Notas sobre Regresión Logística
Breve introducción con aplicaciones

Carlos E Martinez-Rodriguez

Contents

1 Análisis de Regresion Lineal (RL)

En muchos problemas hay dos o más variables relacionadas, para medir el grado de relación se utiliza el **análisis de regresión**. Supongamos que se tiene una única variable dependiente y, y varias variables independientes, $x_1, x_2, \ldots, x_n$. La variable y es una varaible aleatoria, y las variables independientes pueden ser distribuidas independiente o conjuntamente. A la relación entre estas variables se le denomina modelo regresión de y en $x_1, x_2, \ldots, x_n$, por ejemplo $y = \phi(x_1, x_2, \ldots, x_n)$, lo que se busca es una función que aproxime a $\phi(\cdot)$.

Supongamos que de momento solamente se tienen una variable independiente x, para la variable de respuesta y. Y supongamos que la relación que hay entre x y y es una línea recta, y que para cada observación de x, y es una variable aleatoria. El valor esperado de y para cada valor de x es

$$E(y|x) = \beta_0 + \beta_1 x, \tag{1}$$

β_0 es la ordenada al orígen y β_1 la pendiente de la recta en cuestión, ambas constantes desconocidas.

Supongamos que cada observación y se puede describir por el modelo

$$y = \beta_0 + \beta_1 x + \epsilon \tag{2}$$

donde ϵ es un error aleatorio con media cero y varianza σ^2. Para cada valor y_i se tiene ϵ_i variables aleatorias no correlacionadas, cuando se incluyen en el modelo 2, este se le llama *modelo de regresión lineal simple*. Supongamos además que se tienen n pares de observaciones $(x_1, y_1), (x_2, y_2), \ldots, (x_n, y_n)$, estos datos pueden utilizarse para estimar los valores de β_0 y β_1. Esta estimación se realiza utilizando el **métodos de mínimos cuadrados**. Entonces la ecuación (2) se puede reescribir como

$$y_i = \beta_0 + \beta_1 x_i + \epsilon_i, \text{ para } i = 1, 2, \ldots, n. \tag{3}$$

Si consideramos la suma de los cuadrados de los errores aleatorios, es decir, el cuadrado de la diferencia entre las observaciones con la recta de regresión

$$L = \sum_{i=1}^{n} \epsilon^2 = \sum_{i=1}^{n} (y_i - \beta_0 - \beta_1 x_i)^2 .$$

Para obtener los estimadores por mínimos cuadrados de β_0 y β_1, $\hat{\beta}_0$ y $\hat{\beta}_1$, es preciso calcular las derivadas parciales con respecto a β_0 y β_1, igualar a cero y resolver el sistema de ecuaciones lineales resultante:

$$\begin{aligned} \frac{\partial L}{\partial \beta_0} &= 0; \\ \frac{\partial L}{\partial \beta_1} &= 0, \end{aligned}$$

Evaluando en $\hat{\beta}_0$ y $\hat{\beta}_1$, se tiene

$$\begin{aligned} -2 \sum_{i=1}^{n} \left(y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i \right) &= 0, \\ -2 \sum_{i=1}^{n} \left(y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i \right) x_i &= 0. \end{aligned}$$

Simplificando

$$\begin{aligned} n\hat{\beta}_0 + \hat{\beta}_1 \sum_{i=1}^{n} x_i &= \sum_{i=1}^{n} y_i \\ \hat{\beta}_0 \sum_{i=1}^{n} x_i + \hat{\beta}_1 \sum_{i=1}^{n} x_i^2 &= \sum_{i=1}^{n} x_i y_i. \end{aligned}$$

Las ecuaciones anteriores se les denominan *ecuaciones normales de mínimos cuadrados* con solución

$$\hat{\beta}_0 = \overline{y} - \hat{\beta}_1 \overline{x}; \tag{4}$$

$$\hat{\beta}_1 = \frac{\sum_{i=1}^{n} x_i y_i - \frac{1}{n} \left(\sum_{i=1}^{n} y_i \right) \left(\sum_{i=1}^{n} x_i \right)}{\sum_{i=1}^{n} x_i^2 - \frac{1}{n} \left(\sum_{i=1}^{n} x_i \right)^2}, \tag{5}$$

entonces el modelo de regresión lineal simple ajustado es

$$\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x.$$

Se intrduce la siguiente notación

$$\begin{aligned} S_{xx} &= \sum_{i=1}^{n} (x_i - \overline{x})^2 = \sum_{i=1}^{n} x_i^2 - \frac{1}{n}\left(\sum_{i=1}^{n} x_i\right)^2, \\ S_{xy} &= \sum_{i=1}^{n} y_i (x_i - \overline{x}) = \sum_{i=1}^{n} x_i y_i - \frac{1}{n}\left(\sum_{i=1}^{n} x_i\right)\left(\sum_{i=1}^{n} y_i\right). \end{aligned}$$

Por lo tanto

$$\hat{\beta}_1 = \frac{S_{xy}}{S_{xx}}. \tag{6}$$

1.1 Propiedades de los Estimadores $\hat{\beta}_0$ y $\hat{\beta}_1$

Las propiedades estadísticas de los estimadores de mínimos cuadrados son útiles para evaluar la suficiencia del modelo. Dado que $\hat{\beta}_0$ y $\hat{\beta}_1$ son combinaciones lineales de las variables aleatorias y_i, también resultan ser variables aleatorias. A saber

$$\begin{aligned} E\left(\hat{\beta}_1\right) &= E\left(\frac{S_{xy}}{S_{xx}}\right) = \frac{1}{S_{xx}} E\left(\sum_{i=1}^{n} y_i (x_i - \overline{x})\right) = \frac{1}{S_{xx}} E\left(\sum_{i=1}^{n} (\beta_0 + \beta_1 x_i + \epsilon_i)(x_i - \overline{x})\right) \\ &= \frac{1}{S_{xx}}\left[\beta_0 E\left(\sum_{k=1}^{n} (x_k - \overline{x})\right) + E\left(\beta_1 \sum_{k=1}^{n} x_k (x_k - \overline{x})\right) + E\left(\sum_{k=1}^{n} \epsilon_k (x_k - \overline{x})\right)\right] \\ &= \frac{1}{S_{xx}} \beta_1 S_{xx} = \beta_1, \end{aligned}$$

por lo tanto

$$E\left(\hat{\beta}_1\right) = \beta_1. \tag{7}$$

Es decir, $\hat{\beta}_1$ es un estimador insesgado. Ahora calculemos la varianza:

$$\begin{aligned} V\left(\hat{\beta}_1\right) &= V\left(\frac{S_{xy}}{S_{xx}}\right) = \frac{1}{S_{xx}^2} V\left(\sum_{k=1}^{n} y_k (x_k - \overline{x})\right) \\ &= \frac{1}{S_{xx}^2} \sum_{k=1}^{n} V\left(y_k (x_k - \overline{x})\right) = \frac{1}{S_{xx}^2} \sum_{k=1}^{n} \sigma^2 (x_k - \overline{x})^2 \\ &= \frac{\sigma^2}{S_{xx}^2} \sum_{k=1}^{n} (x_k - \overline{x})^2 = \frac{\sigma^2}{S_{xx}}, \end{aligned}$$

por lo tanto

$$V\left(\hat{\beta}_1\right) = \frac{\sigma^2}{S_{xx}}. \tag{8}$$

Se tiene la siguiente proposición

Proposición 1.

$$E\left(\hat{\beta}_0\right) = \beta_0, \tag{9}$$

$$V\left(\hat{\beta}_0\right) = \sigma^2\left(\frac{1}{n}+\frac{\overline{x}^2}{S_{xx}}\right), \tag{10}$$

$$Cov\left(\hat{\beta}_0,\hat{\beta}_1\right) = -\frac{\sigma^2\overline{x}}{S_{xx}}. \tag{11}$$

Para estimar σ^2 es preciso definir la diferencia entre la observación y_k, y el valor predecido $\hat{y}_k$, es decir

$$e_k = y_k - \hat{y}_k \text{ (**residuo**)}.$$

La suma de los cuadrados de los errores de los residuos, *suma de cuadrados del error*,

$$SC_E = \sum_{k=1}^{n} e_k^2 = \sum_{k=1}^{n}\left(y_k-\hat{y}_k\right)^2,$$

sustituyendo $\hat{y}_k = \hat{\beta}_0 + \hat{\beta}_1 x_k$ se obtiene

$$\begin{aligned} SC_E &= \sum_{k=1}^{n} y_k^2 - n\overline{y}^2 - \hat{\beta}_1 S_{xy} = S_{yy} - \hat{\beta}_1 S_{xy}, \\ E\left(SC_E\right) &= \left(n-2\right)\sigma^2, \end{aligned}$$

por lo tanto

$$\hat{\sigma}^2 = \frac{SC_E}{n-2} = MC_E \tag{12}$$

que es un estimador insesgado de σ^2.

1.2 Prueba de Hipótesis en RLS

Para evaluar la suficiencia del modelo de regresión lineal simple, es necesario lleva a cabo una prueba de hipótesis respecto de los parámetros del modelo así como de la construcción de intervalos de confianza.

Para poder realizar la prueba de hipótesis sobre la pendiente y la ordenada al orígen de la recta de regresión es necesario hacer el supuesto de que el error ϵ_i se distribuye normalmente, es decir $\epsilon_i \sim N\left(0,\sigma^2\right)$. Suponga que se desea probar la hipótesis de que la pendiente es igual a una constante, $\beta_{0,1}$ las hipótesis Nula y Alternativa son:

H_0: $\beta_1 = \beta_{1,0}$,

H_1: $\beta_1 \neq \beta_{1,0}$.

donde dado que las $\epsilon_i \sim N\left(0,\sigma^2\right)$, se tiene que y_i son variables aleatorias normales $N\left(\beta_0+\beta_1 x_1,\sigma^2\right)$.

De las ecuaciones (4) se desprende que $\hat{\beta}_1$ es combinación lineal de variables aleatorias normales independientes, es decir, $\hat{\beta}_1 \sim N\left(\beta_1,\sigma^2/S_{xx}\right)$, recordar las ecuaciones (7) y (8). Entonces se tiene que el estadístico de prueba apropiado es

$$t_0 = \frac{\hat{\beta}_1 - \hat{\beta}_{1,0}}{\sqrt{MC_E/S_{xx}}}, \tag{13}$$

que se distribuye t con $n-2$ grados de libertad bajo $H_0 : \beta_1 = \beta_{1,0}$. Se rechaza H_0 si

$$t_0 > t_{\alpha/2,n-2}. \tag{14}$$

Para β_0 se puede proceder de manera análoga para

$H_0 : \beta_0 = \beta_{0,0}$,

$H_1 : \beta_0 \neq \beta_{0,0}$,

con $\hat{\beta}_0 \sim N\left(\beta_0, \sigma^2\left(\frac{1}{n} + \frac{\overline{x}^2}{S_{xx}}\right)\right)$, por lo tanto

$$t_0 = \frac{\hat{\beta}_0 - \beta_{0,0}}{MC_E\left[\frac{1}{n} + \frac{\overline{x}^2}{S_{xx}}\right]}, \tag{15}$$

con el que rechazamos la hipótesis nula si

$$t_0 > t_{\alpha/2,n-2}. \tag{16}$$

No rechazar $H_0 : \beta_1 = 0$ es equivalente a decir que no hay relación lineal entre x y y. Alternativamente, si $H_0 : \beta_1 = 0$ se rechaza, esto implica que x explica la variabilidad de y, es decir, podría significar que la línea recta esel modelo adecuado. El procedimiento de prueba para $H_0 : \beta_1 = 0$ puede realizarse de la siguiente manera:

