UNIVERSITAS Miguel Hernández

Tema 2. Regresión Logística.

José L. Sainz-Pardo Auñón

TÉCNICAS ESTADÍSTICAS PARA EL APRENDIZAJE II

Máster Universitario en Estadística Computacional y Ciencia de Datos para la Toma de Decisiones.

Índice

¿Qué es la Regresión Logística?

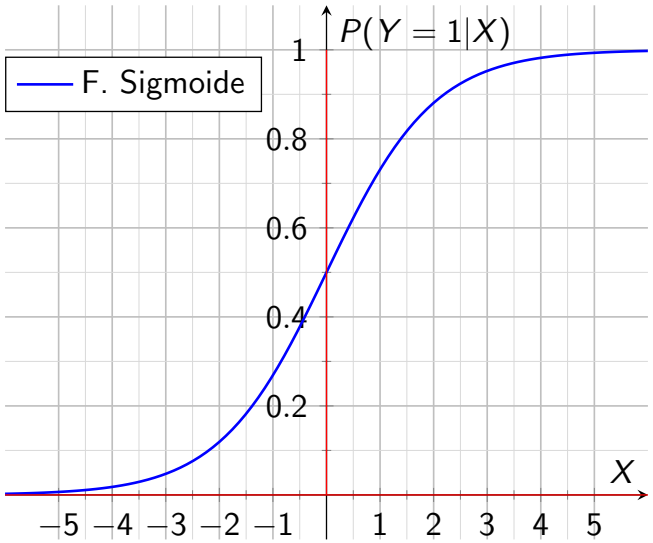
- La regresión logística es un modelo estadístico utilizado para predecir una variable dependiente binaria.
- Es una extensión de la regresión lineal para problemas de clasificación.
- La respuesta es categórica: $Y \in \{0, 1\}$.

La Función Sigmoide

La regresión logística utiliza la **función sigmoide** para modelar probabilidades:

$$P(Y = 1|X) = \frac{1}{1 + e^{-\beta_0 - \beta_1 X_1 - \cdots - \beta_n X_n}}$$

- Convierte cualquier valor real en un rango entre 0 y 1.
- Ideal para modelar probabilidades de eventos binarios.



Estimación de Parámetros: Máxima Verosimilitud

- Los parámetros $\beta_0, \beta_1, \ldots, \beta_n$ se estiman mediante el método de máxima verosimilitud.
- El objetivo es maximizar la probabilidad de observar los datos dados los parámetros.

$$L(\beta_0, \beta_1, \ldots, \beta_n) = \prod_{i=1}^{n} P(y_i|X_i; \beta)$$

- Aplicando logaritmos:

$$\ell(\beta_0, \beta_1, \ldots, \beta_n) = \sum_{i=1}^{n} [y_i \log(P(y_i|X_i; \beta)) + (1 - y_i) \log(1 - P(y_i|X_i; \beta)$$

Derivación de la Log-Verosimilitud

- Para estimar los parámetros óptimos, derivamos la función de log-verosimilitud con respecto a cada parámetro β_j.
- Para β_0:

$$\frac{\partial \ell(\beta)}{\partial \beta_0} = \sum_{i=1}^{n} (y_i - P(y_i|X_i;\beta))$$

- Para β_1:

$$\frac{\partial \ell(\beta)}{\partial \beta_1} = \sum_{i=1}^{n} (y_i - P(y_i|X_i;\beta))\, X_{i1}$$

- Igualamos a cero para obtener las ecuaciones de máxima verosimilitud:

$$\frac{\partial \ell(\beta)}{\partial \beta_j} = 0 \quad \text{para cada } \beta_j$$

Interpretación de los coeficientes

- Los coeficientes β_j en la regresión logística representan el cambio en el **logaritmo de las probabilidades (log-odds)** por unidad de cambio en la variable independiente X_j.
- El log-odds o función logit viene dado por:

$$\log\left(\frac{p_i}{1-p_i}\right) = \log\left(\frac{P(Y=1|X)}{1-P(Y=1|X)}\right) = \beta_0 + \beta_1 X_1 + \cdots + \beta_n X_n$$

Solución Numérica: Gradiente Descendente

- Las ecuaciones resultantes de la derivación son no lineales, lo que impide obtener soluciones analíticas directas.
- En su lugar, utilizamos métodos numéricos, como el gradiente descendente.
- El gradiente descendente ajusta los parámetros β_j iterativamente, reduciendo el error paso a paso.

