MODELOS DE REGRESIÓN PARA DATOS CATEGÓRICOS Y DE CONTEO

A continuación, nos vamos a centrar en los modelos de regresión cuya variable objetivo toma un **número reducido de valores** (puede ser cuantitativa discreta o cualitativa).

Dependiendo de los **valores que tome dicha variable** hablaremos de modelos binarios (e.g. un individuo que debe decidir si compra o no un producto), multinomiales (e.g. un elector que debe decidir a qué partido vota en unas elecciones), ordinales (e.g. la frecuencia con la que un individuo realiza deporte) o de conteo (e.g. número de hijos).

Estos modelos se enmarcan dentro del denominado **modelo lineal generalizado**, que no es más que una generalización del modelo de regresión lineal, que acabamos de estudiar, aplicable cuando la variable objetivo toma valores restringidos.

De nuevo, partimos de la ecuación ya conocida de los modelos supervisados, que persiguen **predecir** una variable Y a partir de un conjunto de p variables X_i :

$$Y = g(X_1, X_2, \ldots, X_p) + \epsilon,$$

donde ϵ recoge todos los factores que no pueden ser modelizados y que se asumen aleatorios y $g(\cdot)$ es cualquier función matemática.

El **modelo lineal generalizado** (GLM) parte de la ecuación anterior pero con las siguientes asunciones (lo mismo ocurre con la regresión lineal, pues es un caso particular de este modelo):

- Se modeliza el **comportamiento medio** de la Y, ignorando indirectamente la componente aleatoria, pues se asume que su esperanza es 0.
- Se asume que la variable objetivo tiene una **distribución conocida concreta**, que varía en función del tipo de la variable.
- El efecto de las variables explicativas se puede resumir con una **combinación lineal** de todas ellas.

Por lo que la ecuación anterior se reescribe como (con $f(\cdot)$ denominada **función de enlace**):

$$E[Y|X_1, X_2, \ldots, X_p] = f(\beta_0 + \beta_1 X_1 + \cdots + \beta_p X_p)$$

Por todo lo anterior, el GLM tiene las siguientes característica:

- Todas las variables explicativas deben estar **codificadas con números** (pues no se podría calcular la combinación lineal de otra manera). La forma habitual de proceder con las variables cualitativas es a través de variables dummy (codificación *one-hot*).
- Se asume que la **relación bivariante** entre la variable objetivo y las explicativas es monótona y viene dada por la función de enlace.
- No puede haber **datos ausentes** en las variables pues, de lo contrario, las observaciones con algún dato faltante son ignoradas completamente.
- La presencia de **datos atípicos** tiene un gran impacto en los resultados.
- La presencia de **multicolinealidad** aumenta la variabilidad de los estimadores de los parámetros, reduciendo su credibilidad.

En el caso de la **regresión lineal** clásica, la variable Y es **continua**, se asume que sigue una distribución normal y $f(\cdot)$ es la función identidad, por lo que el modelo queda así:

$$E[Y|X_1, X_2, \ldots, X_p] = \beta_0 + \beta_1 X_1 + \cdots + \beta_p X_p.$$

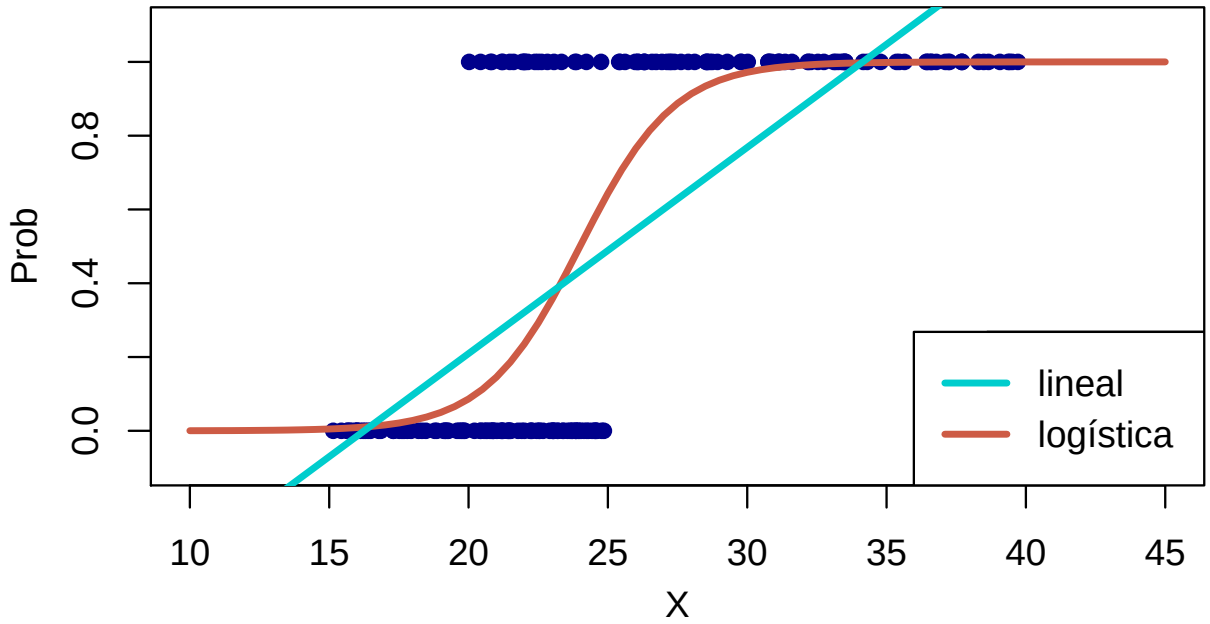
Cuando la variable dependiente es **binaria** (es decir, se puede codificar como $Y = 1$ ó $Y = 0$), es habitual asumir que la distribución de la misma es **Bernouilli** por lo que:

