitada a extensiones clásicas	Extensible a modelos jerárquicos, no lineales, mixtos

Conexiones con OLS y Ridge

- **OLS como caso límite:** Si $\Sigma_0^{-1} \to 0$ y $a_0, b_0 \to 0$ (prior débil / no informativo), entonces

$$\boldsymbol{\beta}_n \to (X^\top X)^{-1} X^\top y \quad \text{(Mínimos Cuadrados Ordinarios).}$$

Conexiones con OLS y Ridge

- **OLS como caso límite:** Si $\Sigma_0^{-1} \to 0$ y $a_0, b_0 \to 0$ (prior débil / no informativo), entonces

$$\boldsymbol{\beta}_n \to (X^\top X)^{-1} X^\top y \quad \text{(Mínimos Cuadrados Ordinarios)}.$$

- **Ridge como MAP:** Con prior gaussiano isotrópico

$$\boldsymbol{\beta} \mid \sigma^2 \sim \mathcal{N}(0,\, \sigma^2 \lambda^{-1} I),$$

el *modo a posteriori* satisface

$$\hat{\beta}_{\text{MAP}} = \arg\min_{\beta} \|y - X\beta\|^2 + \lambda \|\beta\|^2,$$

que coincide con la **regresión Ridge**.

Regresión Ridge: Motivación

- En Mínimos Cuadrados Ordinarios (OLS), los estimadores pueden tener alta varianza si:
  - Las variables están fuertemente correlacionadas.
  - El número de predictores p es grande comparado con n.
- Ridge busca reducir la varianza mediante una **penalización cuadrática**.
- **El parámetro λ (lambda):**
  - Controla la **fuerza de la penalización** en los coeficientes
  - $\lambda \geq 0$ (siempre un valor no negativo)

Regresión Ridge en Bayes

- Enfoque bayesiano: asumimos un prior normal isotrópico sobre los coeficientes:

$$\beta \mid \sigma^2 \sim \mathcal{N}\big(0,\, \sigma^2 \lambda^{-1} I\big).$$

Regresión Ridge en Bayes

- Enfoque bayesiano: asumimos un prior normal isotrópico sobre los coeficientes:

$$\beta \mid \sigma^2 \sim \mathcal{N}\big(0,\, \sigma^2 \lambda^{-1} I\big).$$

- La función de densidad es

$$p(\beta \mid \sigma^2) \propto \exp\big(-\tfrac{\lambda}{2\sigma^2}\|\beta\|^2\big)\,.$$

- Combinada con la verosimilitud gaussiana, el **estimador Maximo a Posterior (MAP)** es:

$$\hat{\beta}_{MAP} = \arg\min_{\beta} \|y - X\beta\|^2 + \lambda\|\beta\|^2.$$

Interpretación Bayesiana de Ridge

- El prior normal isotrópico refleja la creencia de que los coeficientes tienden a estar **cerca de cero**, pero sin forzarlos exactamente.
- La penalización λ controla el grado de contracción:
- **El parámetro λ (lambda):**
  - Controla la **fuerza de la penalización** en los coeficientes
  - $\lambda \geq 0$ (siempre un valor no negativo)
  - Cuando $\lambda = 0$: recuperamos el modelo OLS estándar
  - Cuando $\lambda \to \infty$: todos los coeficientes se reducen hacia cero
  - Valores intermedios de λ encuentran equilibrio entre sesgo y varianza
- La regresión Ridge es equivalente a imponer un prior gaussiano y tomar el MAP.
- Extiende OLS reduciendo varianza a cambio de un sesgo controlado (*bias–variance tradeoff*).
  ¿Cómo elegir λ? Normalmente mediante validación cruzada.

Conexiones con Lasso

- **Lasso como MAP:** Si asumimos que los coeficientes siguen una distribución Laplace:

$$\beta_j \sim \text{Laplace}(0, \tau),$$

entonces obtenemos:

$$\hat{\beta}_{MAP} = \arg\min_{\beta} \|y - X\beta\|^2 + \lambda\|\beta\|_1,$$

que es exactamente la **regresión Lasso** [3].

- **¿Qué hace el parámetro λ aquí?**
  - Controla el **nivel de penalización** en los coeficientes
  - λ grande: coeficientes más pequeños, más variables se hacen cero
  - λ pequeño: coeficientes más grandes, menos variables se eliminan
- **Características clave:**
  - Penaliza la suma de valores absolutos (no cuadrados)
  - *Sparsity*: produce coeficientes exactamente cero (elimina variables)
  - **En palabras simples:** Lasso no solo reduce coeficientes, sino que elimina variables innecesarias

Regresión Lasso: Motivación

- **Problema con Ridge:** Reduce coeficientes pero nunca los hace exactamente cero.
- **Solución con Lasso:** Puede eliminar variables completamente.
- **¿Cuándo usar Lasso?**
  - Cuando tenemos muchas variables y queremos identificar las más importantes
  - Cuando buscamos modelos más simples y fáciles de interpretar
  - Cuando creemos que solo algunas variables realmente importan
- **La fórmula de Lasso:**

$$\hat{\beta}^{\text{Lasso}} = \arg\min_{\beta} \|y - X\beta\|^2 + \lambda\|\beta\|_1,$$

  donde $\|\beta\|_1 = \sum_j |\beta_j|$ (suma de valores absolutos).
- **El rol de λ:**
  - Controla cuántas variables se eliminan
  - λ muy grande: casi todos los coeficientes son cero
  - λ muy pequeño: casi como OLS tradicional