Ejemplo Práctico: Predicción de Compra Basada en Ingreso

Datos de entrada:

Ingreso (X)	Compra (Y)
1	0
2	0
3	0
4	1
5	1

Queremos predecir $P(\text{Compra} = 1|\text{Ingreso})$ utilizando la fórmula de la regresión logística:

$$P(Y = 1|X) = \frac{1}{1 + e^{-(\beta_0+\beta_1 X)}}$$

Paso 1: Valores Iniciales y Probabilidades

Usamos $\beta_0 = 0$ y $\beta_1 = 0$ como suposiciones iniciales:

$$P(Y = 1|X) = \frac{1}{1 + e^0} = 0.5 \quad \text{(para todos los valores de } X\text{)}$$

Las probabilidades iniciales son:

$$\hat{P}(Y = 1|X) = [0.5, 0.5, 0.5, 0.5, 0.5]$$

Paso 2: Cálculo de los Gradientes

Calculamos los gradientes para actualizar los coeficientes:

$$\text{Gradiente de } \beta_0 = \sum_{i=1}^{5}(Y_i - \hat{P}_i) = -0.5$$

$$\text{Gradiente de } \beta_1 = \sum_{i=1}^{5}(Y_i - \hat{P}_i)X_i = 1.5$$

Paso 3: Actualización de Coeficientes

Con una tasa de aprendizaje $\alpha = 0.1$, actualizamos los coeficientes:

$$\beta_0^{(1)} = 0 + 0.1 \cdot (-0.5) = -0.05$$

$$\beta_1^{(1)} = 0 + 0.1 \cdot (1.5) = 0.15$$

Luego recalculamos las probabilidades usando estos nuevos coeficientes:

$$\hat{P}(Y = 1|X) = \left[\frac{1}{1 + e^{-(-0.05+0.15 \cdot X)}} \right]$$

Paso 4: Segunda Iteración

Ahora, con $\beta_0 = -0.05$ y $\beta_1 = 0.15$, recalculamos las probabilidades:

$$\hat{P}(Y = 1|X) = [0.47, 0.51, 0.55, 0.59, 0.63]$$

Calculamos los nuevos gradientes:

$$\text{Gradiente de } \beta_0 = \sum_{i=1}^{5}(Y_i - \hat{P}_i) = -0.35$$

$$\text{Gradiente de } \beta_1 = \sum_{i=1}^{5}(Y_i - \hat{P}_i)X_i = 1.2$$

Nuevas actualizaciones de los coeficientes:

$$\beta_0^{(2)} = -0.05 + 0.1 \cdot (-0.35) = -0.085$$

$$\beta_1^{(2)} = 0.15 + 0.1 \cdot (1.2) = 0.27$$

Regla de Parada en el Gradiente Descendente

- El proceso iterativo del gradiente descendente debe detenerse en algún punto.
- La regla de parada comúnmente utilizada es cuando el cambio en los coeficientes entre iteraciones consecutivas es muy pequeño.

$$|\beta_0^{(t+1)} - \beta_0^{(t)}| < \epsilon \quad \text{y} \quad |\beta_1^{(t+1)} - \beta_1^{(t)}| < \epsilon$$

- ϵ es un umbral pequeño, por ejemplo $\epsilon = 10^{-4}$.
- Alternativamente, se puede detener cuando la función de pérdida (log-verosimilitud) cambia muy poco entre iteraciones.

Supuestos de la Regresión Logística

- **Relación lineal en el logit**: La relación entre las variables independientes y el logit de la probabilidad debe ser lineal.
- **Independencia de observaciones**: Las observaciones deben ser independientes entre sí.
- **Ausencia de multicolinealidad**: No debe haber colinealidad significativa entre las variables predictoras.
- **Suficiente tamaño de muestra**: Se recomienda tener al menos 10 eventos por predictor (no como en el ejemplo anterior).
- **No hay valores extremos influyentes**: Los outliers no deben tener un impacto desproporcionado en el modelo.