$$E[Y|X_1, \ldots, X_p] = P(Y = 1|X_1, \ldots, X_p) = f(\beta_0 + \beta_1 X_1 + \cdots + \beta_p X_p).$$

En ese caso, asumir una función $f(\cdot)$ identidad tiene el inconveniente de que no asegura dar como resultados **valores en el intervalo** $(0, 1)$, incumpliendo la definición de probabilidad.

Por ello, la función de enlace más habitual es la **función de distribución logística**:

$$P(Y = 1|X_1, X_2, \ldots, X_p) = \frac{e^{(\beta_0+\beta_1 X_1+\cdots+\beta_p X_p)}}{1 + e^{(\beta_0+\beta_1 X_1+\cdots+\beta_p X_p)}}$$



El principal motivo por el cual la función de enlace más habitual es la logística es que **los parámetros β pueden interpretarse**, aportando información sobre el efecto que tienen las variables independientes sobre las probabilidades a través de los ***odds ratio***.

INTERPRETACIÓN DE LOS PARÁMETROS: VARIABLE REGRESORA DICOTÓMICA

Para interpretar correctamente el parámetro de una variable regresora dicotómica debemos tener en cuenta que la **inclusión de variables cualitativas** se lleva a cabo a partir de la **creación de variables *dummy***.

Supongamos un modelo univariante donde la única variable regresora sea **dicotómica**. Calculamos los ***odds*** de aquellos individuos para los que $x = 1$ y para los que $x = 0$:

$$odds(evento|x = 1) = \frac{P(evento|x = 1)}{1 - P(evento|x = 1)} = \frac{\frac{e^{(\beta_0+\beta_1)}}{1+e^{(\beta_0+\beta_1)}}}{\frac{1}{1+e^{(\beta_0+\beta_1)}}} = e^{\beta_0+\beta_1}$$

$$odds(evento|x = 0) = \frac{P(evento|x = 0)}{1 - P(evento|x = 0)} = \frac{\frac{e^{\beta_0}}{1+e^{\beta_0}}}{\frac{1}{1+e^{\beta_0}}} = e^{\beta_0}$$

El ***odds ratio*** se define como el cociente entre los *odds* con $x = 1$ y los *odds* con $x = 0$:

$$OR = \frac{odds(evento|x = 1)}{odds(evento|x = 0)} = \frac{e^{\beta_0+\beta_1}}{e^{\beta_0}} = e^{\beta_1}$$

INTERPRETACIÓN DE LOS PARÁMETROS: VARIABLE REGRESORA DICOTÓMICA

Por lo tanto, la **exponencial del parámetro** representa el *odds ratio* asociado a la variable y, en consecuencia, permite aproximar **cuánto más probable** (o improbable) es que se de el **evento** entre los individuos con $x = 1$ frente a los individuos con $x = 0$.

INTERPRETACIÓN DE LOS PARÁMETROS: VARIABLE REGRESORA CATEGÓRICA

Cuando se desea analizar el efecto de una variable regresora categórica es necesario tener en cuenta que el modelo tendrá **tantos parámetros β como categorías menos uno**.

La interpretación de cada uno de estos parámetros **coincide con la de las variables dicotómicas** pero teniendo en cuenta que la comparación a la que se refiere el *odds ratio* es la de la **categoría correspondiente y la de referencia**.