Regresión Lasso en Bayes

- **Enfoque bayesiano:** Asumimos que cada coeficiente sigue una distribución Laplace:

$$\beta_j \sim \text{Laplace}(0, \tau), \qquad p(\beta_j) \propto \exp\left(-\frac{|\beta_j|}{\tau}\right).$$

- **¿Por qué Laplace?**
  - Esta distribución favorece que muchos coeficientes sean exactamente cero
  - Es como decir: "creemos que muchas variables probablemente no son importantes"
- **Resultado importante:** El máximo a posteriori (MAP) bajo este prior es:

$$\hat{\beta}_{MAP} = \arg\min_{\beta} \|y - X\beta\|^2 + \lambda\|\beta\|_1.$$

- **Conclusión:** Lasso = solución bayesiana con distribuciones Laplace
- **Relación λ - τ:**
  - $\lambda = \frac{1}{\tau}$ (parámetros inversamente relacionados)
  - λ grande $\Leftrightarrow$ τ pequeño: más penalización, más coeficientes cero

Interpretación Bayesiana de Lasso

- **¿Por qué Lasso elimina variables?**
  - La distribución Laplace tiene "picos" en cero
  - Esto hace que sea más probable que los coeficientes sean exactamente cero
  - **Analogía:** Es como tener un "prejuicio" de que muchas variables son innecesarias
- **Ventajas de Lasso:**
  - **Selección automática:** Elige qué variables incluir y cuáles eliminar
  - **Modelos simples:** Produce modelos más fáciles de interpretar
  - **Elimina redundancias:** Ayuda cuando hay variables correlacionadas
- **Limitaciones:**
  - Si varias variables están muy correlacionadas, Lasso puede elegir arbitrariamente una
  - Puede ser inestable con datos muy correlacionados
  - λ debe elegirse cuidadosamente (usualmente por validación cruzada)

g-prior de Zellner: Motivación

- **Problema con Ridge y Lasso:**
  - Tratan todos los coeficientes como igualmente importantes
  - No consideran cómo se relacionan las variables entre sí
  - **Analogía:** Es como dar el mismo peso a todas las materias sin importar su dificultad
- **Solución del g-prior:**
  - Usa la información de los datos mismos para definir el prior
  - Considera que algunas direcciones en los datos son más importantes que otras
  - **En palabras simples:** Ajusta el prior según la "geometría" de tus datos
- **Objetivo:** Balancear "inteligentemente" entre lo que creemos antes de ver los datos y lo que los datos nos dicen

g-prior de Zellner: La Fórmula

- **La fórmula clave:**

$$\boldsymbol{\beta} \mid \sigma^2 \sim \mathcal{N}\big(0,\, g\sigma^2 (X^\top X)^{-1}\big).$$

- **¿Qué significa cada parte?**
  - $\boldsymbol{\beta}$: coeficientes del modelo
  - σ^2: varianza del error
  - $(X^\top X)^{-1}$: información sobre cómo se relacionan las variables
  - g: **parámetro clave** que controla la fuerza del prior
- **El parámetro g explicado:**
  - Controla cuánto "confiamos" en nuestro prior vs. en los datos
  - g pequeño: más confianza en el prior (coeficientes cerca de cero)
  - g grande: más confianza en los datos (coeficientes más libres)
  - $g \to \infty$: recuperamos OLS tradicional

¿Por qué el g-prior es "inteligente" ?

- **Considera las correlaciones:**
  - Si dos variables están correlacionadas, el g-prior lo tiene en cuenta
  - No trata cada variable como si fuera independiente de las demás
- **Ajusta por la escala:**
  - Automáticamente considera que algunas variables tienen más varianza que otras
  - No necesitamos estandarizar las variables manualmente
- **Ejemplo intuitivo:**
  - Imagina predecir el precio de una casa
  - El g-prior reconoce que "área" y "número de habitaciones" están relacionadas
  - No las trata como variables completamente separadas
- **Ventaja práctica:** Cálculos más eficientes y resultados más estables

Aplicaciones y Consideraciones Prácticas

- **Selección de modelos:**
  - Diferentes valores de g nos llevan a preferir modelos distintos
  - Podemos comparar qué tan bueno es cada modelo usando factores de Bayes
- **Valores típicos de g:**
  - $g = n$ (tamaño de muestra): "prior de información unitaria"
  - $g = p^2$ o $g = n/p^2$: otras opciones comunes
  - En la práctica: probamos varios valores y elegimos el mejor
- **Ventajas:**
  - Muy bueno para comparar modelos
  - Considera la estructura de correlación de los datos
  - Cálculos relativamente simples
- **Precauciones:**
  - La elección de g es importante
  - Puede ser sensible si hay variables muy correlacionadas
  - Necesitamos que $X^T X$ sea invertible

Referencias I

Gauss, C. F. (1821).
Teoría de la combinación de observaciones sujeta a los errores mínimos
[Título original en latín: *Theoria combinationis observationum erroribus minimis obnoxiae*].

Commentationes Societatis Regiae Scientiarum Gottingensis Recentiores, Vol. V.

Hoerl, A. E., & Kennard, R. W. (1970).
Ridge regression: Biased estimation for nonorthogonal problems.
Technometrics, 12(1), 55–67.

Tibshirani, R. (1996).
Regression shrinkage and selection via the Lasso.
Journal of the Royal Statistical Society: Series B, 58(1), 267–288.