Supuestos No Necesarios en Regresión Logística

En comparación con la regresión lineal, la regresión logística no requiere:

- **Homocedasticidad**: La varianza de los residuos no necesita ser constante.
- **Normalidad de los errores**: Los residuos no necesitan seguir una distribución normal.
- **Linealidad entre predictores y respuesta**: No se requiere que la relación sea lineal, ya que se modela el logit.

Cómo Validar los Supuestos de la Regresión Logística

Para validar los supuestos, se pueden realizar las siguientes acciones (sólo se enumeran, la mayoría caen fuera del objeto de esta asignatura):

- **Relación lineal en el logit**:
  - Gráficos de residuos devianza.
  - Box-Tidwell test para la relación logit.
- **Independencia de observaciones**:
  - Revisar el diseño del estudio.
  - Usar modelos jerárquicos si hay dependencia.
- **Ausencia de multicolinealidad**:
  - Calcular el Factor de Inflación de la Varianza (VIF).
  - $VIF > 10$ indica problemas de multicolinealidad.
- **Valores extremos influyentes**:
  - Análisis de la distancia de Cook.
  - Evaluar leverage.

Validación del Modelo: Tabla de Confusión

- La **tabla de confusión** permite evaluar el rendimiento de un modelo de clasificación.
- Muestra cuántas veces el modelo predice correctamente o falla al clasificar cada clase.

Ejemplo: Predicción de Compra Basada en Ingreso

Supongamos que usamos un umbral de decisión de 0.5 para clasificar si alguien hará una compra ($Y = 1$) o no ($Y = 0$). El modelo predice la clase positiva si la probabilidad estimada es mayor que 0.5.

Datos Predichos:

$$\hat{P}(Y = 1|X) = [0.47, 0.51, 0.55, 0.59, 0.63]$$

Predicciones finales (para un umbral $\theta = 0.5$):

$$Y_{\text{pred}} = [0, 1, 1, 1, 1]$$

Datos Reales:

$$Y = [0, 0, 0, 1, 1]$$

Tabla de Confusión para el Ejemplo

A continuación, presentamos la tabla de confusión correspondiente a este ejemplo:

	Compra Real = 1	Compra Real = 0
Compra Predicha = 1	2 (VP)	1 (FP)
Compra Predicha = 0	0 (FN)	2 (VN)

Definiciones:

- **Verdaderos Positivos (VP)**: La cantidad de casos predichos correctamente como positivos (compra).
- **Falsos Positivos (FP)**: La cantidad de casos predichos como positivos, pero que en realidad no ocurrieron (error tipo I).
- **Verdaderos Negativos (VN)**: Casos correctamente predichos como negativos (no compra).
- **Falsos Negativos (FN)**: Casos predichos como negativos, pero que en realidad ocurrieron (error tipo II).

Métricas Derivadas de la Tabla de Confusión

A partir de la tabla de confusión, podemos calcular varias métricas:

- **Precisión (Accuracy)**: Proporción de predicciones correctas.

$$\text{Precisión} = \frac{\text{VP} + \text{VN}}{\text{Total}} = \frac{2+2}{5} = 0.8$$

- **Precisión Positiva (Precision)**: Proporción de verdaderos positivos entre todas las predicciones positivas.

$$\text{Precisión Positiva} = \frac{\text{VP}}{\text{VP} + \text{FP}} = \frac{2}{2+1} = 0.67$$

- **Sensibilidad (Recall o Tasa de Verdaderos Positivos)**: Proporción de verdaderos positivos entre todas las instancias positivas reales.

$$\text{Sensibilidad} = \frac{\text{VP}}{\text{VP} + \text{FN}} = \frac{2}{2+0} = 1.0$$

- **Especificidad**: Proporción de verdaderos negativos entre todas las instancias negativas reales.

$$\frac{\text{VN}}{\quad} \quad 2$$

Evaluación del Modelo: Curva ROC

La curva **ROC** (Receiver Operating Characteristic) permite evaluar el rendimiento de un modelo de clasificación, como la regresión logística. Mide la capacidad del modelo para discriminar entre las clases positivas y negativas en diferentes umbrales de decisión.