INTERPRETACIÓN DE LOS PARÁMETROS: VARIABLE REGRESORA CONTINUA

En el caso de variables regresoras continuas, la exponencial del parámetro coincide con el *odds ratio* asociado al **efecto de un incremento unitario** en la variable:

$$OR = \frac{odds(evento|x = a + 1)}{odds(evento|x = a)} = \frac{e^{\beta_0 + \beta_1 (a+1)}}{e^{\beta_0 + \beta_1 a}} = e^{\beta_1}$$

ESTIMACIÓN DE LOS PARÁMETROS

Para estimar los parámetros del modelo se utiliza el método de **máxima verosimilitud**, que consiste en determinar los valores β_j que maximizan la correspondiente función de verosimilitud (calculada considerando que Y se distribuye según una Bernouilli):

$$\mathcal{L}(\boldsymbol{\beta}) = \prod_{i=1}^{n} p_{1i}^{y_i}(1-p_{1i})^{1-y_i} = \prod_{i/y_i=1} p_{1i} \prod_{j/y_j=0} (1-p_{1j}),$$

donde p_{1i} representa la **probabilidad del evento** para la observación i, dada por:

$$p_{1i} = P(Y=1|x_{1i}, x_{2i}, \ldots, x_{pi}) = \frac{e^{(\beta_0+\beta_1 x_{1i}+\cdots+\beta_p x_{pi})}}{1+e^{(\beta_0+\beta_1 x_{1i}+\cdots+\beta_p x_{pi})}}.$$

No existe una fórmula explicita para la obtención de los parámetros que maximizan la verosimilitud, por lo que será necesario recurrir a **métodos iterativos de optimización**.

ANÁLISIS DEL MODELO

- **Contrastes de los parámetros**: permiten determinar si cada uno de los parámetros del modelo son significativamente distintos de 0 ($H_0 : \beta_i = 0$ vs. $H_1 : \beta_i \neq 0$).

  Dado que estos se obtienen por el método de máxima verosimilitud, se tiene que son **asintóticamente normales** (sin necesidad de hipótesis adicionales). Como, además, su varianza es desconocida y debe estimarse, se puede utilizar el correspondiente *test t*.

- **Análisis de tipo II**: en este caso, para contrastar la utilidad de las variables independientes del modelo, se cuantifica **la disminución de la verosimilitud del modelo** debida a la eliminación de cada una de ellas.
  - En particular, cuando esta cantidad es **pequeña**, la variable **carece de poder predictivo** pues el modelo no empeora al eliminarla del mismo.
  - Se trata de un contraste en el que se compara el modelo construido con un **modelo simplificado** en el que se elimine únicamente la variable en cuestión.
  - Se recurre al **contraste de razón de verosimilitudes**, basado en el estadístico $\lambda = -2 \ln \left(\frac{\mathcal{L}(\text{modelo sin})}{\mathcal{L}(\text{modelo con})} \right) = -2 \left(\ln(\mathcal{L}(\text{modelo sin})) - \ln(\mathcal{L}(\text{modelo con})) \right)$ con distribución χ^2 con tantos g.l. como parámetros tenga la variable.
  - Adicionalmente, la magnitud de estas reducciones en la verosimilitud permiten generar una **ordenación de las variables** por su importancia.

EVALUACIÓN DEL MODELO

- **Matriz de confusión**: una vez obtenidas las probabilidades a través del modelo, se pueden **clasificar las observaciones** como eventos (1) o no eventos (0) en función de si su probabilidad es superior/inferior a cierto **punto de corte** (generalmente $\mathbf{0.5}$ pero, en variables desbalanceadas es preferible recurrir a la proporción de eventos).

  Una vez obtenida esta clasificación, se puede enfrentar a la clasificación original para determinar los **aciertos y errores** cometidos:

	Predicción = 0	Predicción = 1
Realidad = 0	**VN** Verdadero negativo	**FP** Falso positivo
Realidad = 1	**FN** Falso negativo	**VP** Verdadero positivo

Utilizando esta información, se pueden definir las siguientes **medidas de clasificación**:

Tasa de acierto: $\frac{VN+VP}{VN+FP+FN+VP}$ Tasa de fallo: $\frac{FP+FN}{VN+FP+FN+VP}$

Sensibilidad: $\frac{VP}{FN+VP}$ Especificidad: $\frac{VN}{VN+FP}$

EVALUACIÓN DEL MODELO

- **Índice Kappa**: este índice surge como una **alternativa a la tasa de acierto** (acc, por su nombre en inglés *accuracy*) para reducir la influencia de la frecuencia relativa del "evento", para lo cual se elimina el efecto de los aciertos que se pueden producir al azar.