Referencias II

Zellner, A. (1986).
On assessing prior distributions and Bayesian regression analysis with g-prior distributions.
In *Bayesian Inference and Decision Techniques*, pp. 233–243.

Referencias adicionales I


Bishop, C. M. (2006).
Pattern Recognition and Machine Learning.
Springer. (**Cap. 3**: Regresión lineal, conexión Bayes–MAP con Ridge y Lasso)


Gelman, A., Carlin, J. B., Stern, H. S., Dunson, D., Vehtari, A., & Rubin, D. (2013).
Bayesian Data Analysis (3rd ed.).
Chapman & Hall/CRC. (**Cap. 14**: Modelos lineales bayesianos, priors conjugados y g-priors)


Hastie, T., Tibshirani, R., & Friedman, J. (2009).
The Elements of Statistical Learning (2nd ed.).
Springer. (**Cap. 3**: Regularización, Ridge y Lasso, desde el punto de vista clásico)


Murphy, K. P. (2023).
Probabilistic Machine Learning: Advanced Topics.
MIT Press. (**Cap. 12**: Bayesian linear regression, priors y conexión con regularización)

Motivación: Cuando la respuesta es sí/no

- **Ejemplos del mundo real:**
  - ¿Aprobará el examen? *Sí/No*
  - ¿Tendrá la enfermedad? *Sí/No*
- **El problema fundamental:**
  - Regresión lineal: $y = \beta_0 + \beta_1 x + \epsilon$
  - Puede predecir: $y = 1{,}2$ o $y = -0{,}3$
  - **¡Esto no tiene sentido como probabilidad!**

Recordatorio importante

En problemas de clasificación binaria, no modelamos la categoría directamente, sino la **probabilidad** de pertenecer a cada categoría.

Construyendo el modelo: La necesidad de una transformación

- **Problema central:** Necesitamos conectar el predictor lineal $\eta_i = x_i^\top \beta$ con una probabilidad $\pi_i = P(y_i = 1)$.
- **¿Por qué no usar directamente?** $\pi_i = x_i^\top \beta$
  - Podría dar $\pi_i < 0$ o $\pi_i > 1$
  - **¡Las probabilidades deben estar entre 0 y 1!**
- **Propuesta natural:** Buscar una función f que:
  - Transforme cualquier número real a [0,1]
  - Sea creciente (a mayor η_i, mayor probabilidad)
  - Sea invertible (para poder interpretar)
- **Candidatas comunes:**
  - Función logística
  - Función de distribución normal (probit)
  - Función log-log

Pregunta para reflexionar

¿Qué propiedades debería tener la función ideal para modelar probabilidades?

La función logística: una elección natural

- **Función logística:**

$$f(z) = \frac{1}{1 + e^{-z}} = \frac{e^z}{1 + e^z}$$

- **Propiedades clave:**
  - Cuando $z \to -\infty$, $f(z) \to 0$
  - Cuando $z \to +\infty$, $f(z) \to 1$
  - $f(0) = 0{,}5$ (punto medio)
  - Forma de "S" suave (sigmoide)
- **Aplicando a nuestro modelo:**

$$\pi_i = P(y_i = 1 \mid x_i) = \frac{1}{1 + e^{-x_i^\top \beta}}$$

- **Interpretación:** La probabilidad depende del predictor lineal a través de esta transformación

Visualización

La curva logística transforma valores entre $-\infty$ y $+\infty$ al rango [0,1] de manera suave.

El logit: la transformación inversa

- **Problema:** ¿Cómo recuperar la relación lineal?
- **Despejemos:**

$$\pi_i = \frac{1}{1 + e^{-x_i^\top \beta}}$$

$$1 - \pi_i = \frac{e^{-x_i^\top \beta}}{1 + e^{-x_i^\top \beta}}$$

$$\frac{\pi_i}{1 - \pi_i} = e^{x_i^\top \beta}$$

- **Definición del logit:**

$$\text{logit}(\pi_i) = \log\left(\frac{\pi_i}{1 - \pi_i}\right) = x_i^\top \beta$$

- **Interpretación:**
  - $\frac{\pi_i}{1-\pi_i}$ = **odds** (razón de probabilidades)
  - $\log(\text{odds})$ = **log-odds**
  - ¡El logit nos devuelve la estructura lineal!

Modelo logístico y función logit

Definición

Para variable binaria $y_i \in \{0, 1\}$, la regresión logística modela la probabilidad $p_i = \Pr(y_i = 1)$ mediante

$$\text{logit}(p_i) = \log\left(\frac{p_i}{1 - p_i}\right) = \mathbf{x}_i^\top \beta.$$

La inversa logística transforma de regreso a la escala de probabilidades: $p_i = \frac{e^{\mathbf{x}_i^\top \beta}}{1+e^{\mathbf{x}_i^\top \beta}}$.

- Los coeficientes β representan cambios en los *log-odds*: un incremento en x_j multiplica las odds por e^{β_j}.
- La relación entre probabilidad y predictor es no lineal; los efectos se moderan en los extremos de la curva sigmoidal.

Estimación por máxima verosimilitud

Log-verosimilitud

La verosimilitud para datos $\{y_i\}$ bajo independencia es

$$\ell(\beta) = \sum_{i=1}^{n} \left[y_i \log(p_i) + (1 - y_i) \log(1 - p_i) \right].$$

No existe solución analítica para maximizar $\ell(\beta)$; se utiliza el método de Newton–Raphson o iterativamente reponderación de mínimos cuadrados (IRLS).