$$\begin{aligned}
S_{yy} &= \sum_{k=1}^{n}(y_k - \overline{y})^2 = \sum_{k=1}^{n}(\hat{y}_k - \overline{y})^2 + \sum_{k=1}^{n}(y_k - \hat{y}_k)^2 \\
&= \sum_{k=1}^{n}(y_k - \hat{y}_k + \hat{y}_k - \overline{y})^2 = \sum_{k=1}^{n}[(\hat{y}_k - \overline{y}) + (y_k - \hat{y}_k)]^2 \\
&= \sum_{k=1}^{n}\left[(\hat{y}_k - \overline{y})^2 + 2(\hat{y}_k - \overline{y})(y_k - \hat{y}_k) + (y_k - \hat{y}_k)^2\right] \\
&= \sum_{k=1}^{n}(\hat{y}_k - \overline{y})^2 + 2\sum_{k=1}^{n}(\hat{y}_k - \overline{y})(y_k - \hat{y}_k) + \sum_{k=1}^{n}(y_k - \hat{y}_k)^2 \\
&\quad \sum_{k=1}^{n}(\hat{y}_k - \overline{y})(y_k - \hat{y}_k) = \sum_{k=1}^{n}\hat{y}_k(y_k - \hat{y}_k) - \sum_{k=1}^{n}\overline{y}(y_k - \hat{y}_k) \\
&= \sum_{k=1}^{n}\hat{y}_k(y_k - \hat{y}_k) - \overline{y}\sum_{k=1}^{n}(y_k - \hat{y}_k) \\
&= \sum_{k=1}^{n}\left(\hat{\beta}_0 + \hat{\beta}_1 x_k\right)\left(y_k - \hat{\beta}_0 - \hat{\beta}_1 x_k\right) - \overline{y}\sum_{k=1}^{n}\left(y_k - \hat{\beta}_0 - \hat{\beta}_1 x_k\right)
\end{aligned}$$

$$
\begin{aligned}
&= \sum_{k=1}^{n} \hat{\beta}_0 \left(y_k - \hat{\beta}_0 - \hat{\beta}_1 x_k\right) + \sum_{k=1}^{n} \hat{\beta}_1 x_k \left(y_k - \hat{\beta}_0 - \hat{\beta}_1 x_k\right) - \overline{y} \sum_{k=1}^{n} \left(y_k - \hat{\beta}_0 - \hat{\beta}_1 x_k\right) \\
&= \hat{\beta}_0 \sum_{k=1}^{n} \left(y_k - \hat{\beta}_0 - \hat{\beta}_1 x_k\right) + \hat{\beta}_1 \sum_{k=1}^{n} x_k \left(y_k - \hat{\beta}_0 - \hat{\beta}_1 x_k\right) - \overline{y} \sum_{k=1}^{n} \left(y_k - \hat{\beta}_0 - \hat{\beta}_1 x_k\right) \\
&= 0 + 0 + 0 = 0.
\end{aligned}
$$

Por lo tanto, efectivamente se tiene

$$S_{yy} = \sum_{k=1}^{n} (\hat{y}_k - \overline{y})^2 + \sum_{k=1}^{n} (y_k - \hat{y}_k)^2 , \tag{17}$$

entonces se tiene que

$$SC_E = \sum_{k=1}^{n} (\hat{y}_k - \overline{y})^2 \cdots \text{Suma de Cuadrados del Error.} \tag{18}$$

$$SC_R = \sum_{k=1}^{n} (y_k - \hat{y}_k)^2 \cdots \text{ Suma de Regresión de Cuadrados.} \tag{19}$$

Por lo tanto la ecuación (17) se puede reescribir como

$$S_{yy} = SC_R + SC_E, \tag{20}$$

recordemos que $SC_E = S_{yy} - \hat{\beta}_1 S_{xy}$, entonces

$$
\begin{aligned}
S_{yy} &= SC_R + \left(S_{yy} - \hat{\beta}_1 S_{xy}\right), \\
S_{xy} &= \frac{1}{\hat{\beta}_1} SC_R,
\end{aligned}
$$

luego, S_{xy} tiene $n-1$ grados de libertad y SC_R y SC_E tienen 1 y $n-2$ grados de libertad respectivamente.

Proposición 2.

$$E\left(SC_R\right) = \sigma^2 + \beta_1 S_{xx}, \tag{21}$$

además, SC_E y SC_R son independientes.

Recordemos que $\hat{\beta}_1 = \frac{S_{xy}}{S_{xx}}$. Para $H_0 : \beta_1 = 0$ verdadera,

$$F_0 = \frac{SC_R/1}{SC_E/(n-2)} = \frac{MC_R}{MC_E},$$

se distribuye $F_{1,n-2}$, y se rechazaría H_0 si $F_0 > F_{\alpha,1,n-2}$. El procedimiento de prueba de hipótesis puede presentarse como la tabla de análisis de varianza 1:

Table 1: Tabla de análisis de varianza

Fuente de variación	Suma de Cuadrados	Grados de Libertad	Media Cuadrática	F_0
Regresión	SC_R	1	MC_R	MC_R/MC_E
Error Residual	SC_E	$n-2$	MC_E	
Total	S_{yy}	$n-1$		

La prueba para la significación de la regresión puede desarrollarse basándose en la expresión (13), con $\hat{\beta}_{1,0} = 0$, es decir

$$t_0 = \frac{\hat{\beta}_1}{\sqrt{MC_E/S_{xx}}}. \tag{22}$$

Elevando al cuadrado ambos términos:

$$t_0^2 = \frac{\hat{\beta}_1^2 S_{xx}}{MC_E} = \frac{\hat{\beta}_1 S_{xy}}{MC_E} = \frac{MC_R}{MC_E}. \tag{23}$$

Observar que $t_0^2 = F_0$, por tanto la prueba que se utiliza para t_0 es la misma que para F_0.

1.3 Estimación de Intervalos en RLS

Además de la estimación puntual para los parámetros β_1 y β_0, es posible obtener estimaciones del intervalo de confianza de estos parámetros. El ancho de estos intervalos de confianza es una medida de la calidad total de la recta de regresión. Si los ϵ_k se distribuyen normal e independientemente, entonces

$$\frac{\left(\hat{\beta}_1 - \beta_1\right)}{\sqrt{\frac{MC_E}{S_{xx}}}} \quad y \quad \frac{\left(\hat{\beta}_0 - \beta_0\right)}{\sqrt{MC_E\left(\frac{1}{n} + \frac{\overline{x}^2}{S_{xx}}\right)}}$$

se distribuyen t con $n-2$ grados de libertad. Por tanto un intervalo de confianza de $100\,(1-\alpha)\,\%$ para β_1 está dado por

$$\hat{\beta}_1 - t_{\alpha/2,n-2}\sqrt{\frac{MC_E}{S_{xx}}} \leq \beta_1 \leq \hat{\beta}_1 + t_{\alpha/2,n-2}\sqrt{\frac{MC_E}{S_{xx}}}. \tag{24}$$

De igual manera, para β_0 un intervalo de confianza al $100\,(1-\alpha)\,\%$ es

$$\hat{\beta}_0 - t_{\alpha/2,n-2}\sqrt{MC_E\left(\frac{1}{n} + \frac{\overline{x}^2}{S_{xx}}\right)} \leq \beta_0 \leq \hat{\beta}_0 + t_{\alpha/2,n-2}\sqrt{MC_E\left(\frac{1}{n} + \frac{\overline{x}^2}{S_{xx}}\right)} \tag{25}$$

1.4 Predicción

Supongamos que se tiene un valor x_0 de interés, entonces la estimación puntual de este nuevo valor

$$\hat{y}_0 = \hat{\beta}_0 + \hat{\beta}_1 x_0. \tag{26}$$

Esta nueva observación es independiente de las utilizadas para obtener el modelo de regresión, por tanto, el intervalo en torno a la recta de regresión es inapropiado, puesto que se basa únicamente en los datos empleados para ajustar el modelo de regresión. El intervalo de confianza en torno a la recta de regresión se refiere a la

respuesta media verdadera $x = x_0$, no a observaciones futuras. Sea y_0 la observación futura en $x = x_0$, y sea $\hat{y}_0$ dada en la ecuación anterior, el estimador de y_0. Si se define la variable aleatoria

$$w = y_0 - \hat{y}_0,$$

esta se distribuye normalmente con media cero y varianza

$$V(w) = \sigma^2 \left[1 + \frac{1}{n} + \frac{(x - x_0)^2}{S_{xx}} \right],$$

dado que y_0 es independiente de $\hat{y}_0$, por lo tanto el intervalo de predicción al nivel α para futuras observaciones x_0 es

$$\hat{y}_0 - t_{\alpha/2,n-2} \sqrt{MC_E \left[1 + \frac{1}{n} + \frac{(x - x_0)^2}{S_{xx}} \right]} \leq y_0$$
$$\leq \hat{y}_0 + t_{\alpha/2,n-2} \sqrt{MC_E \left[1 + \frac{1}{n} + \frac{(x - x_0)^2}{S_{xx}} \right]}.$$

Es común encontrar que el modelo ajustado no satisface totalmente el modelo necesario para los datos, en este caso es preciso saber qué tan bueno es el modelo propuesto. Para esto se propone la siguiente prueba de hipótesis:

H_0 : El modelo propuesto se ajusta adecuademente a los datos.

H_1 : El modelo NO se ajusta a los datos.

La prueba implica dividir la suma de cuadrados del eror o del residuo en las siguientes dos componentes:

$$SC_E = SC_{EP} + SC_{FDA},$$

donde SC_{EP} es la suma de cuadrados atribuibles al error puro, y SC_{FDA} es la suma de cuadrados atribuible a la falta de ajuste del modelo. La cantidad

$$R^2 = \frac{SC_R}{S_{yy}} = 1 - \frac{SC_E}{S_{yy}}, \tag{27}$$

se denomina coeficiente de determinación y se utiliza para saber si el modelo de regresión es suficiente o no. Se puede demostrar que $0 \leq R^2 \leq 1$, una manera de interpretar este valor es que si $R^2 = k$, entonces el modelo de regresión explica el $k * 100\%$ de la variabilidad en los datos. En lo que corresponde a R^2

- No mide la magnitud de la pendiente de la recta de regresión.
- Un valor grande de R^2 no implica una pendiente empinada.
- No mide la suficiencia del modelo.
- Valores grandes de R^2 no implican necesariamente que el modelo de regresión proporcionará predicciones precisas para futuras observaciones.

2 Regresión múltiple

La regresión logística es una técnica de modelado estadístico utilizada para predecir la probabilidad de un evento binario (es decir, un evento que tiene dos posibles resultados) en función de una o más variables independientes. Un modelo de regresión logística describe cómo una variable dependiente binaria Y (que puede tomar los valores 0 o 1) está relacionada con una o más variables independientes $X_1, X_2, \ldots, X_n$. A diferencia de la regresión lineal, que predice un valor continuo, la regresión logística predice una probabilidad que puede ser interpretada como la probabilidad de que $Y = 1$ dado un conjunto de valores para $X_1, X_2, \ldots, X_n$, la regresión logística está diseñada para manejar situaciones donde la respuesta es categórica.

En su forma más común, la regresión logística binaria, el modelo predice la probabilidad de que un evento ocurra en función de una o más variables independientes. Este tipo de regresión toma la forma de un modelo no lineal, debido a la naturaleza discreta de la variable dependiente. La regresión lineal busca modelar la relación entre una variable dependiente continua Y y una o más variables independientes $X_1, X_2, \ldots, X_n$ mediante una ecuación de la forma:

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \ldots + \beta_n X_n + \epsilon, \tag{28}$$

donde:

a) Y es la variable dependiente.

b) β_0 es la intersección con el eje Y o término constante.

c) $\beta_1, \beta_2, \ldots, \beta_n$ son los coeficientes que representan la relación entre las variables independientes y la variable dependiente.

d) $X_1, X_2, \ldots, X_n$ son las variables independientes.

e) ϵ es el término de error, que representa la desviación de los datos observados de los valores predichos por el modelo.

El objetivo de la regresión lineal es encontrar los valores de los coeficientes $\beta_0, \beta_1, \ldots, \beta_n$ que minimicen la suma de los cuadrados de las diferencias entre los valores observados y los valores predichos. Este método se conoce como mínimos cuadrados ordinarios (OLS, por sus siglas en inglés). La función de costo a minimizar es:

$$J(\beta_0, \beta_1, \ldots, \beta_n) = \sum_{i=1}^{n} (y_i - \hat{y}_i)^2, \tag{29}$$

donde:

a) y_i es el valor observado de la variable dependiente para la i-ésima observación.

b) $\hat{y}_i$ es el valor predicho por el modelo para la i-ésima observación, dado por:

$$\hat{y}_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \ldots + \beta_n x_{in}. \tag{30}$$

Para encontrar los valores óptimos de los coeficientes, se toman las derivadas parciales de la función de costo con respecto a cada coeficiente y se igualan a cero:

$$\frac{\partial J}{\partial \beta_j} = 0 \quad \text{para } j = 0, 1, \ldots, n.$$

Resolviendo este sistema de ecuaciones, se obtienen los valores de los coeficientes que minimizan la función de costo.