  Sean $p_{i\cdot}$ y $p_{\cdot i}$ las proporciones de observaciones en la categoría i real y predicha, respectivamente. La proporción de **aciertos que ocurren al azar** en la categoría i vienen dados por el producto de dichas cantidades y, por ende, la proporción total de aciertos debidos al azar vendrá dado por la suma de todas las categorías:

$$\kappa = \frac{acc - \sum_{i=1}^{k} p_{i\cdot} p_{\cdot i}}{1 - \sum_{i=1}^{k} p_{i\cdot} p_{\cdot i}}$$

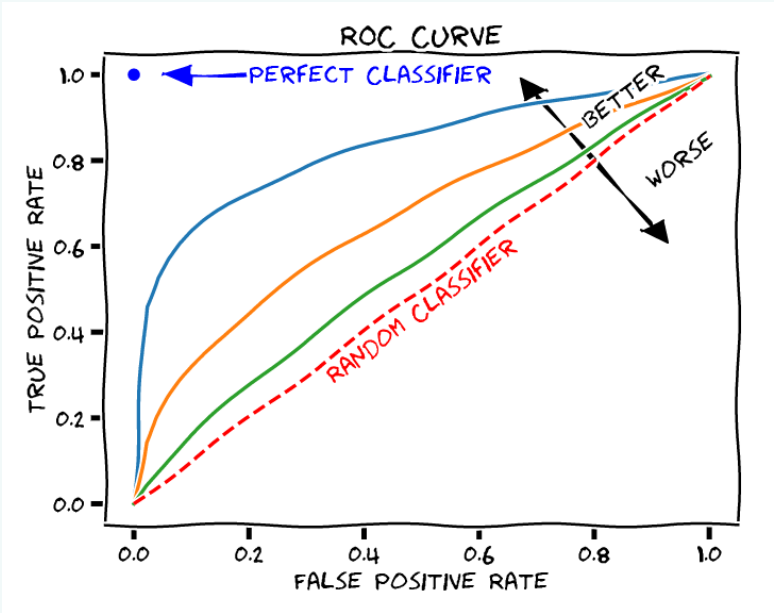
  La siguiente tabla contiene una guía sobre la **interpretación de este índice**:

Índice Kappa	Interpretación
0	Equivalente al azar
0.01 - 0.20	Pobre
0.21 - 0.40	Justo
0.41 - 0.60	Moderado
0.61 - 0.80	Bueno
0.81 - 1	Excelente

EVALUACIÓN DEL MODELO

- **Curva ROC**: El principal inconveniente de las medidas derivadas de la matriz de confusión es que **dependen del punto de corte aplicado a la probabilidad** para clasificar las observaciones como evento o no evento. La Curva ROC permite obtener una medida de evaluación del modelo **sin necesidad de fijar este punto de corte**.

  Para ello, se grafica la **sensibilidad frente a 1 - especificidad** derivadas de todos los posibles puntos de corte y las correspondientes clasificaciones que se obtienen mediante las probabilidades predichas.



Teniendo en cuenta que un modelo **perfecto** es aquel con una sensibilidad y especificidad del 100 %, cuanto **más cóncava** sea esta curva, **mejor** será el modelo que se está evaluando.

En ocasiones es **dificil discernir** visualmente qué curva es más cóncava, para lo que se define el **área bajo la curva** ROC (AUC) como medida de evaluación del modelo. Un modelo **perfecto tendrá un AUC de 1**, mientras que un “mal” modelo tendrá un AUC de 0.5.

Como en el modelo de regresión lineal, la **combinación las variables independientes** que componen los modelos de regresión logística binaria determina la calidad de los mismos.

Para determinar dicha combinación, de nuevo recurrimos a los ya estudiados **métodos automáticos de selección de variables**:

- **Backward o hacia atrás**: Partiendo del modelo que contiene todos los posibles efectos, ir eliminando **una a una** las variables que menos influyan en el modelo hasta que el modelo empeore con la eliminación de cualquiera de las variables restantes.
- **Forward o hacia delante**: Partiendo desde cero, ir introduciendo **uno a uno** los efectos que mayor mejora produzcan en el modelo hasta que no haya ningún efecto más fuera del modelo que **aporte información**.
- **Stepwise o paso a paso**: Este método es una **mezcla** de los anteriores. El método es similar al forward, salvo por que **se pueden eliminar** los efectos que han entrado en el modelo (ya que la entrada de alguno puede hacer no significativo el aporte de otro). La eliminación de los efectos se hace de acuerdo al método backward.