- La función de verosimilitud es cóncava; Newton–Raphson converge rápidamente.
- Los errores estándar de $\hat{\beta}$ se obtienen a partir de la matriz de información de Fisher.

De máxima verosimilitud a enfoque bayesiano

- **Enfoque clásico (MLE):**
  - Encontramos $\hat{\beta}$ que maximiza la verosimilitud
  - Respuesta única sin cuantificar incertidumbre
- **Problema:** ¿Y si queremos incorporar conocimiento previo o medir incertidumbre?
- **Solución bayesiana:** Tratamos β como variable aleatoria

$$\underbrace{p(\beta \mid y, X)}_{\text{posterior}} \propto \underbrace{\prod_{i=1}^{n} \pi_i^{y_i}(1-\pi_i)^{1-y_i}}_{\text{verosimilitud}} \times \underbrace{p(\beta)}_{\text{prior}}$$

- **Ventajas:**
  - Incorpora conocimiento previo mediante el prior
  - Cuantifica incertidumbre completa sobre los parámetros
  - Predicciones más robustas

El corazón del enfoque bayesiano

Actualizamos nuestro conocimiento sobre β combinando lo que sabíamos antes (prior) con lo que nos dicen los datos (verosimilitud).

¿Cómo elegir nuestros priors?

- **Prior Normal (Gaussiano):**

$$\beta \sim \mathcal{N}(0, \sigma^2 I)$$

  - **Equivalente frecuentista:** Regularización Ridge
  - **Uso:** Cuando queremos coeficientes pequeños pero no cero
- **Prior Laplace (Doble Exponencial):**

$$\beta_j \sim \text{Laplace}(0, \tau)$$

  - **Equivalente frecuentista:** Regularización Lasso
  - **Uso:** Para selección automática de variables
- **Priors informativos:** Cuando tenemos conocimiento experto específico

Reto computacional

A diferencia de la regresión lineal normal, aquí no existe un prior conjugado que simplifique los cálculos.

El desafío computacional y sus soluciones

- **Problema:** La posterior no tiene forma analítica cerrada

$$p(\beta \mid y, X) \propto \left[\prod_{i=1}^{n} \pi_i^{y_i} (1 - \pi_i)^{1-y_i} \right] \times p(\beta)$$

- **Solución**

MCMC (Monte Carlo vía Cadenas de Markov)

- Herramientas prácticas: `Stan`, `PyMC`, `JAGS` and `T-walk`.
- Computacionalmente costoso

Predicciones bayesianas: Más allá del punto único

- Recordemos: $\pi(z) = \frac{1}{1+e^{-z}}$ (función logística)
- **Enfoque clásico:** Predecimos con $\hat{\beta}_{MLE}$

$$P(y_* = 1 \mid x_*) \approx \pi(x_*^\top \hat{\beta}_{MLE})$$

- **Enfoque bayesiano:** Promediamos sobre toda la incertidumbre

$$P(y_* = 1 \mid x_*, \textit{datos}) = \int \pi(x_*^\top \beta)\, p(\beta \mid \textit{datos})\, d\beta$$

En la práctica

Usamos software como `Stan`, `Twalk`, o `JAGS` que automatizan estos cálculos complejos.

¿Por qué necesitamos diagnosticar los residuales?

- Al ajustar una regresión lineal con Mínimos Cuadrados, hacemos inferencias sobre los coeficientes β_j:
  - ¿Son significativamente distintos de 0?
  - ¿Explican realmente la variación de y?
- Estas pruebas (t, F, intervalos de confianza) se basan en supuestos sobre los errores:

$$e_i \sim N(0, \sigma^2), \quad \text{independientes y homocedásticos}$$

- **Si los supuestos no se cumplen**:
  - Los errores estándar pueden estar sesgados.
  - Los valores t y F pueden ser incorrectos.
  - La inferencia estadística deja de ser válida.
- **Diagnósticos de residuales** = herramientas para verificar que el modelo es confiable.

Normalidad de los Residuales: Verificación

- Supuesto: $e_i \sim N(0, \sigma^2)$.
- Métodos:
  - **QQ-plot**: los residuales deberían seguir una línea recta.
  - **Prueba de Shapiro-Wilk**: hipótesis H_0: normalidad.

Normalidad de los Residuales: Problemas y Soluciones

- Problema: Si no hay normalidad, las pruebas t y F pierden validez.
- Posibles causas:
  - Datos con colas pesadas.
  - Outliers.
- Soluciones:
  - Transformar la variable ($\log$, $\sqrt{\cdot}$).

Homocedasticidad: Verificación

- Supuesto: $\text{Var}(e_i) = \sigma^2$ constante.
- Verificación:
  - Gráfico de residuales vs. valores ajustados.
  - Residuales deben estar distribuidos aleatoriamente sin patrón.

Homocedasticidad: Problemas y Soluciones

- Problema: heterocedasticidad $\rightarrow$ errores estándar incorrectos.
- Consecuencia: intervalos y pruebas de hipótesis poco fiables.
- Soluciones:
  - Transformar y (log, Box-Cox).
  - Modelar explícitamente la varianza (GLS).

Independencia de Residuales: Verificación

- Supuesto: no hay correlación entre residuales.
- Importante en series de tiempo.
- Método:
  - **Prueba Durbin-Watson**:

$$DW = \frac{\sum (e_i - e_{i-1})^2}{\sum e_i^2}$$

  - Valores cercanos a 2 indican independencia.