La deducción de la fórmula de la regresión logística comienza con la necesidad de modelar la probabilidad de un evento binario. Queremos encontrar una función que relacione las variables independientes con la probabilidad de que la variable dependiente tome el valor 1. La probabilidad de que el evento ocurra, $P(Y = 1)$, se denota como p. La probabilidad de que el evento no ocurra, $P(Y = 0)$, es $1-p$. Los **odds** (chances/momios) de que ocurra el evento se definen como:

$$\text{odds} = \frac{p}{1-p}. \tag{31}$$

Los momios indican cuántas veces más probable es que ocurra el evento frente a que no ocurra. Para simplificar el modelado de los *momios*, se aplica el logaritmo natural, obteniendo la función **logit**:

$$\text{logit}(p) = \log\left(\frac{p}{1-p}\right). \tag{32}$$

La transformación logit es útil porque convierte el rango de la probabilidad (0, 1) al rango de números reales $(-\infty, \infty)$. La idea clave de la regresión logística es modelar la transformación logit de la probabilidad como una combinación lineal de las variables independientes:

$$\text{logit}(p) := \log\left(\frac{p}{1-p}\right) = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \ldots + \beta_n X_n. \tag{33}$$

Aquí, β_0 es el término constante y $\beta_1, \beta_2, \ldots, \beta_n$ son los coeficientes asociados con las variables independientes $X_1, X_2, \ldots, X_n$. Para expresar p en función de una combinación lineal de las variables independientes, invertimos la transformación logit. Partimos de la ecuación 33, aplicando la función exponencial en ambos lados:

$$\frac{p}{1-p} = e^{\beta_0+\beta_1 X_1+\beta_2 X_2+\ldots+\beta_n X_n}.$$

Despejando p:

$$p = \frac{e^{\beta_0+\beta_1 X_1+\beta_2 X_2+\ldots+\beta_n X_n}}{1+e^{\beta_0+\beta_1 X_1+\beta_2 X_2+\ldots+\beta_n X_n}}.$$

La expresión final que obtenemos es conocida como la **función logística**:

$$p = \frac{1}{1+e^{-(\beta_0+\beta_1 X_1+\beta_2 X_2+\ldots+\beta_n X_n)}}. \tag{34}$$

Esta función describe cómo las variables independientes se relacionan con la probabilidad de que el evento de interés ocurra. Los coeficientes $\beta_0, \beta_1, \ldots, \beta_n$ se estiman a partir de los datos utilizando el método de máxima verosimilitud.

3 Método de Máxima Verosimilitud

Para estimar los coeficientes $\beta_0, \beta_1, \ldots, \beta_n$ en la regresión logística, utilizamos el método de máxima verosimilitud. La idea es encontrar los valores de los coeficientes que maximicen la probabilidad de observar los datos dados. Esta probabilidad se expresa mediante la función de verosimilitud L. La función de verosimilitud $L(\beta_0, \beta_1, \ldots, \beta_n)$ para un conjunto de n observaciones se define como el producto de las probabilidades de las observaciones dadas las variables independientes:

$$L(\beta_0, \beta_1, \ldots, \beta_n) = \prod_{i=1}^{n} p_i^{y_i}(1-p_i)^{1-y_i} \tag{35}$$

donde:

- p_i es la probabilidad predicha de que $Y_i = 1$,
- y_i es el valor observado de la variable dependiente para la i-ésima observación.

Trabajar directamente con esta función de verosimilitud puede ser complicado debido al producto de muchas probabilidades, especialmente si n es grande. Para simplificar los cálculos, se utiliza el logaritmo de la función de verosimilitud, conocido como la función de *log-verosimilitud*. El uso del logaritmo simplifica significativamente la diferenciación y maximización de la función. La función de log-verosimilitud se define como:

$$\log L(\beta_0, \beta_1, \ldots, \beta_n) = \sum_{i=1}^{n} \left[y_i \log(p_i) + (1 - y_i) \log(1 - p_i) \right]. \tag{36}$$

Aquí, log representa el logaritmo natural. Esta transformación es válida porque el logaritmo es una función monótona creciente, lo que significa que maximizar la log-verosimilitud es equivalente a maximizar la verosimilitud original. En la regresión logística, la probabilidad p_i está dada por la función logística:

$$p_i = \frac{1}{1 + e^{-(\beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + \ldots + \beta_n X_{in})}}.$$

Sustituyendo esta expresión en la función de log-verosimilitud, obtenemos:

$$\begin{aligned} \log L(\beta_0, \beta_1, \ldots, \beta_n) \quad &= \textstyle\sum_{i=1}^{n} \left[y_i \log \left(\frac{1}{1+e^{-(\beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + \ldots + \beta_n X_{in})}} \right) + \right. \\ &\quad \left. (1 - y_i) \log \left(1 - \frac{1}{1+e^{-(\beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + \ldots + \beta_n X_{in})}} \right) \right]. \end{aligned}$$

Simplificando esta expresión, notamos que:

$$\log \left(\frac{1}{1 + e^{-z}} \right) = -\log(1 + e^{-z}),$$

y

$$\log \left(1 - \frac{1}{1 + e^{-z}} \right) = \log \left(\frac{e^{-z}}{1 + e^{-z}} \right) = -z - \log(1 + e^{-z}).$$

Aplicando estas identidades, la función de log-verosimilitud se convierte en:

$$\begin{aligned} \log L(\beta_0, \beta_1, \ldots, \beta_n) \quad &= \quad \sum_{i=1}^{n} \Big[y_i (-\log(1 + e^{-(\beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + \ldots + \beta_n X_{in})})) \\ &+ \quad (1 - y_i) \left(-(\beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + \ldots + \beta_n X_{in}) \right) \\ &- \quad (1 - y_i) \log(1 + e^{-(\beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + \ldots + \beta_n X_{in})}) \Big]. \end{aligned}$$

Simplificando aún más, obtenemos:

$$\begin{aligned} \log L(\beta_0, \beta_1, \ldots, \beta_n) \quad &= \quad \sum_{i=1}^{n} \left[y_i (\beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + \ldots + \beta_n X_{in}) \right. \\ &- \quad \left. \log(1 + e^{\beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + \ldots + \beta_n X_{in}}) \right]. \end{aligned}$$

Para simplificar aún más la notación, podemos utilizar notación matricial. Definimos la matriz $\mathbf{X}$ de tamaño $n \times (k + 1)$ y el vector de coeficientes $\boldsymbol{\beta}$ de tamaño $(k + 1) \times 1$ como sigue:

$$\mathbf{X} = \begin{bmatrix} 1 & x_{11} & x_{12} & \dots & x_{1k} \\ 1 & x_{21} & x_{22} & \dots & x_{2k} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 1 & x_{n1} & x_{n2} & \dots & x_{nk} \end{bmatrix}, \quad \boldsymbol{\beta} = \begin{bmatrix} \beta_0 \\ \beta_1 \\ \beta_2 \\ \vdots \\ \beta_k \end{bmatrix}. \tag{37}$$

Entonces, la expresión para la función de log-verosimilitud es:

$$\log L(\boldsymbol{\beta}) = \sum_{i=1}^{n} \left[y_i(\mathbf{X}_i\boldsymbol{\beta}) - \log(1 + e^{\mathbf{X}_i\boldsymbol{\beta}}) \right], \tag{38}$$

donde $\mathbf{X}_i$ es la i-ésima fila de la matriz $\mathbf{X}$. Esta notación matricial simplifica la implementación y la derivación de los estimadores de los coeficientes en la regresión logística. Utilizando métodos numéricos, como el algoritmo de *Newton-Raphson*, se pueden encontrar los coeficientes que maximizan la función de log-verosimilitud. Para maximizar la función de log-verosimilitud, derivamos esta función con respecto a cada uno de los coeficientes β_j y encontramos los puntos críticos. La derivada parcial de la función de log-verosimilitud con respecto a β_j es:

$$\frac{\partial \log L(\boldsymbol{\beta})}{\partial \beta_j} = \sum_{i=1}^{n} \left[y_i X_{ij} - \frac{X_{ij} e^{\mathbf{X}_i\boldsymbol{\beta}}}{1 + e^{\mathbf{X}_i\boldsymbol{\beta}}} \right]. \tag{39}$$

Simplificando, esta derivada se puede expresar como:

$$\frac{\partial \log L(\boldsymbol{\beta})}{\partial \beta_j} = \sum_{i=1}^{n} X_{ij}(y_i - p_i), \text{ donde } p_i = \frac{1}{1 + e^{-\mathbf{X}_i\boldsymbol{\beta}}}. \tag{40}$$

Para encontrar los coeficientes que maximizan la log-verosimilitud, resolvemos el sistema de ecuaciones

$$\frac{\partial \log L(\boldsymbol{\beta})}{\partial \beta_j} = 0 \text{ para todos los } j = 0, 1, \dots, k.$$

Este sistema de ecuaciones no tiene una solución analítica cerrada, por lo que se resuelve numéricamente utilizando métodos iterativos como el algoritmo de Newton-Raphson.

4 Método de Newton-Raphson

El método de Newton-Raphson es un algoritmo iterativo que se utiliza para encontrar las raíces de una función. En el contexto de la regresión logística, se utiliza para maximizar la función de log-verosimilitud encontrando los valores de los coeficientes $\beta_0, \beta_1, \dots, \beta_n$. Este método se basa en una aproximación de segundo orden de la función objetivo. Dado un valor inicial de los coeficientes $\boldsymbol{\beta}^{(0)}$, se actualiza iterativamente el valor de los coeficientes utilizando la fórmula:

$$\boldsymbol{\beta}^{(t+1)} = \boldsymbol{\beta}^{(t)} - \left[\mathbf{H}(\boldsymbol{\beta}^{(t)}) \right]^{-1} \nabla \log L(\boldsymbol{\beta}^{(t)}), \tag{41}$$

donde:

- $\boldsymbol{\beta}^{(t)}$ es el vector de coeficientes en la t-ésima iteración.
- $\nabla \log L(\boldsymbol{\beta}^{(t)})$ es el gradiente de la función de log-verosimilitud con respecto a los coeficientes $\boldsymbol{\beta}$:

$$\nabla \log L(\boldsymbol{\beta}) = \mathbf{X}^T(\mathbf{y} - \mathbf{p}), \tag{42}$$

  donde $\mathbf{y}$ es el vector de valores observados y $\mathbf{p}$ es el vector de probabilidades.

- $\mathbf{H}(\boldsymbol{\beta}^{(t)})$ es la matriz Hessiana (matriz de segundas derivadas) evaluada en $\boldsymbol{\beta}^{(t)}$:

$$\mathbf{H}(\boldsymbol{\beta}) = -\mathbf{X}^T\mathbf{W}\mathbf{X}, \tag{43}$$

  donde $\mathbf{W}$ es una matriz diagonal de pesos con elementos $w_i = p_i(1 - p_i)$.