- El eje **Y** representa la **Tasa Verdaderos Positivos** o **Sensibilidad**:

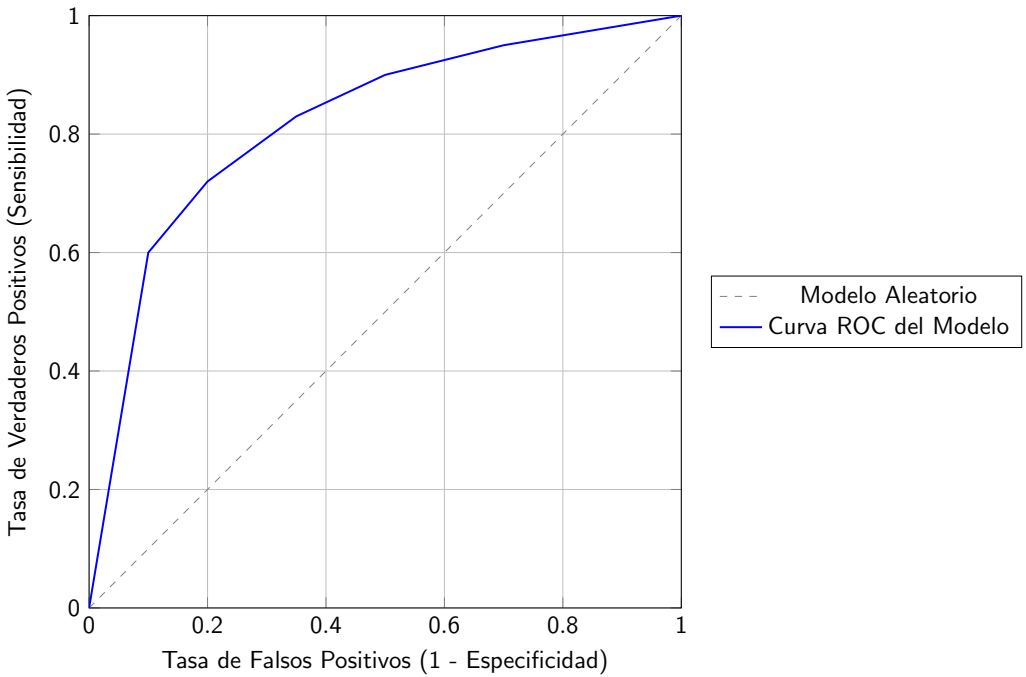
$$\text{Sensibilidad} = \frac{\text{VP}}{\text{VP} + \text{FN}}$$

- El eje **X** representa la **Tasa de Falsos Positivos**:

$$\text{Tasa Falsos Positivos} = \frac{\text{FP}}{\text{FP} + \text{VN}}$$

- Un modelo perfecto tiene una curva ROC que pasa por el punto (0,1), indicando que tiene una sensibilidad del 100% y una tasa de falsos positivos del 0%.
- El área bajo la curva (AUC) mide el rendimiento general: cuanto más cercano a 1, mejor es el modelo.

Curva ROC: Ejemplo



- Se ha trazado la curva para los umbrales $\theta = \{0.5, 0.6, 0.7, 0.8, 0.9\}$
- El área bajo la curva (**AUC**) es aproximadamente 0.9, lo que indica que el modelo tiene un buen desempeño.
- El modelo tiene una alta tasa de verdaderos positivos y una baja tasa de falsos positivos.

Conclusiones

- La regresión logística permite modelar variables dependientes binarias y por tanto es una **técnica de clasificación**.
- El método de máxima verosimilitud permite estimar los coeficientes del modelo (trata de hallar los betas más probables).
- La validación de supuestos no es tan exigente como en Regresión Lineal.