Independencia de Residuales: Problemas y Soluciones

- Problema: autocorrelación $\Rightarrow$ subestimación de incertidumbre.
- Consecuencias: estadísticos t y F inflados.
- Soluciones:
  - Incluir rezagos de la variable dependiente o regresores adicionales.
  - Usar modelos de series de tiempo (AR, ARIMA).

Leverage y Outliers: Verificación

- Algunos puntos tienen gran influencia en el ajuste.
- Diagnósticos:
  - **Distancia de Cook**.
  - **Residuales studentizados**.
- Visualización: gráficos de leverage vs. residuales.

Leverage y Outliers: Consecuencias y Soluciones

- Problema: un punto puede distorsionar los coeficientes.
- Consecuencias:
  - Parámetros sesgados.
  - Inferencia inválida.
- Soluciones:
  - Revisar datos y posible error de medición.
  - Métodos robustos a outliers (regresión robusta).
  - Considerar transformaciones.

¿Por qué necesitamos evaluar modelos?

- Después de ajustar un modelo (lineal o logístico, clásico o bayesiano) surgen preguntas clave:
  - ¿Qué tan bien se ajusta a los datos observados?
  - ¿Predice correctamente nuevos datos?
  - ¿Es mejor este modelo que otro más simple o más complejo?
- Problema central: **no basta con ajustar un modelo, debemos comparar su desempeño**.
- Distintos enfoques:
  - **Clásico**: R^2, pruebas F y t, AIC/BIC, diagnósticos de residuales.
  - **Bayesiano**: DIC, WAIC, LOO, Posterior Predictive Checks.
- **Objetivo**: Equilibrar entre **ajuste** y **complejidad**, integrando la **incertidumbre** de manera adecuada.

R^2: Definición e Intuición

- Mide la **proporción de variabilidad** de y explicada por el modelo.

$$R^2 = 1 - \frac{SS_{res}}{SS_{tot}}$$

- Donde:
  - $SS_{res} = \sum (y_i - \hat{y}_i)^2$ (suma de cuadrados residuales).
  - $SS_{tot} = \sum (y_i - \bar{y})^2$ (suma total de cuadrados).
- Valores posibles:

$$0 \leq R^2 \leq 1$$

R^2: Interpretación

- $R^2 = 0$: el modelo no explica nada mejor que la media.
- $R^2 = 1$: el modelo explica toda la variación.
- Ejemplo:
  - $R^2 = 0{,}85 \rightarrow$ el 85 % de la variación de y es explicada por el modelo.

Limitaciones

- Siempre aumenta al agregar variables (aunque no sean relevantes).
- No mide calidad predictiva fuera de muestra.

R^2 Ajustado: Motivación

- Problema: R^2 siempre sube al agregar variables.
- Solución: Penalizar por número de predictores p.
- Fórmula:

$$R^2_{adj} = 1 - \frac{SS_{res}/(n-p-1)}{SS_{tot}/(n-1)}$$

R^2 Ajustado: Propiedades

- Puede ser menor que R^2.
- Útil para comparar modelos con distinto número de predictores.
- Si una nueva variable no mejora lo suficiente el ajuste, R^2_{adj} **disminuye**.

Interpretación

- Mide proporción de varianza explicada ajustada por complejidad.
- Más realista para modelos con muchos predictores.

R^2 Ajustado: Ejemplo

- Modelo 1: $R^2 = 0{,}80$, $R^2_{adj} = 0{,}79$.
- Modelo 2 (con 5 variables extra): $R^2 = 0{,}82$, $R^2_{adj} = 0{,}75$.
- Conclusión: el Modelo 1 es preferible, pese al mayor R^2 del Modelo 2.

Error Estándar de los Residuales: Definición

- Mide la **desviación promedio** de las predicciones respecto a los valores reales.
- Fórmula:

$$s = \sqrt{\frac{SS_{res}}{n - p - 1}}$$

donde:

- $n =$ número de observaciones,
- $p =$ número de predictores.

Error Estándar de los Residuales: Interpretación

- Valores más bajos $\Rightarrow$ mejor ajuste.
- Se expresa en las mismas unidades que y.
- Indica cuánto, en promedio, se desvían las predicciones de los valores observados.

Ejemplo

Si y es ingreso anual en miles de pesos y $s = 2{,}5$, en promedio el modelo se equivoca $\pm 2,500$ pesos.

Error Estándar de los Residuales: Uso Práctico

- Permite comparar modelos sobre la misma variable.
- Complementa R^2: dos modelos con igual R^2 pueden tener distinto s.
- Diagnóstico útil cuando la escala de y es relevante.

Prueba Global: Estadístico F

- Objetivo: verificar si el modelo es mejor que usar solo la media de y.
- Hipótesis:

$$H_0 : \beta_1 = \beta_2 = \cdots = \beta_p = 0$$

$$H_a : \text{Al menos un } \beta_j \neq 0$$

- Estadístico F:

$$F = \frac{(SS_{tot} - SS_{res})/p}{SS_{res}/(n-p-1)}$$

Interpretación de la Prueba F

- Si F es grande $\Rightarrow$ el modelo explica mucha más variabilidad que la media.
- Valor-p:
  - p-valor pequeño $\Rightarrow$ rechazamos H_0.
  - Conclusión: el modelo es significativo en conjunto.
- Limitación: no indica qué variables son responsables del efecto.

Ejemplo: Prueba F

- Modelo lineal con $p = 3$ predictores, $n = 50$ datos.
- Resultado: $F = 12{,}4$, p-valor $< 0{,}001$.
- Conclusión: el modelo es significativamente mejor que usar solo la media.