Algoritmo 1. *El algoritmo Newton-Raphson para la regresión logística se puede resumir en los siguientes pasos:*

1. *Inicializar el vector de coeficientes $\boldsymbol{\beta}^{(0)}$ (por ejemplo, con ceros o valores pequeños aleatorios).*
2. *Calcular el gradiente $\nabla \log L(\boldsymbol{\beta}^{(t)})$ y la matriz Hessiana $\mathbf{H}(\boldsymbol{\beta}^{(t)})$ en la iteración t.*
3. *Actualizar los coeficientes utilizando la fórmula:*

$$\boldsymbol{\beta}^{(t+1)} = \boldsymbol{\beta}^{(t)} - \left[\mathbf{H}(\boldsymbol{\beta}^{(t)})\right]^{-1} \nabla \log L(\boldsymbol{\beta}^{(t)}) \tag{44}$$

4. *Repetir los pasos 2 y 3 hasta que la diferencia entre $\boldsymbol{\beta}^{(t+1)}$ y $\boldsymbol{\beta}^{(t)}$ sea menor que un umbral predefinido (criterio de convergencia).*

El método de Newton-Raphson permite encontrar los coeficientes que maximizan la función de log-verosimilitud de manera eficiente. En específico para un conjunto de n observaciones, la función de verosimilitud L se define como el producto de las probabilidades individuales de observar cada dato:

$$L(\beta_0, \beta_1, \ldots, \beta_n) = \prod_{i=1}^{n} p_i^{y_i}(1 - p_i)^{1-y_i},$$

donde y_i es el valor observado de la variable dependiente para la i-ésima observación y p_i es la probabilidad predicha de que $Y_i = 1$. Aquí, p_i es dado por la función logística:

$$p_i = \frac{1}{1 + e^{-(\beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + \ldots + \beta_n X_{in})}}.$$

Tomando el logaritmo:

$$\log L(\beta_0, \beta_1, \ldots, \beta_n) = \sum_{i=1}^{n} \left[y_i \log(p_i) + (1 - y_i) \log(1 - p_i) \right].$$

Sustituyendo p_i:

$$\log L(\beta_0, \beta_1, \ldots, \beta_n) = \sum_{i=1}^{n} \left[y_i(\beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + \ldots + \beta_n X_{in}) - \log(1 + e^{\beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + \ldots + \beta_n X_{in}}) \right].$$

Dado que el objetivo es encontrar los valores de $\beta_0, \beta_1, \ldots, \beta_n$ que maximicen la función de log-verosimilitud. Para β_j, la derivada parcial de la función de log-verosimilitud es:

$$\frac{\partial \log L}{\partial \beta_j} = \sum_{i=1}^{n} \left[y_i X_{ij} - \frac{X_{ij} e^{\beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + \ldots + \beta_n X_{in}}}{1 + e^{\beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + \ldots + \beta_n X_{in}}} \right].$$

Esto se simplifica a (comparar con la ecuación 39):

$$\frac{\partial \log L}{\partial \beta_j} = \sum_{i=1}^{n} X_{ij}(y_i - p_i). \tag{45}$$

Para maximizar la log-verosimilitud, resolvemos el sistema de ecuaciones $\frac{\partial \log L}{\partial \beta_j} = 0$ para todos los j de 0 a n, mismo que se resuelve numéricamente utilizando métodos el algoritmo de Newton-Raphson. El método de Newton-Raphson se basa en una aproximación de segundo orden de la función objetivo. Dado un valor inicial de los coeficientes $\beta^{(0)}$, se iterativamente actualiza el valor de los coeficientes utilizando la fórmula:

$$\beta^{(k+1)} = \beta^{(k)} - \left[\mathbf{H}(\beta^{(k)})\right]^{-1} \mathbf{g}(\beta^{(k)}), \tag{46}$$

donde:

- $\beta^{(k)}$ es el vector de coeficientes en la k-ésima iteración.
- $\mathbf{g}(\beta^{(k)})$ es el gradiente (vector de primeras derivadas) evaluado en $\beta^{(k)}$:

$$\mathbf{g}(\beta) = \frac{\partial \log L}{\partial \beta} = \sum_{i=1}^{n} \mathbf{X}_i(y_i - p_i), \tag{47}$$

  donde $\mathbf{X}_i$ es el vector de valores de las variables independientes para la i-ésima observación (comparar con ecuación 42).

- $\mathbf{H}(\beta^{(k)})$ es la matriz Hessiana (matriz de segundas derivadas) evaluada en $\beta^{(k)}$:

$$\mathbf{H}(\beta) = \frac{\partial^2 \log L}{\partial \beta \partial \beta^T} = -\sum_{i=1}^{n} p_i(1 - p_i)\mathbf{X}_i\mathbf{X}_i^T, \tag{48}$$

  comparar con ecuación 43.

Algoritmo 2. *Los pasos del algoritmo Newton-Raphson para la regresión logística son:*

1. *Inicializar el vector de coeficientes $\beta^{(0)}$ (por ejemplo, con ceros o valores pequeños aleatorios).*
2. *Calcular el gradiente $\mathbf{g}(\beta^{(k)})$ y la matriz Hessiana $\mathbf{H}(\beta^{(k)})$ en la iteración k.*
3. *Actualizar los coeficientes utilizando la fórmula:*

$$\beta^{(k+1)} = \beta^{(k)} - \left[\mathbf{H}(\beta^{(k)})\right]^{-1} \mathbf{g}(\beta^{(k)}). \tag{49}$$

4. *Repetir los pasos 2 y 3 hasta que la diferencia entre $\beta^{(k+1)}$ y $\beta^{(k)}$ sea menor que un umbral predefinido (criterio de convergencia).*

Como se puede observar, la diferencia entre el Algoritmo 1 y el Algoritmo 2 son mínimas.

5 Desarrollo

El análisis de regresión estima la variable dependiente y, dado el rango de valores de la variable x. El modelo de regresión se plantea:

$$\begin{aligned} y &= \beta_0 + \beta_1 x + \epsilon, \text{ caso univariado,} \\ y &= \sum_{j=0}^{q} \beta_j x_j; x_0 = 1, \text{ caso multivariado.} \end{aligned}$$

Para resolver el problema de regresión lineal univariado se requiere encontrar β_0 y β_1 minimizando la función de versosimilitud:

$$L(\beta_0^*, \beta_1) = \sum_{i=1}^{n} [y_i - (\beta_0^* + \beta_1(x_i - \bar{x}))]^2;$$

Que se obtiene resolviendo

$$\begin{aligned} \frac{\partial}{\partial \beta_0} \sum [y_i - (\beta_0 + \beta_1 x_i)]^2 = 0, \\ \frac{\partial}{\partial \beta_1} \sum [y_i - (\beta_0 + \beta_1 x_i)]^2 = 0. \end{aligned}$$

Si se define

$$y_i = \beta_0^* + \beta_1(x_i - \bar{x}) + \varepsilon_i, \text{ donde } \beta_0 = \beta_0^* + \beta_1 \bar{x}.$$

Entonces se tiene

$$\begin{aligned} \frac{\partial}{\partial \beta_0^*} \sum [y_i - (\beta_0^* + \beta_1 \bar{x})]^2 &= 0, \\ \frac{\partial}{\partial \beta_1^*} \sum [y_i - (\beta_0^* + \beta_1 \bar{x})]^2 &= 0. \end{aligned}$$

Sabemos que:

$$\beta_0 = \beta_0^* + \beta_1 \bar{x}, \text{ entonces} \quad \beta_0^* = \beta_0 - \beta_1 \bar{x}, \text{ por lo tanto}$$

$$\begin{aligned} \frac{\partial}{\partial \beta_0} \sum [y_i - (\beta_0 + \beta_1 \bar{x})]^2 &= \sum \frac{\partial}{\partial \beta_0} [y_i - (\beta_0 + \beta_1 \bar{x})]^2 = \sum \frac{\partial}{\partial \beta_0} [y_i - (\beta_0^* - \beta_1 \bar{x} + \beta_1 x_i)]^2 \\ &= \sum \frac{\partial}{\partial \beta_0} [y_i - (\beta_0^* - (x_i - \bar{x})\beta_1)]^2 = -\sum 2 [y_i - (\beta_0^* - (x_i - \bar{x})\beta_1)] (-1) \\ &= 0, \end{aligned}$$

entonces

$$\begin{aligned} 2 \sum [y_i - (\beta_0^* + (x_i - \bar{x})\beta_1)] (-1) &= -2 \sum [y_i - \beta_0^* - \beta_1 (x_i - \bar{x})] \\ &= \sum y_i - n\beta_0^* - \beta_1 \sum (x_i - \bar{x}) = \sum y_i - n\beta_0^* = 0, \end{aligned}$$

entonces, $\sum y_i = n\beta_0^*$, ya que $\sum(x_i - \bar{x}) = 0$, por lo tanto

$$\beta_0^* = \bar{y}. \tag{50}$$

Por otra parte, la derivada respecto a β_1:

$$\begin{aligned}\frac{\partial}{\partial\beta_1}\sum\left[y_i - (\beta_0^* + \beta_1 x_i)\right]^2 &= \frac{\partial}{\partial\beta_1}\sum\left[y_i - (\beta_0^* - \beta_1\bar{x} + \beta_1 x_i)\right]^2 = \frac{\partial}{\partial\beta_1}\sum\left[y_i - (\beta_0^* + \beta_1(x_i - \bar{x}))\right]^2 \\ &= 2\sum\left[y_i - (\beta_0^* + \beta_1(x_i - \bar{x}))\right](x_i - \bar{x})(-1) = 0,\end{aligned}$$

de aquí que

$$\sum\left[y_i - (\beta_0^* + \beta_1(x_i - \bar{x}))\right](x_i - \bar{x}) = \sum y_i(x_i - \bar{x}) - \beta_0^*\sum(x_i - \bar{x}) - \beta_1\sum(x_i - \bar{x})^2 = 0.$$

Recordar que $\beta_0^* = \bar{y}$, entonces:

$$\sum y_i(x_i - \bar{x}) - \bar{y}\sum(x_i - \bar{x}) - \beta_1\sum(x_i - \bar{x})^2 = \sum(y_i - \bar{y})(x_i - \bar{x}) - \beta_1\sum(x_i - \bar{x})^2 = 0$$

entonces

$$\beta_1 = \frac{\sum(y_i - \bar{y})(x_i - \bar{x})}{\sum(x_i - \bar{x})^2} = \frac{S_{xy}}{S_{xx}}, \tag{51}$$

por lo tanto

$$\boxed{\beta_1 = \frac{S_{xy}}{S_{xx}},} \quad \text{y} \quad \boxed{\beta_0 = \bar{y} - \beta_1\bar{x}.} \tag{52}$$

Ejemplo 1. ***Caso Bernoulli***

Sea $X \in \mathbb{R}^{n\times d}$, donde n es el número de instancias; d el número de características; y un vector binario de resultados. Para todo $x_i \in \mathbb{R}^d$, la salida es $y_i \in \{0, 1\}$. El objetivo es clasificar la instancia x_i como positiva o negativa. Una instancia se puede pensar como un intento Bernoulli con esperanza $\mathbb{E}[y_i|x_i]$ o probabilidad ρ_i. Se propone el modelo

$$y = X\beta + \varepsilon, \quad \textit{donde } \varepsilon \textit{es el vector error.}$$

$$y = \begin{pmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{pmatrix}, \quad X = \begin{bmatrix} 1 & x_{11} & x_{12} & \cdots & x_{1d} \\ 1 & x_{21} & x_{22} & \cdots & x_{2d} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 1 & x_{n1} & x_{n2} & \cdots & x_{nd} \end{bmatrix}, \quad \varepsilon = \begin{pmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \vdots \\ \varepsilon_n \end{pmatrix},$$

y

$$\beta = \begin{pmatrix} \beta_0 \\ \beta_1 \\ \vdots \\ \beta_d \end{pmatrix} \quad \textit{es el vector de parámetros.}$$

Supongamos que $X_i = [1, x_i^\top]$ y $\beta = [\beta_0, \boldsymbol{\beta}^\top]$. Como y es una variable aleatoria Bernoulli con probabilidad ρ_i, se tiene:

$$P(y_i) = \begin{cases} \rho_i, & si\ y_i = 1, \\ 1-\rho_i, & si\ y_i = 0. \end{cases}$$

Entonces

$$\begin{aligned} \mathbb{E}[y_i] &= 1 \cdot \rho_i + 0 \cdot (1-\rho_i) = \rho_i = X_i^\top \beta, \\ \mathbb{V}[y_i] &= \rho_i(1-\rho_i). \end{aligned}$$

Por lo tanto se tiene

$$y_i = X_i^\top \beta + \varepsilon_i, \quad donde \quad \varepsilon_i = \begin{cases} 1-\rho_i, & si\ y_i = 1, \\ \rho_i, & si\ y_i = 0. \end{cases}$$

donde $\varepsilon_i \sim$ Binomial, con esperanza:

$$\mathbb{E}[\varepsilon_i] = (1-\rho_i)(\rho_i + (1-\rho_i))(1-\rho_i) = 0;$$

y varianza:

$$\mathbb{V}[\varepsilon_i] = \mathbb{E}[\varepsilon_i^2] - (\mathbb{E}[\varepsilon_i])^2 = (1-\rho_i)^2\rho_i + (-\rho_i)^2(1-\rho_i) \neq 0 = \rho_i(1-\rho_i).$$

Ahora, se sabe que

$$\mathbb{E}[Y_i = 1 \mid x_i, \beta] = \rho_i = \frac{e^{x_i^\top \beta}}{1+e^{x_i^\top \beta}} = \frac{1}{1+e^{-x_i^\top \beta}};$$

si se define

$$\eta_i = g(\rho_i) \quad = \quad \ln\left(\frac{\rho_i}{1-\rho_i}\right) = x_i^\top \beta \text{ entonces } \eta = X.$$