CÁLCULO DE LA MEJORA EN LOS MODELOS DEL PROCESO ITERATIVO

Como ya se ha comentado previamente, desde el punto de vista del conjunto de datos de **entrenamiento**, cuantas **más variables** tiene un modelo, menor error comete. Por lo tanto, si evaluamos los modelos en dicho conjunto de datos en cada paso del proceso iterativo, la conclusión siempre será que el **mejor modelo** es el que más variables contiene.

Una alternativa sería evaluar los modelos que se van generando mediante **validación cruzada**. El problema de esta estrategia es que computacionalmente es bastante **intensa**, por lo que no se suele aplicar en este contexto.

Así pues, lo que se suele hacer en esos casos es recurrir a las métricas ya estudiadas, que, por un lado, evalúan la mejora producida en la verosimilitud (L) y, por otro, tienen en cuenta la mayor **complejidad del modelo**:

- **AIC** (Akaike information criterion): $-2\ln(L) + 2\tau$
- **BIC ó SBC** (Bayesian o Schwarz information criterion): $-2\ln(L) + \tau\ln(n)$

De nuevo, su diferencia radica en el segundo sumando y se reduce a la **penalización del número de parámetros** (τ), siendo el BIC/SBC el que más penaliza.

Como en el caso lineal, lo recomendable es generar las 6 posibles combinaciones de métodos y métricas para posteriormente comparar ese número más reducido de modelos con **validación cruzada repetida**.

De nuevo, en presencia de multicolinealidad y cuando el número de observaciones no es suficientemente grande, el enfoque anterior puede no ser válido a la hora de seleccionar variables. La mejor alernativa es la regresión penalizada/regularizada.

REGRESIÓN PENALIZADA

Los modelos de **regresión penalizada** se basan en la misma formulación del modelo de regresión logística binaria, salvo que los parámetros se obtienen como:

$$\boldsymbol{\beta} = \arg\max \left\{ \mathcal{L}(\boldsymbol{\beta}) - \lambda \|\boldsymbol{\beta}\| \right\},$$

donde $\mathcal{L}(\boldsymbol{\beta})$ es la función de verosimilitud ya vista, $\lambda \geq 0$ es el parámetro de penalización y $\|\boldsymbol{\beta}\|$ es una norma aplicada al vector de parámetros del modelo.

El modelo penalizado más utilizado es el llamado ***Elastic Net*** (ENet), que viene dado por:

$$\boldsymbol{\beta} = \arg\max \left\{ \mathcal{L}(\boldsymbol{\beta}) - \lambda \left(\alpha \sum_{i=1}^{t} |\beta_i| + (1-\alpha) \sum_{i=1}^{t} \beta_i^2 \right) \right\},$$

con t el número de parámetros del modelo saturado (sin contar con el término independiente) y $\alpha \in [0, 1]$.

Cuando $\alpha = 0$, el modelo recibe el nombre de ***Ridge*** (penalización cuadrática), mientras que si $\alpha = 1$, nos encontramos con el modelo ***LASSO*** (penalización absoluta). Valores intermedios proporcionan modelos con penalización mixta.

REGRESIÓN PENALIZADA

- Los modelos penalizados persiguen **reducir la magnitud de los parámetros** estimados y, por ende, el posible sobreajuste.
- La magnitud de la **penalización** viene dada por el valor del parámetro λ, dando lugar al modelo clásico si $\lambda = 0$ y al *no modelo* si $\lambda = \infty$.
- La **penalización absoluta** propicia una selección de variables "real", pues permite obtener coeficientes nulos (si la penalización es suficientemente grande).
- Las variables *input* han de **estandarizarse** para evitar que haya variables "dominantes" debido a la diferencia de escala. Esto hace que se pierda cierta interpretabilidad.
- Dependiendo de las características concretas de los datos, los **valores óptimos de λ y α** varían, por lo que se debe probar con varios valores y seleccionar el que mejores resultados ofrezca en la validación cruzada.
  - Este proceso recibe el nombre de afinación (aunque se suele recurrir al anglicismo *tuneo*).
  - Dado que la validación cruzada es un proceso aleatorio, para reducir aún más las posibilidades de sobreajuste, se puede recurrir a la **regla 1se** (*one standard error rule*), que consiste en seleccionar el modelo más sencillo que proporcione una métrica de calidad que se encuentre a menos de una desviación típica del óptimo.