Pruebas t para Coeficientes

- Objetivo: evaluar si cada predictor contribuye significativamente.
- Hipótesis:

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

- Estadístico t:

$$t = \frac{\hat{\beta}_j}{SE(\hat{\beta}_j)}$$

Interpretación de la Prueba t

- Si $|t|$ es grande $\Rightarrow$ evidencia contra H_0.
- Valor-p:
  - p-valor pequeño $\Rightarrow$ el predictor j es significativo.
- Limitación: múltiples pruebas aumentan el riesgo de falsos positivos.

Ejemplo: Pruebas t

- Predictor: nivel educativo en un modelo de ingresos.
- Resultado: $\hat{\beta}_{edu} = 2{,}3,\ SE = 0{,}7,\ t = 3{,}29,\ p = 0{,}002$.
- Conclusión: el nivel educativo tiene un efecto significativo sobre ingresos.

Intervalos de Confianza para Coeficientes

- Definición: rango plausible para el valor verdadero de β_j.
- Fórmula:

$$IC_{(1-\alpha)}: \quad \hat{\beta}_j \pm t_{\alpha/2,\, n-p-1} \cdot SE(\hat{\beta}_j)$$

- Nivel común: 95 %.

Interpretación de Intervalos de Confianza

- Si el intervalo incluye 0 $\rightarrow$ no hay evidencia fuerte de efecto.
- Si el intervalo excluye 0 $\rightarrow$ predictor significativo.
- Relación con prueba t: equivalentes en su conclusión al mismo nivel α.

Ejemplo

$\beta_1 \in [0{,}5,\ 1{,}2] \Rightarrow$ significativo. $\beta_2 \in [-0{,}3,\ 0{,}8] \Rightarrow$ no significativo.

Evaluación Bayesiana de Coeficientes

- En lugar de pruebas F o t, usamos la **distribución posterior** de cada coeficiente β_j.
- Posterior:

$$p(\beta_j \mid y) \propto p(y \mid \beta_j)\, p(\beta_j)$$

- Pregunta clave:

$$P(\beta_j \approx 0 \mid y)$$

- Si la mayor parte de la masa posterior está lejos de 0, concluimos que β_j es importante.

Probabilidad de Efecto Nulo vs. Diferente de Cero

- Podemos calcular:

$$P(\beta_j > 0 \mid y) \quad \text{o} \quad P(\beta_j < 0 \mid y)$$

- En vez de un p-valor, obtenemos probabilidades directas.
- También usamos intervalos creíbles (HPD):

$$CI_{95\,\%} = [a, b] \quad \text{tal que } P(\beta_j \in [a, b] \mid y) = 0{,}95$$

- Interpretación:
  - Si 0 está fuera del intervalo $\rightarrow$ β_j es significativo.
  - Si 0 está dentro $\rightarrow$ evidencia débil.

AIC: Definición

- Criterio propuesto por Hirotugu Akaike (1973).
- Mide el compromiso entre:
  - **Bondad de ajuste**: qué tan bien explica los datos.
  - **Complejidad**: número de parámetros k.
- Fórmula:

$$AIC = 2k - 2\ln(L)$$

donde L = verosimilitud máxima.

AIC: Interpretación

- Menor AIC $\Rightarrow$ mejor modelo.
- Comparación relativa:

$$\Delta AIC_i = AIC_i - \min(AIC)$$

- Regla práctica:
  - $\Delta AIC \leq 2$: modelos comparables.
  - $4 \leq \Delta AIC \leq 7$: evidencia moderada contra el modelo.
  - $\Delta AIC > 10$: modelo muy poco probable.

AIC: Uso Práctico

- Útil para comparar modelos no anidados.
- Ampliamente usado en selección de modelos (ej. ecología, biología).
- Limitaciones:
  - No busca el "verdadero modelo", sino el que mejor predice.
  - Puede favorecer modelos más complejos.

BIC: Definición

- También conocido como criterio de Schwarz (1978).
- Relacionado con la aproximación bayesiana al cálculo de evidencias.
- Fórmula:

$$BIC = \ln(n)k - 2\ln(L)$$

donde $n =$ tamaño de muestra, $k =$ número de parámetros.

BIC: Interpretación

- Penaliza más fuertemente la complejidad que AIC.
- Menor BIC $\Rightarrow$ mejor modelo.
- Con n grande, favorece modelos más simples.
- Basado en aproximaciones asintóticas bayesianas al Bayes Factor.

BIC: Uso Práctico

- Adecuado cuando se busca identificar el modelo "verdadero" entre candidatos.
- Muy usado en contextos de inferencia estadística más que de predicción.
- Regla práctica de comparación:
  - $\Delta BIC > 10$: evidencia muy fuerte contra el modelo.

AIC vs BIC: Comparación

- Ambos penalizan modelos complejos, pero de forma distinta:

$$AIC: 2k \qquad BIC: \ln(n)k$$

- Diferencias clave:
  - **AIC**: Mejor para predicción (elige modelos que generalizan).
  - **BIC**: Mejor para identificación del modelo verdadero.
- Regla general:
  - AIC tiende a elegir modelos más complejos.
  - BIC tiende a elegir modelos más simples.