Ahora definamos la **Función de verosimilitud:**

$$\mathcal{L}(\beta) = \prod_{i=1}^{n} \rho_i^{y_i}(1-\rho_i)^{1-y_i} = \prod_{i=1}^{n} \left(\frac{e^{x_i^\top \beta}}{1+e^{x_i^\top \beta}}\right)^{y_i} \left(\frac{1}{1+e^{x_i^\top \beta}}\right)^{1-y_i},$$

aplicando el logaritmo natural

$$\ln \mathcal{L}(\beta) = \sum_{i=1}^{n} \left[y_i \ln\left(\frac{e^{x_i^\top \beta}}{1+e^{x_i^\top \beta}}\right) + (1-y_i)\ln\left(\frac{1}{1+e^{x_i^\top \beta}}\right)\right].$$

Calculando el gradiente y la matriz Hessiana:

$$\frac{\partial}{\partial \beta_j} \ln \mathcal{L}(\beta) = \sum \left[y_i \left(\frac{x_{ij}}{1+e^{x_i^\top \beta}}\right) + (1-y_i)\left(\frac{-x_{ij}e^{x_i^\top \beta}}{1+e^{x_i^\top \beta}}\right)\right].$$

Recordemos que:

$$\frac{\partial}{\partial \beta_j} \ln\left(\frac{1}{1+e^{x_i^\top \beta}}\right) \quad = \quad -\frac{1}{1+e^{x_i^\top \beta}} \cdot e^{x_i^\top \beta} \cdot x_{ij} = -\frac{e^{x_i^\top \beta}}{1+e^{x_i^\top \beta}} x_{ij},$$

entonces

$$\frac{\partial}{\partial \beta_j} \ln \mathcal{L}(\beta) = y_i \cdot \frac{1}{1+e^{x_i^\top \beta}} \cdot x_i = y_i \cdot \frac{x_i}{1+e^{x_i^\top \beta}}.$$

Por lo tanto

$$\begin{aligned} \frac{\partial}{\partial \beta_j} \ln \mathcal{L}(\beta) &= \sum_i \frac{\partial}{\partial \beta_j} \left[y_i \ln \left(\frac{e^{x_i^\top \beta}}{1+e^{x_i^\top \beta}} \right) + (1-y_i) \ln \left(\frac{1}{1+e^{x_i^\top \beta}} \right) \right] \\ &= \sum_i \left[y_i \frac{\partial}{\partial \beta_j} \ln \left(\frac{e^{x_i^\top \beta}}{1+e^{x_i^\top \beta}} \right) + (1-y_i) \frac{\partial}{\partial \beta_j} \ln \left(\frac{1}{1+e^{x_i^\top \beta}} \right) \right]. \end{aligned}$$

Dado que

$$\begin{aligned} \frac{\partial}{\partial \beta_j} \ln \left(\frac{e^{x_i^\top \beta}}{1+e^{x_i^\top \beta}} \right) &= \frac{\partial}{\partial \beta_j} \left[x_{ij} - \ln \left(1+e^{x_i^\top \beta} \right) \right] = x_{ij} - \frac{1}{1+e^{-x_i^\top \beta}} \cdot x_{ij} \\ &= x_{ij} \left[1 - \frac{e^{x_i^\top \beta}}{1+e^{x_i^\top \beta}} \right] = x_{ij} \cdot \frac{1}{1+e^{x_i^\top \beta}}, \end{aligned}$$

por otra parte

$$\begin{aligned} \frac{\partial}{\partial \beta_j} \ln \left(\frac{1}{1+e^{x_i^\top \beta}} \right) &= \frac{\partial}{\partial \beta_j} \ln \left(1+e^{x_i^\top \beta} \right) = -\frac{1}{1+e^{x_i^\top \beta}} \cdot e^{x_i^\top \beta} \cdot x_{ij} \\ &= -x_{ij} \cdot \frac{e^{x_i^\top \beta}}{1+e^{x_i^\top \beta}}, \end{aligned}$$

por lo tanto

$$\begin{aligned} \frac{\partial}{\partial \beta_j} \ln \mathcal{L}(\beta) &= \sum y_i x_{ij} \cdot \frac{1}{1+e^{x_i^\top \beta}} - (1-y_i) x_{ij} \cdot \frac{e^{x_i^\top \beta}}{1+e^{x_i^\top \beta}} \\ &= \sum y_i x_{ij} (1-\rho_i) - (1-y_i) x_{ij} \rho_i = \sum x_{ij} (y_i - \rho_i) = 0. \end{aligned}$$

En forma matricial se puede reescribir como

$$g(\beta) = \nabla_\beta \ln \mathcal{L}(\beta) = X^\top (y - \rho) = 0.$$

Ahora calculemos la segunda derivada de β

$$\frac{\partial^2}{\partial \beta_j \partial \beta_k} \ln \mathcal{L}(\beta) = \frac{\partial^2}{\partial \beta_j \partial \beta_k} \sum \left(\frac{-x_{ij} x_{ik} e^{x_i^\top \beta}}{(1+e^{x_i^\top \beta})^2} \right) = \frac{\partial^2}{\partial \beta_j \partial \beta_k} \sum x_{ij} x_{ik} \, \rho_i \, (1-\rho_i),$$

entonces

$$\frac{\partial^2}{\partial \beta_j \partial \beta_k} \ln \mathcal{L}(\beta) = \frac{\partial}{\partial \beta_k} \sum (x_{ij})(y_i - \rho_i) = -\sum x_{ij} \frac{\partial}{\partial \beta_k} \rho_i,$$

donde

$$\frac{\partial}{\partial \beta_k}\rho_i = \rho_i(1-\rho_i)x_{ik}$$

por lo tanto

$$\frac{\partial^2}{\partial \beta_j \partial \beta_k}\ln \mathcal{L}(\beta) = -\sum x_{ij}\rho_i(1-\rho_i)x_{ik} = -\sum x_{ij}x_{ik}\rho_i(1-\rho_i).$$

Si $v_i := \rho_i(1-\rho_i)$ y $\mathbb{V} = \text{diag}(v_1, v_2, \ldots, v_n)$, entonces

$$\mathcal{H}(\beta) = \nabla^2_\beta \ln \mathcal{L}(\beta) = -X^\top \mathbb{V} X$$

que es negativa definida, es decir, es cóncava con un máximo global. La matriz de información LR está dada por:

$$\mathcal{I}(\beta) = -\mathbb{E}[\mathcal{H}(\beta)] = X^\top \mathbb{V} X,$$

con Varianza:

$$\mathbb{V}(\hat{\beta}) = \mathcal{I}^{-1}(\beta) = (X^\top \mathbb{V} X)^{-1}.$$

La *log-verosimilitud regularizada* se define por

$$\begin{aligned}\ln \mathcal{L}(\beta) &= \sum_i y_i \ln\left(\frac{e^{x_i^\top \beta}}{1+e^{x_i^\top \beta}}\right) + (1-y_i)\ln\left(\frac{1}{1+e^{x_i^\top \beta}}\right) - \frac{\lambda}{2}\|\beta\|^2 \\ &= \sum_i \ln\left(\frac{e^{y_i x_i^\top \beta}}{1+e^{x_i^\top \beta}}\right) - \frac{\lambda}{2}\|\beta\|^2,\end{aligned}$$

donde λ es el *parámetro de regularización.* Retomando,

$$\begin{aligned}\nabla_\beta \ln \mathcal{L}(\beta) &= X^\top(y-p) = 0, \\ \nabla^2_\beta \ln \mathcal{L}(\beta) &= -X^\top \mathbb{V} X - \Sigma^{-1}.\end{aligned}$$

Ejemplo 2. *Supongamos que se tienen $\mathcal{D} = \left\{u^{(1)}, u^{(2)}, \ldots, u^{(N)}\right\}$ observaciones, supongamos además que se tienen datos generados con distrubución $U \sim (U;\theta)$. Calculemos la función de verosimlitud.*

$$\mathcal{L}(\theta) = \prod_{i=1}^{N} p\left(u^{(i)};\theta\right)$$

donde

$$\theta_{ML} = \arg\max_\theta \mathcal{L}(\theta) = \arg\max_\theta \sum_{i=1}^{N} \log p\left(u^{(i)};\theta\right)$$

donde tanto $log(f(x))$ y $\arg\max_\theta$ son funciones monótonas crecientes. supongamos que se tiene un ruido gaussiano com media 0 y varianza σ^2, entonces

$$y^{(i)} = h_\theta\left(x^{(i)}\right) + \epsilon^{(i)} = \theta^\top \mathbf{X}^{(i)} + \epsilon^{(i)},$$

por lo tanto

$$y^{(i)} \sim N\left(\theta^\top \mathbf{X}^{(i)}, \sigma^2\right),$$

entonces

$$\begin{aligned} p\left(y|\mathbf{X}, \theta, \sigma^2\right) &= \prod_{i=1}^{N} p\left(y|\mathbf{x}^{(i)}, \theta, \sigma^2\right) = \prod_{i=1}^{N} \left(2\pi\sigma^2\right) e^{-\frac{1}{2\sigma^2}\left(y^{(i)} - \theta^\top \mathbf{x}^{(i)}\right)^2} \\ &= \left(2\pi\sigma^2\right)^{-\frac{N}{2}} e^{-\frac{1}{2\sigma^2}\sum_{i=1}^{2}\left(y^{(i)} - \theta^\top \mathbf{x}^{(i)}\right)^2} \\ &= \left(2\pi\sigma^2\right)^{-\frac{N}{2}} e^{-\frac{1}{2\sigma^2}(y - \mathbf{X}\theta)^\top (y - \mathbf{X}\theta)} \end{aligned}$$

entonces la verosimilitud es

$$p\left(y|\mathbf{X}, \theta, \sigma^2\right) = \left(2\pi\sigma^2\right)^{-\frac{N}{2}} e^{-\frac{1}{2\sigma^2}(y - \mathbf{X}\theta)^\top (y - \mathbf{X}\theta)}$$

y la log-verosimilitud es

$$\mathcal{L}\left(\theta, \sigma^2\right) = -\frac{N}{2}\log\left(2\pi\sigma^2\right)\left[-\frac{1}{2\sigma^2}\left(y - \mathbf{X}\theta\right)^\top \left(y - \mathbf{X}\theta\right)\right].$$

Maximizar la log-verosimilitud con respecto aθ es equivalente a maximizar $-\left(y - \mathbf{X}\theta\right)^\top \left(y - \mathbf{X}\theta\right)$ que a su vez es equivalente a minimizar $\left(y - \mathbf{X}\theta\right)^\top \left(y - \mathbf{X}\theta\right)$.

Ejemplo 3. *Se define la función sigmoide*

$$\sigma(u) = \frac{1}{1 + e^{-u}} \quad \Rightarrow \textit{clasificador de regresión logística}$$

donde la regla de decisión para y

$$y = \sigma(h_{\boldsymbol{\theta}}(x)) = \sigma(\boldsymbol{\theta}^\top x)$$

Matemáticamente, la probabilidad de que un ejemplo pertenezca a la clase 1 es:

$$\begin{aligned} p\left(y^{(i)} = 1 \mid x^{(i)}; \boldsymbol{\theta}\right) &= \sigma(\boldsymbol{\theta}^\top x^{(i)}) \\ p\left(y^{(i)} = 0 \mid x^{(i)}; \boldsymbol{\theta}\right) &= 1 - \sigma(\boldsymbol{\theta}^\top x^{(i)}) \end{aligned}$$

la probabilidad conjunta en función de $y^{(i)}$

$$p\left(y^{(i)} \mid x^{(i)}; \boldsymbol{\theta}\right) = \sigma(\boldsymbol{\theta}^\top x^{(i)})^{y^{(i)}} \cdot \left[1 - \sigma(\boldsymbol{\theta}^\top x^{(i)})\right]^{1 - y^{(i)}}$$

mientras que la probabilidad conjunta de todas las etiquetas

$$\prod_{i=1}^{N} \sigma(\boldsymbol{\theta}^\top x^{(i)})^{y^{(i)}} \left(1 - \sigma(\boldsymbol{\theta}^\top x^{(i)})\right)^{(1 - y^{(i)})}$$

La log-verosimilitud para regresión logística está dada por:

$$\ell(\boldsymbol{\theta}) = \sum_{i=1}^{N} y^{(i)} \log\left(\sigma(\boldsymbol{\theta}^\top x^{(i)})\right) + (1 - y^{(i)}) \log\left(1 - \sigma(\boldsymbol{\theta}^\top x^{(i)})\right)$$

Antes de calcular la derivada, recordemos:

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

con derivada

$$\frac{d}{dz}\sigma(z) = \frac{d}{dz}\left(1+e^{-z}\right)^{-1} = -(1+e^{-z})^{-2}\cdot(-e^{-z}) = \frac{e^{-z}}{(1+e^{-z})^2}$$

además:

$$\begin{aligned}
\sigma(z) &= \frac{1}{1+e^{-z}} \Rightarrow 1-\sigma(z) = \frac{e^{-z}}{1+e^{-z}} \\
&\Rightarrow \sigma(z)(1-\sigma(z)) = \frac{e^{-z}}{(1+e^{-z})^2} \\
&\therefore \frac{d}{dz}\sigma(z) = \sigma(z)(1-\sigma(z))
\end{aligned}$$

Derivando la log-verosimilitud respecto a theta$_j$ la función $\ell(\boldsymbol{\theta})$:

$$\begin{aligned}
\frac{\partial \ell(\boldsymbol{\theta})}{\partial \theta_j} &= \sum_{i=1}^{N}\left[y^{(i)}\log\sigma(\boldsymbol{\theta}^\top x^{(i)}) + (1-y^{(i)})\log(1-\sigma(\boldsymbol{\theta}^\top x^{(i)}))\right] \\
&= \sum_{i=1}^{N}\left[\frac{y^{(i)}}{\sigma(\boldsymbol{\theta}^\top x^{(i)})} - \frac{1-y^{(i)}}{1-\sigma(\boldsymbol{\theta}^\top x^{(i)})}\right]\cdot\frac{d}{d\theta_j}\sigma(\boldsymbol{\theta}^\top x^{(i)}) \\
&= \sum_{i=1}^{N}\left[\frac{y^{(i)}}{\sigma(\boldsymbol{\theta}^\top x^{(i)})} - \frac{1-y^{(i)}}{1-\sigma(\boldsymbol{\theta}^\top x^{(i)})}\right]\cdot\sigma(\boldsymbol{\theta}^\top x^{(i)})\sigma(1-\boldsymbol{\theta}^\top x^{(i)})x_j^{(i)} \\
&= \sum_{i=1}^{N}\left[\frac{y^{(i)}-\sigma(\boldsymbol{\theta}^\top x^{(i)})}{\sigma(\boldsymbol{\theta}^\top x^{(i)})(1-\sigma(\boldsymbol{\theta}^\top x^{(i)}))}\right]\cdot\sigma(\boldsymbol{\theta}^\top x^{(i)})\sigma(1-\boldsymbol{\theta}^\top x^{(i)})x_j^{(i)} \\
&= \sum_{i=1}^{N}\left[y^{(i)}-\sigma(\boldsymbol{\theta}^\top x^{(i)})\right]x_j^{(i)}
\end{aligned}$$

la cual es la función recursiva para calcular el gradiente.

Ejemplo 4. *El modelo lineal univariado se expresa como:*

$$h_\theta(x) = \theta_0 + \theta_1 x$$

donde la función de Costo (SSE - Error Cuadrático medio) está definida por:

$$J(\theta_0,\theta_1) = \frac{1}{2N}\sum_{i=1}^{N}\left(h_\theta(x^{(i)}) - y^{(i)}\right)^2$$

Nuestro objetivo es encontrar los valores de θ_0 y θ_1 que minimicen la función de costo

$$\min_{\theta_0,\theta_1} J(\theta_0,\theta_1)$$

El gradiente de la función de costo respecto a θ_0

$$\frac{\partial J(\theta_0,\theta_1)}{\partial \theta_0} = \frac{1}{N}\sum_{i=1}^{N}\left(h_\theta(x^{(i)}) - y^{(i)}\right)$$

y el gradiente de la función de costo respecto a θ_1

$$\frac{\partial J(\theta_0,\theta_1)}{\partial \theta_1} = \frac{1}{N}\sum_{i=1}^{N} x^{(i)}\left(h_\theta(x^{(i)}) - y^{(i)}\right).$$

Para minimizar $J(\theta_0,\theta_1)$*, se pueden utilizar métodos iterativos, como el de gradiente descentiente*

$$\theta_j := \theta_j - \alpha \cdot \frac{\partial J(\theta)}{\partial \theta_j}, \quad j = 0,1;$$

donde α *es la corrección de la dirección de descenso.*

$$\begin{aligned}
\theta_0 &= \frac{1}{N}\left\{\sum_{i=1}^{N} y^{(i)} - \theta_1 \sum_{i=1}^{N} x^{(i)}\right\}, \\
\theta_1 &= \frac{N\sum_{i=1}^{N} y^{(i)}x^{(i)} - \sum_{i=1}^{N} y^{(i)}\sum_{i=1}^{N} x^{(i)}}{N\sum_{i=1}^{N}(x^{(i)})^2 - (\sum_{i=1}^{N} x^{(i)})^2}.
\end{aligned}$$

Para el caso multivariado sería

$$h_\theta(x) = \sum_{i=1}^{d}\theta_i x_i + \theta_0 = \sum_{i=0}^{d}\theta_i x_i, \text{ con } x_0 = 1,$$

es decir, en forma matricial se puede ver como

$$h_\theta(x) = \theta^T X,\ X = \begin{pmatrix} x_0 \\ x_1 \\ \vdots \\ x_d \end{pmatrix},$$

con

$$J(\theta) = J(\theta_0,\theta_1,\ldots,\theta_d) = \tfrac{1}{2N}\sum_{i=1}^{N}\left(\theta^T x^{(i)} - y^{(i)}\right)^2.$$

y

$$h_\theta(\mathbf{X}) = \theta^T\mathbf{X} = \mathbf{X}^T\theta.$$

Por lo tanto

$$\hat{\boldsymbol{y}} = \mathbf{X}\theta \quad \Leftrightarrow \quad \begin{bmatrix} \hat{y}^{(1)} \\ \hat{y}^{(2)} \\ \vdots \\ \hat{y}^{(N)} \end{bmatrix} = \begin{bmatrix} h_\theta\mathbf{x}^{(1)} \\ h_\theta\mathbf{x}^{(2)} \\ \vdots \\ h_\theta\mathbf{x}^{(N)} \end{bmatrix} = \begin{bmatrix} x_0^{(1)} & x_1^{(1)} & \cdots & x_d^{(1)} \\ x_0^{(2)} & x_1^{(2)} & \cdots & x_d^{(2)} \\ \vdots & \vdots & \ddots & \vdots \\ x_0^{(N)} & x_1^{(N)} & \cdots & x_d^{(N)} \end{bmatrix} \begin{bmatrix} \theta_0 \\ \theta_1 \\ \vdots \\ \theta_d \end{bmatrix},$$

donde $\mathbf{X} \in \mathbb{R}^{N\times(d+1)}$, $\hat{\mathbf{y}} \in \mathbb{R}^{N\times 1}$ *y* $\theta \in \mathbb{R}^{(d+1)\times 1}$. *Entonces*

$$\begin{aligned} J(\theta) &= \frac{1}{2N}\sum_{i=1}^{N}\left(\theta^T\mathbf{x}^{(i)} - y^{(i)}\right)^2 = \frac{1}{2N}\sum_{i=1}^{N}\left(\hat{y}^{(i)} - y^{(i)}\right)^2 \\ &= \frac{1}{2N}|\hat{\mathbf{y}} - \mathbf{y}|_2^2 = \frac{1}{2N}(\hat{\mathbf{y}} - \mathbf{y})^T(\hat{\mathbf{y}} - \mathbf{y}) = \frac{1}{2N}\left(\mathbf{X}\theta - y\right)^\top\left(\mathbf{X}\theta - y\right) \\ &= \frac{1}{2N}\left\{\theta^\top\left(\mathbf{X}^\top\mathbf{X}\right)\theta - \theta^\top\mathbf{X}^\top y - y^\top\mathbf{X}\theta + y^\top y\right\} \\ &= \frac{1}{2N}\left\{\theta^\top\left(\mathbf{X}^\top\mathbf{X}\right)\theta - \left(\mathbf{X}^\top y\right)^\top\theta - \left(\mathbf{X}^\top y\right)^\top\theta + y^\top y\right\} \\ &= \frac{1}{2N}\left\{\theta^\top\left(\mathbf{X}^\top\mathbf{X}\right)\theta - 2\left(\mathbf{X}^\top y\right)^\top\theta + y^\top y\right\} \end{aligned}$$

por lo tanto

$$J(\theta) = \frac{1}{2N}(\mathbf{X}\theta - \mathbf{y})^\top(\mathbf{X}\theta - \mathbf{y}).$$

Recordemos que $\theta^\top\mathbf{X}^\top\mathbf{y} = (\mathbf{X}^\top\mathbf{y})^\top\theta$, $(\mathbf{X}^\top\mathbf{y})^\top = \mathbf{y}^\top\mathbf{X}$ *y* $(\mathbf{a}^\top\mathbf{b}) = (\mathbf{b}^\top\mathbf{a})$, *por lo tanto podemos reescribir:*

$$J(\theta) = \frac{1}{2N}\left(\theta^\top\left(\mathbf{X}^\top\mathbf{X}\right)\theta - 2\left(\mathbf{X}^\top y\right)^\top\theta + \mathbf{y}^\top\mathbf{y}\right).$$

Calculando el gradiente e igualando a cero:

$$\begin{aligned} \nabla_{\boldsymbol{\theta}} J(\boldsymbol{\theta}) &= -\frac{1}{2N}\left\{\boldsymbol{\theta}^\top(X^\top X)\boldsymbol{\theta} - 2(X^\top\mathbf{y})^\top\boldsymbol{\theta} + \mathbf{y}^\top\mathbf{y}\right\} \\ &= \frac{1}{2N}\left\{2X^\top X\boldsymbol{\theta} - 2X^\top\mathbf{y}\right\}\nabla_{\boldsymbol{\theta}} \\ J(\boldsymbol{\theta}) &= 0 \Leftrightarrow X^\top X\boldsymbol{\theta} = X^\top\mathbf{y} \Leftrightarrow \boldsymbol{\theta} = (X^\top X)^{-1}X^\top\mathbf{y} \end{aligned}$$

Alternativamente (gradiente descendente)

$$\frac{\partial J(\boldsymbol{\theta})}{\partial\theta_j} = \frac{1}{N}\sum_{i=1}^{N}\left(h_\theta(x^{(i)}) - y^{(i)}\right)x_j^{(i)}.$$

Nota 1. $(X^\top X)^\top = X^\top(X^\top)^\top = X^\top X$.

6 Notas finales

Nota 2. *En el contexto de la regresión logística, los vectores* $X_1, X_2, \ldots, X_n$ *representan las variables independientes. Cada* X_j *es un vector columna que contiene los valores de la variable independiente* j *para cada una de las* n *observaciones. Es decir,*

$$X_j = \begin{bmatrix} x_{1j} \\ x_{2j} \\ \vdots \\ x_{nj} \end{bmatrix}. \tag{53}$$

Para simplificar la notación y los cálculos, a menudo combinamos todos los vectores de variables independientes en una única matriz de diseño $\mathbf{X}$ *de tamaño* $n \times (k+1)$*, donde* n *es el número de observaciones y* $k+1$ *es el número de variables independientes más el término de intercepto. La primera columna de* $\mathbf{X}$ *corresponde a un vector de unos para el término de intercepto, y las demás columnas corresponden a los valores de las variables independientes:*

$$\mathbf{X} = \begin{bmatrix} 1 & x_{11} & x_{12} & \dots & x_{1k} \\ 1 & x_{21} & x_{22} & \dots & x_{2k} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 1 & x_{n1} & x_{n2} & \dots & x_{nk}, \end{bmatrix} \tag{54}$$

revisar la ecuación 37. De esta forma, el modelo logit puede ser escrito de manera compacta utilizando la notación matricial:

$$\mathit{logit}(p) = \log\left(\frac{p}{1-p}\right) = \mathbf{X}\boldsymbol{\beta}, \tag{55}$$

donde $\boldsymbol{\beta}$ *es el vector de coeficientes:*

$$\boldsymbol{\beta} = \begin{bmatrix} \beta_0 \\ \beta_1 \\ \beta_2 \\ \vdots \\ \beta_k \end{bmatrix}. \tag{56}$$