Evaluación de la Regresión Logística: Similitudes y Diferencias

- **Similitudes con la regresión lineal**:
  - Buscamos explicar/predicir una variable respuesta a partir de predictores.
  - Podemos evaluar ajuste global y significancia de coeficientes.
  - Usamos criterios de información (AIC, BIC) para comparar modelos.
- **Diferencias clave**:
  - La respuesta Y es binaria (0/1), no continua.
  - No existen R^2 ni supuestos de normalidad/homocedasticidad de residuales.
  - La inferencia se basa en la **verosimilitud**, no en cuadrados mínimos.
- **Herramientas específicas**:
  - **Bondad de ajuste**: Log-verosimilitud, pruebas de razón de verosimilitud, pseudo-R^2.
  - **Clasificación**: matriz de confusión, sensibilidad, especificidad, precisión, exactitud.
  - **Capacidad discriminativa**: curva ROC y AUC.
  - **Calibración**: prueba Hosmer–Lemeshow.

Log-Verosimilitud en Regresión Logística

- La estimación se basa en la verosimilitud:

$$L(\beta) = \prod_{i=1}^{n} \pi(x_i)^{y_i} \left[1 - \pi(x_i)\right]^{1-y_i}$$

  donde $\pi(x_i) = \frac{e^{x_i\beta}}{1+e^{x_i\beta}}$.
- Se trabaja con la **log-verosimilitud**:

$$\ell(\beta) = \sum_{i=1}^{n} \Big[y_i \ln(\pi(x_i)) + (1 - y_i) \ln(1 - \pi(x_i)) \Big]$$

- Valores más altos (menos negativos) indican mejor ajuste.

Prueba de Razón de Verosimilitud

- Compara un modelo completo con un modelo reducido (ej. nulo).
- Estadístico:

$$G^2 = -2\Big(\ell_{\text{reducido}} - \ell_{\text{completo}}\Big)$$

- Bajo H_0: $G^2 \sim \chi^2_{df}$.
- p-valor pequeño $\rightarrow$ el modelo completo mejora significativamente al nulo.

Matriz de Confusión

	Predicho 0	Predicho 1
Real 0	TN	FP
Real 1	FN	TP

- Resume los aciertos y errores de clasificación.

Métricas Derivadas

- **Precisión**: $\frac{TP}{TP+FP}$
- **Sensibilidad (Recall)**: $\frac{TP}{TP+FN}$
- **Especificidad**: $\frac{TN}{TN+FP}$
- **Exactitud**: $\frac{TP+TN}{N}$

Punto de corte

Por defecto $= 0.5$, pero puede ajustarse según el balance entre sensibilidad y especificidad.

Curva ROC: Concepto

- ROC = Receiver Operating Characteristic.
- Compara la capacidad de un modelo para distinguir entre clases 0 y 1.
- Se basa en dos métricas:

$$\text{Sensibilidad} = \frac{TP}{TP + FN}$$

$$\text{1 - Especificidad} = \frac{FP}{FP + TN}$$

- Ejes del gráfico:
  - x = tasa de falsos positivos (1 - especificidad).
  - y = tasa de verdaderos positivos (sensibilidad).

Curva ROC: Construcción Paso a Paso

1. El modelo produce probabilidades $\hat{p}_i$ para cada observación.
2. Elegimos un **punto de corte** (ej. 0.5).
3. Clasificamos:
   - $\hat{p}_i \geq 0{,}5 \Rightarrow$ clase 1.
   - $\hat{p}_i < 0{,}5 \Rightarrow$ clase 0.
4. Calculamos:
   - Sensibilidad.
   - Especificidad.
5. Obtenemos un punto en el plano ROC.

De un Punto a la Curva Completa

- Repetimos el proceso para muchos puntos de corte (ej. 0.1, 0.2, ..., 0.9).
- Cada corte genera un par (FPR, TPR).
- Conectamos los puntos para formar la curva ROC.

Referencias visuales

- Línea diagonal = clasificador aleatorio.
- Curva por encima de la diagonal = modelo con poder discriminativo.

ROC paso a paso: 1) Datos y probabilidades

- Supón un modelo logístico que entrega probabilidades $\hat{p}_i = P(Y = 1 \mid x_i)$.
- Tenemos 10 observaciones, con 5 positivos (1) y 5 negativos (0).
- Ordenamos por $\hat{p}$ de mayor a menor (esto facilita recorrer umbrales).

ID	$\mathbf{y}$ (real)	$\hat{\mathbf{p}}$ (pred.)
1	1	0.95
2	1	0.85
3	0	0.80
4	1	0.70
5	0	0.60
6	1	0.55
7	0	0.40
8	0	0.30
9	1	0.20
10	0	0.10

$\Rightarrow$ Totales: Positivos $= 5$, Negativos $= 5$. Recorremos umbrales t desde 1.00 $\downarrow$ 0.00 y en cada t:

$$\hat{y}_i(t) = \mathbb{I}\{\hat{p}_i \geq t\}, \quad \text{TPR}(t) = \frac{TP}{TP + FN}, \quad \text{FPR}(t) = \frac{FP}{FP + TN}.$$

ROC paso a paso: 2) Primeros umbrales, conteos y tasas

- $\mathbf{t > 0{,}95}$ (ningún caso predicho 1): $TP = 0, FP = 0, FN = 5, TN = 5 \Rightarrow \text{TPR} = 0/5 = 0$, $\text{FPR} = 0/5 = 0$.
- $\mathbf{t = 0{,}95}$: predichos 1 $= \{\text{ID 1}\,(y = 1)\} \Rightarrow TP = 1, FP = 0, FN = 4, TN = 5 \Rightarrow$ $\text{TPR} = 1/5 = 0{,}20$, $\text{FPR} = 0$.
- $\mathbf{t = 0{,}85}$: predichos 1 $= \{1(1), 2(1)\} \Rightarrow TP = 2, FP = 0 \Rightarrow \text{TPR} = 0{,}40$, $\text{FPR} = 0$.
- $\mathbf{t = 0{,}80}$: predichos 1 $= \{1(1), 2(1), 3(0)\} \Rightarrow TP = 2, FP = 1 \Rightarrow \text{TPR} = 0{,}40$, $\text{FPR} = 1/5 = 0{,}20$.
- $\mathbf{t = 0{,}70}$: añade 4(1) $\Rightarrow TP = 3, FP = 1 \Rightarrow \text{TPR} = 0{,}60$, $\text{FPR} = 0{,}20$.