Así, la probabilidad p *se puede expresar como:*

$$p = \frac{1}{1 + e^{-\mathbf{X}\boldsymbol{\beta}}}. \tag{57}$$

Comparar la ecuación anterior con la ecuación 34. Esta notación matricial simplifica la implementación y la derivación de los estimadores de los coeficientes en la regresión logística. Para estimar los coeficientes $\boldsymbol{\beta}$ *en la regresión logística, se utiliza el método de máxima verosimilitud. La función de verosimilitud* $L(\boldsymbol{\beta})$ *se define como el producto de las probabilidades de las observaciones dadas las variables independientes, recordemos la ecuación 35:*

$$L(\boldsymbol{\beta}) = \prod_{i=1}^{n} p_i^{y_i}(1-p_i)^{1-y_i}, \tag{58}$$

donde y_i *es el valor observado de la variable dependiente para la* i*-ésima observación, y* p_i *es la probabilidad predicha de que* $Y_i = 1$*. La función de log-verosimilitud, que es más fácil de maximizar, se obtiene tomando el logaritmo natural de la función de verosimilitud (38):*

$$\log L(\boldsymbol{\beta}) = \sum_{i=1}^{n} \left[y_i \log(p_i) + (1-y_i)\log(1-p_i) \right]. \tag{59}$$

Sustituyendo $p_i = \frac{1}{1+e^{-\mathbf{X}_i\boldsymbol{\beta}}}$*, donde* $\mathbf{X}_i$ *es la* i*-ésima fila de la matriz de diseño* $\mathbf{X}$*, obtenemos:*

$$\log L(\boldsymbol{\beta}) = \sum_{i=1}^{n} \left[y_i(\mathbf{X}_i\boldsymbol{\beta}) - \log(1 + e^{\mathbf{X}_i\boldsymbol{\beta}}) \right]. \tag{60}$$

Para encontrar los valores de $\boldsymbol{\beta}$ *que maximizan la función de log-verosimilitud, se utiliza un algoritmo iterativo como el método de Newton-Raphson. Este método requiere calcular el gradiente y la matriz Hessiana de la función de log-verosimilitud. El gradiente de la función de log-verosimilitud con respecto a* $\boldsymbol{\beta}$ *es (42 y 47):*

$$\nabla \log L(\boldsymbol{\beta}) = \mathbf{X}^T(\mathbf{y} - \mathbf{p}), \tag{61}$$

donde $\mathbf{y}$ *es el vector de valores observados y* $\mathbf{p}$ *es el vector de probabilidades predichas. La matriz Hessiana de la función de log-verosimilitud es (43 y 48):*

$$\mathbf{H}(\boldsymbol{\beta}) = -\mathbf{X}^T\mathbf{W}\mathbf{X}, \tag{62}$$

donde $\mathbf{W}$ *es una matriz diagonal de pesos con elementos* $w_i = p_i(1 - p_i)$*. El método de Newton-Raphson actualiza los coeficientes* $\boldsymbol{\beta}$ *de la siguiente manera:*

$$\boldsymbol{\beta}^{(t+1)} = \boldsymbol{\beta}^{(t)} - [\mathbf{H}(\boldsymbol{\beta}^{(t)})]^{-1}\nabla \log L(\boldsymbol{\beta}^{(t)}). \tag{63}$$

Iterando este proceso hasta que la diferencia entre $\boldsymbol{\beta}^{(t+1)}$ *y* $\boldsymbol{\beta}^{(t)}$ *sea menor que un umbral predefinido (41, 44, 46 y 49), se obtienen los estimadores de máxima verosimilitud para los coeficientes de la regresión logística.*

Nota 3. ***Selección de Variables***

La selección de variables es el proceso de elegir las variables más relevantes para el modelo. Existen varias técnicas para la selección de variables:

Métodos de Filtrado: *Los métodos de filtrado seleccionan variables basadas en criterios estadísticos, como la correlación o la chi-cuadrado. Algunas técnicas comunes incluyen:*

- ***Análisis de Correlación****: Se seleccionan variables con alta correlación con la variable dependiente y baja correlación entre ellas.*
- ***Pruebas de Chi-cuadrado****: Se utilizan para variables categóricas para determinar la asociación entre la variable independiente y la variable dependiente.*

Métodos de Wrapper: *Los métodos de wrapper evalúan múltiples combinaciones de variables y seleccionan la combinación que optimiza el rendimiento del modelo. Ejemplos incluyen:*

- ***Selección hacia Adelante****: Comienza con un modelo vacío y agrega variables una por una, seleccionando la variable que mejora más el modelo en cada paso.*
- ***Selección hacia Atrás****: Comienza con todas las variables y elimina una por una, removiendo la variable que tiene el menor impacto en el modelo en cada paso.*
- ***Selección Paso a Paso****: Combina la selección hacia adelante y hacia atrás, agregando y eliminando variables según sea necesario.*

Métodos Basados en Modelos: *Los métodos basados en modelos utilizan técnicas de regularización como Lasso y Ridge para seleccionar variables. Estas técnicas añaden un término de penalización a la función de costo para evitar el sobreajuste.*

Regresión Lasso: *La regresión Lasso (Least Absolute Shrinkage and Selection Operator) añade una penalización* L_1 *a la función de costo:*

$$J(\beta) = \sum_{i=1}^{n}(y_i - \hat{y}_i)^2 + \lambda \sum_{j=1}^{p} |\beta_j|,$$

donde λ *es el parámetro de regularización que controla la cantidad de penalización.*

Regresión Ridge: *La regresión Ridge añade una penalización* L_2 *a la función de costo:*

$$J(\beta) = \sum_{i=1}^{n}(y_i - \hat{y}_i)^2 + \lambda \sum_{j=1}^{p} \beta_j^2,$$

donde λ *es el parámetro de regularización.*

Nota 4. *Evaluar la calidad y el rendimiento de un modelo de regresión logística es crucial para asegurar que las predicciones sean precisas y útiles. Este capítulo se centra en las técnicas y métricas utilizadas para evaluar modelos de clasificación binaria, así como en la validación cruzada, una técnica para evaluar la generalización del modelo. Las métricas de evaluación permiten cuantificar la precisión y el rendimiento de un modelo. Algunas de las métricas más comunes incluyen:*

Curva ROC y AUC: *La curva ROC (Receiver Operating Characteristic) es una representación gráfica de la sensibilidad (verdaderos positivos) frente a 1 - especificidad (falsos positivos). El área bajo la curva (AUC) mide la capacidad del modelo para distinguir entre las clases.*

$$\begin{aligned} \textit{Sensibilidad} &= \frac{TP}{TP+FN} \\ \textit{Especificidad} &= \frac{TN}{TN+FP} \end{aligned}$$

Matriz de Confusión: *La matriz de confusión es una tabla que muestra el rendimiento del modelo comparando las predicciones con los valores reales. Los términos incluyen:*

- ***Verdaderos Positivos (TP)****: Predicciones correctas de la clase positiva.*
- ***Falsos Positivos (FP)****: Predicciones incorrectas de la clase positiva.*
- ***Verdaderos Negativos (TN)****: Predicciones correctas de la clase negativa.*
- ***Falsos Negativos (FN)****: Predicciones incorrectas de la clase negativa.*

	Predicción Positiva	**Predicción Negativa**
Real Positiva	TP	FN
Real Negativa	FP	TN

Table 2: Matriz de Confusión

Precisión, Recall y F1-Score: *se define como*

$$\begin{aligned} \textit{Precisión} &= \frac{TP}{TP+FP} \\ \textit{Recall} &= \frac{TP}{TP+FN} \\ \textit{F1-Score} &= 2 \cdot \frac{\textit{Precisión} \cdot \textit{Recall}}{\textit{Precisión} + \textit{Recall}}. \end{aligned}$$

Log-Loss: *La pérdida logarítmica (Log-Loss) mide la precisión de las probabilidades predichas. La fórmula es:*

$$\textit{Log-Loss} = -\frac{1}{n}\sum_{i=1}^{n}\left[y_i \log(p_i) + (1-y_i)\log(1-p_i)\right],$$

donde y_i son los valores reales y p_i son las probabilidades predichas.

Validación Cruzada: *La validación cruzada es una técnica para evaluar la capacidad de generalización de un modelo. Existen varios tipos de validación cruzada:*

K-Fold Cross-Validation: *En K-Fold Cross-Validation, los datos se dividen en K subconjuntos. El modelo se entrena K veces, cada vez utilizando K-1 subconjuntos para el entrenamiento y el subconjunto restante para la validación.*

$$Error\ Medio = \frac{1}{K}\sum_{k=1}^{K} Error_k.$$

Leave-One-Out Cross-Validation (LOOCV): *En LOOCV, cada observación se usa una vez como conjunto de validación y las restantes como conjunto de entrenamiento. Este método es computacionalmente costoso pero útil para conjuntos de datos pequeños.*

Ajuste y Sobreajuste del Modelo: *El ajuste adecuado del modelo es crucial para evitar el sobreajuste (overfitting) y el subajuste (underfitting).*

Sobreajuste: *El sobreajuste ocurre cuando un modelo se ajusta demasiado bien a los datos de entrenamiento, capturando ruido y patrones irrelevantes. Los síntomas incluyen una alta precisión en el entrenamiento y baja precisión en la validación.*

Subajuste: *El subajuste ocurre cuando un modelo no captura los patrones subyacentes de los datos. Los síntomas incluyen baja precisión tanto en el entrenamiento como en la validación.*

Regularización: *La regularización es una técnica para prevenir el sobreajuste añadiendo un término de penalización a la función de costo. Las técnicas comunes incluyen:*

- ***Regresión Lasso (L1)***
- ***Regresión Ridge (L2)***

Nota 5. ***Interpretación de los Resultados:*** *Interpretar correctamente los resultados de un modelo de regresión logística es esencial para tomar decisiones informadas. Este capítulo se centra en la interpretación de los coeficientes del modelo, las odds ratios, los intervalos de confianza y la significancia estadística.*

Coeficientes de Regresión Logística: *Los coeficientes de regresión logística representan la relación entre las variables independientes y la variable dependiente en términos de log-odds. Cada coeficiente β_j en el modelo de regresión logística se interpreta como el cambio en el log-odds de la variable dependiente por unidad de cambio en la variable independiente X_j.*

$$\log\left(\frac{p}{1-p}\right) = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \ldots + \beta_n X_n$$

Signo de los Coeficientes:

- ***Coeficiente Positivo****: Un coeficiente positivo indica que un aumento en la variable independiente está asociado con un aumento en el log-odds de la variable dependiente.*
- ***Coeficiente Negativo****: Un coeficiente negativo indica que un aumento en la variable independiente está asociado con una disminución en el log-odds de la variable dependiente.*

Odds Ratios: *Las odds ratios proporcionan una interpretación más intuitiva de los coeficientes de regresión logística. La odds ratio para una variable independiente X_j se calcula como e^{β_j}. Recordemos que*

$$OR_j = e^{\beta_j}$$

Interpretación de las Odds Ratios:

- *$OR > 1$: Un OR mayor que 1 indica que un aumento en la variable independiente está asociado con un aumento en las odds de la variable dependiente.*
- *$OR < 1$: Un OR menor que 1 indica que un aumento en la variable independiente está asociado con una disminución en las odds de la variable dependiente.*

- *$OR = 1$: Un OR igual a 1 indica que la variable independiente no tiene efecto sobre las odds de la variable dependiente.*

Intervalos de Confianza: *Los intervalos de confianza proporcionan una medida de la incertidumbre asociada con los estimadores de los coeficientes. Un intervalo de confianza del 95% para un coeficiente β_j indica que, en el 95% de las muestras, el intervalo contendrá el valor verdadero de β_j.*

Para calcular un intervalo de confianza del 95% para un coeficiente β_j,

$$\beta_j \pm 1.96 \cdot SE(\beta_j),$$

donde $SE(\beta_j)$ es el error estándar de β_j.

Significancia Estadística: *La significancia estadística se utiliza para determinar si los coeficientes del modelo son significativamente diferentes de cero. Esto se evalúa mediante pruebas de hipótesis. Para cada coeficiente β_j, la hipótesis nula H_0 es que $\beta_j = 0$. La hipótesis alternativa H_a es que $\beta_j \neq 0$.*

P-valor: *El p-valor indica la probabilidad de obtener un coeficiente tan extremo como el observado, asumiendo que la hipótesis nula es verdadera. Un p-valor menor que el nivel de significancia α (típicamente 0.05) indica que podemos rechazar la hipótesis nula.*

Nota 6. *La **regresión** es un método valioso de investigación debido a su versátil aplicación en diferentes áreas. Por ejemplo, se puede utilizar para examinar asociaciones entre un resultado y varias variables independientes (también comúnmente conocidas como covariables, predictores o variables explicativas) o para determinar qué tan bien puede predecirse un resultado a partir de un conjunto de variables independientes. Adicionalmente, uno puede estar interesado en controlar el efecto de variables independientes específicas, particularmente aquellas que actúan como variables de confusión (es decir, cuya relación tanto con el resultado como con otra variable independiente oscurece la relación entre esa variable independiente y el resultado).*

En cuanto a las estrategias de modelado, existen tres tipos generales:

- *directa/estándar,*
- *secuencial/jerárquica y*
- *por pasos/estadística,*

cada una con distinto énfasis y propósito.