$\mathbf{t}$	Pred. 1	TP	FP	FN	TN	(FPR, TPR)
$> 0{,}95$	$\emptyset$	0	0	5	5	(0.00, 0.00)
0.95	{1}	1	0	4	5	(0.00, 0.20)
0.85	{1,2}	2	0	3	5	(0.00, 0.40)
0.80	{1,2,3}	2	1	3	4	(0.20, 0.40)
0.70	{1,2,3,4}	3	1	2	4	(0.20, 0.60)

$\Rightarrow$ Cada fila es un **punto ROC**: eje $x =$ FPR, eje $y =$ TPR.

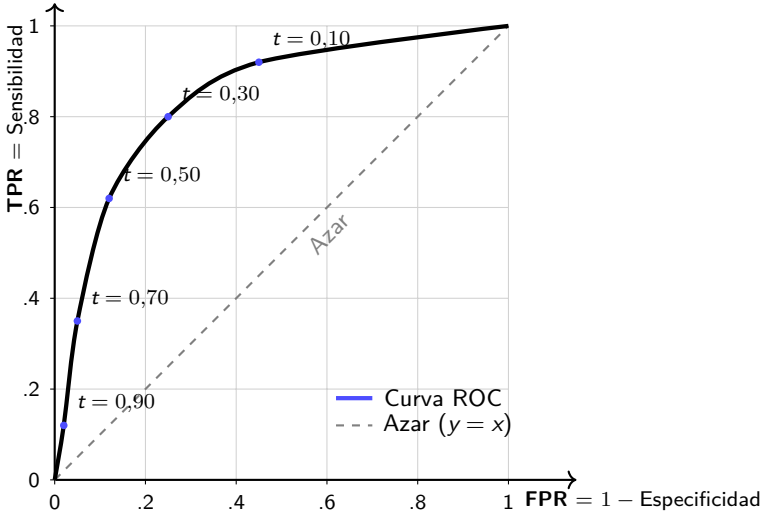
ROC paso a paso: 3) Todos los puntos y la curva

- Continuamos bajando t por cada $\hat{p}$ único:

$\mathbf{t}$	FPR	TPR
$> 0{,}95$	0.00	0.00
0.95	0.00	0.20
0.85	0.00	0.40
0.80	0.20	0.40
0.70	0.20	0.60
0.60	0.40	0.60
0.55	0.40	0.80
0.40	0.60	0.80
0.30	0.80	0.80
0.20	0.80	1.00
0.10	1.00	1.00

- **Cómo trazar**: Grafica los pares (FPR, TPR) en el plano y únelos en orden de t descendente.
- **Referencia**: la línea diagonal $y = x$ es el desempeño aleatorio.
- **Lectura**: curvas más "al noroeste" (mayor TPR con menor FPR) indican mejor discriminación.

Curva ROC: Visualización



Nota: Más arriba/izquierda ⇒ mejor discriminación (mayor TPR con menor FPR).

Área Bajo la Curva (AUC)

- AUC = probabilidad de que el modelo asigne mayor probabilidad a un positivo que a un negativo.
- Valores típicos:
  - 0.5 $\rightarrow$ modelo sin discriminación (equivalente a azar).
  - 0.7–0.8 $\rightarrow$ aceptable.
  - 0.8–0.9 $\rightarrow$ excelente.
  - $>$0.9 $\rightarrow$ sobresaliente.
- Ventaja: independiente del punto de corte elegido.

Bibliografía

- Draper, N. R., & Smith, H. (1998). *Applied Regression Analysis* (3rd ed.). Wiley.
- Weisberg, S. (2005). *Applied Linear Regression* (3rd ed.). Wiley.
- Montgomery, D. C., Peck, E. A., & Vining, G. G. (2021). *Introduction to Linear Regression Analysis* (6th ed.). Wiley.
- Kutner, M. H., Nachtsheim, C. J., & Neter, J. (2004). *Applied Linear Regression Models* (4th ed.). McGraw-Hill/Irwin.
- Akaike, H. (1974). A new look at the statistical model identification. *IEEE Transactions on Automatic Control*, 19(6), 716–723.
- Schwarz, G. (1978). Estimating the dimension of a model. *The Annals of Statistics*, 6(2), 461–464.
- Burnham, K. P., & Anderson, D. R. (2002). *Model Selection and Multimodel Inference: A Practical Information-Theoretic Approach* (2nd ed.). Springer.
- Belsley, D. A., Kuh, E., & Welsch, R. E. (2005). *Regression Diagnostics: Identifying Influential Data and Sources of Collinearity*. Wiley.
- Hosmer, D. W., Lemeshow, S., & Sturdivant, R. X. (2013). *Applied Logistic Regression* (3rd ed.). Wiley.