El ajuste general del modelo de regresión logística a los datos de muestra se evalúa utilizando varias medidas de bondad de ajuste, donde un mejor ajuste se caracteriza por una menor diferencia entre los valores observados y los valores predichos por el modelo. La regresión logística es ideal para predecir la probabilidad de ocurrencia de un evento binario (sí/no) y se basa en la transformación logística del odds ratio (razón de probabilidades). A diferencia de la regresión lineal, no requiere que las variables independientes sigan una distribución normal ni que la relación con la dependiente sea lineal. Los supuestos básicos que deben cumplirse para la regresión logística incluyen

- *independencia de errores,*
- *linealidad en el logit para variables continuas,*
- *ausencia de multicolinealidad y*
- *falta de valores atípicos fuertemente influyentes.*
- *existencia de un número adecuado de eventos por variable independiente para evitar un modelo sobreajustado, con un mínimo comúnmente recomendado de "reglas prácticas" que van de 10 a 20 eventos por covariable.*

Para asegurar que la regresión logística produzca un modelo preciso, se deben considerar factores críticos como la selección de variables independientes y la elección de la estrategia de construcción del modelo.

Criterio 1. ***Criterio de selección*** *Es muy importante seleccionar correctamente las variables independientes. Aunque la regresión logística es bastante flexible y permite distintos tipos de variables (continuas, ordinales y categóricas), alternativamente, uno podría optar por incluir todas las variables independientes relevantes independientemente de sus resultados univariados, ya que puede haber variables clínicamente importantes que merezcan inclusión a pesar de su desempeño estadístico; sin embargo, incluir demasiadas variables independientes en el modelo puede conducir a un modelo matemáticamente inestable, con menor capacidad de generalización más allá de la muestra actual del estudio [2, 3].*

Una parte clave del proceso de selección de variables es reconocer y considerar el papel de los posibles factores de confusión. Como se describió previamente, las variables de confusión son aquellas cuya relación tanto con el resultado como con otra variable independiente oculta la verdadera asociación entre esa variable independiente y el resultado. Independientemente del método para seleccionar las variables independientes, deben cumplirse ciertos supuestos básicos:

Supuesto 1. ***Independencia de los errores*** *Todos los resultados del grupo de muestra deben ser independientes entre sí; si los datos incluyen mediciones repetidas u otros resultados correlacionados, los errores también estarán correlacionados.*

Supuesto 2. ***Linealidad en el logit*** *para las variables continuas independientes, debe existir una relación lineal entre estas variables y sus respectivos resultados transformados en logit. Esto se puede realizar a través de la creación de un término de interacción entre cada variable continua independiente y su logaritmo natural. Si alguno de estos términos es estadísticamente significativo, se considera que el supuesto no se cumple.*

Supuesto 3. ***Ausencia de multicolinealidad****, o redundancia entre variables independientes, un modelo de regresión logística con variables independientes altamente correlacionadas usualmente genera errores estándar grandes para los coeficientes beta estimados. La solución común es eliminar una o más variables redundantes.*

Supuesto 4. ***Ausencia de valores atípicos altamente influyentes****, es decir, casos en los que el resultado predicho para un miembro de la muestra difiere considerablemente de su valor real, si hay demasiados valores atípicos, la precisión general del modelo puede verse comprometida. La detección de valores atípicos se realiza examinando los residuales (diferencia entre los valores predichos y los resultados reales) junto con estadísticas diagnósticas y gráficas; luego, se puede comparar el ajuste general del modelo y los coeficientes beta estimados con y sin los casos atípicos, dependiendo de la magnitud del cambio, uno podría conservar los valores atípicos cuyo efecto no sea alto o eliminar aquellos con una influencia particularmente fuerte sobre el modelo.*

Criterio 2. ***Número de variables a incluir*** *Como parte del proceso de selección de qué variables independientes incluir, también se debe decidir cuántas. El reto es seleccionar el menor número posible de variables independientes que expliquen mejor el resultado sin descuidar las limitaciones del tamaño de muestra. En términos generales, un modelo sobreajustado tiene coeficientes beta estimados para las variables independientes mucho mayores de lo que deberían ser, además de errores estándar más altos de lo esperado. Este tipo de situación genera inestabilidad en el modelo porque la regresión logística requiere más resultados que variables independientes para poder iterar soluciones diferentes en busca del mejor ajuste a través del método de máxima verosimilitud. Aunque no existe un estándar universalmente aceptado, hay algunas reglas generales derivadas en parte de estudios de simulación. Una de estas reglas sugiere que por cada variable independiente, debe haber al menos 10 resultados por cada categoría binaria, siendo el resultado menos frecuente el que determina el número máximo de variables independientes [9, 10]. Algunos estadísticos recomiendan una regla general aún más estricta de 20 resultados por variable independiente, dado que una relación más alta tiende a mejorar la validez del modelo[11].*

Nota 7. *Además de la cuidadosa selección de las variables independientes, se debe elegir el tipo adecuado de modelo de regresión logística para el estudio. De hecho, seleccionar una estrategia de construcción del modelo está estrechamente relacionado con la elección de variables independientes, por lo que estos dos componentes deben considerarse simultáneamente al planear un análisis de regresión logística.*

Existen tres enfoques generales para la construcción del modelo que se aplican a las técnicas de regresión en general, cada uno con un énfasis y propósito diferente:

a) ***Directo*** *(completo, estándar o simultáneo): Este enfoque es una especie de valor por defecto, ya que introduce todas las variables independientes en el modelo al mismo tiempo y no hace suposiciones sobre el orden o la importancia relativa de dichas variables. El enfoque directo es más adecuado si no existen hipótesis previas sobre cuáles variables tienen mayor relevancia que otras.*

b) ***Secuencial*** *(jerárquico): las variables se añaden secuencialmente para evaluar si mejoran el modelo de acuerdo a un orden predeterminado de prioridad. Aunque este enfoque es útil para clarificar patrones causales entre variables independientes y resultados, puede volverse complejo conforme aumentan los patrones causales, dificultando así la obtención de conclusiones definitivas sobre los datos en algunos casos.*

c) ***Paso a paso*** *(estadístico): En contraste con los dos métodos anteriores, la regresión paso a paso identifica variables independientes que deben mantenerse o eliminarse del modelo. Existen distintos tipos de técnicas paso a paso, incluyendo selección hacia adelante y eliminación hacia atrás con una contribución no significativa al resultado son eliminadas una por una hasta que sólo queden las variables estadísticamente significativas. Otra estrategia de construcción del modelo que es conceptualmente similar a la regresión por pasos se llama selección del mejor subconjunto', en la que se comparan modelos separados con diferentes números de variables independientes para determinar el mejor ajuste.*

Estas estrategias de construcción no son necesariamente intercambiables, ya que pueden producir diferentes medidas de ajuste del modelo y diferentes estimaciones puntuales para las variables independientes a partir de los mismos datos. Por lo tanto, identificar el modelo apropiado para los objetivos del estudio es extremadamente importante.

La regresión por pasos se basa en una selección automatizada de variables que tiende a aprovechar factores aleatorios en una muestra dada. Además, puede producir modelos que no parecen completamente razonables desde una perspectiva biológica, algunos argumentan que la regresión por pasos se reserva mejor para el tamizaje preliminar o únicamente para pruebas de hipótesis, como en casos de resultados novedosos y una comprensión limitada de las contribuciones de las variables independientes. Sin embargo, otros señalan que los métodos por pasos no son en sí el problema (y de hecho pueden ser bastante efectivos en ciertos contextos); en cambio, el verdadero problema es una interpretación descuidada de los resultados sin valorar completamente los pros y contras de este enfoque. Por tanto, si uno elige crear un modelo por pasos, es importante validar posteriormente los resultados antes de sacar conclusiones.

Al validar modelos de regresión logística, existen numerosos métodos entre los cuales elegir, cada uno más o menos apropiado según los parámetros del estudio como el tamaño de muestra. Para establecer la validez interna, los métodos comunes incluyen:

a) ***Método de retención, o división de la muestra en dos subgrupos*** *antes de la construcción del modelo, con el grupo de entrenamiento usado para crear el modelo de regresión logística y el grupo de prueba usado para validarlo; [12, 13]*

b) ***Validación cruzada k-fold o división de la muestra en*** *k* ***subgrupos de igual tamaño*** *para propósitos de entrenamiento y validación;[13]*

c) ***Validación cruzada uno fuera (leave-one-out)****, una variante del método k-fold donde el número de particiones es igual al número de sujetos en la muestra;[13] y*

d) ***Bootstrapping*** *es decir, obtener submuestras repetidas con reemplazo de toda la muestra [13, 14].*

Además de validar internamente el modelo, uno debería intentar validarlo externamente en un nuevo entorno de estudio como una prueba adicional de su viabilidad estadística y utilidad clínica [12, 15].

Una vez que se ha creado el modelo de regresión logística, se determina qué tan bien se ajusta a los datos de la muestra en su totalidad. Dos de los métodos más comunes para evaluar el ajuste del modelo son la prueba de chi-cuadrado de Pearson y la desviación residual. Ambas miden la diferencia entre los resultados observados

y los resultados predichos por el modelo, donde un mal ajuste del modelo se indica mediante valores de prueba elevados, lo que señala una diferencia mayor [3, 16, 17].

Otra medida comúnmente utilizada del ajuste del modelo es la prueba de bondad de ajuste de Hosmer-Lemeshow, que divide a los sujetos en grupos iguales (a menudo de 10) según su probabilidad estimada del resultado. El decil más bajo está compuesto por aquellos que tienen menor probabilidad de experimentar el resultado. Si el modelo tiene buen ajuste, los sujetos que experimentaron el resultado principal caerán en su mayoría en los deciles de mayor riesgo. Un modelo con mal ajuste resultará en sujetos distribuidos de manera más uniforme a lo largo de los deciles de riesgo para ambos resultados binarios [2, 3].

Las ventajas de las pruebas de Hosmer-Lemeshow incluyen su aplicación sencilla y facilidad de interpretación, las limitaciones incluyen la dependencia de las pruebas sobre cómo se definen los puntos de corte de los grupos y los algoritmos computacionales utilizados, así como una menor capacidad para identificar modelos con mal ajuste en ciertas circunstancias.

Otras alternativas menos comunes para evaluar el ajuste del modelo son descritas por Hosmer et al [16] y Kuss [17]. Otra opción para ampliar los resultados del ajuste del modelo y de las estadísticas diagnósticas, es evaluando la capacidad del modelo para discriminar entre grupos. Las formas comunes de hacer esto incluyen

a) Tablas de clasificación, donde la pertenencia a un grupo dentro de una categoría binaria del resultado se predice usando probabilidades estimadas y puntos de corte predefinidos, y

*b) Área bajo la curva característica operativa del receptor (**AUROC**), donde un valor de 0.5 significa que el modelo no es mejor que el azar para discriminar entre los sujetos que tienen el resultado y los que no, y un valor de 1.0 indica que el modelo discrimina perfectamente entre sujetos. El AUROC se usa a menudo cuando se desean considerar diferentes puntos de corte para la clasificación y así maximizar tanto la sensibilidad como la especificidad [18].*

Las variables independientes usualmente se presentan como razones de momios (ORs, por sus siglas en inglés), que revelan la fuerza de la contribución de la variable independiente al resultado y se definen como las probabilidades de que ocurra el resultado ($\hat{Y}$) frente a que no ocurra, $(1-\hat{Y})$, para cada variable independiente. La relación entre la razón de momios (OR) y el coeficiente beta estimado de la variable independiente se expresa como $OR = e^{\beta_i}$. Con base en esta fórmula, un cambio de una unidad en la variable independiente multiplica la probabilidad del resultado por la cantidad contenida en e^{β_i}.

*Para un modelo de regresión logística con solo una variable independiente, la OR se considera no ajustada porque no hay otras variables cuya influencia deba ser ajustada o restada. En contraste, si el modelo de regresión logística incluye múltiples variables independientes, las OR ahora son ajustadas porque representan la contribución única de la variable independiente después de ajustar (o restar) los efectos de las otras variables en el modelo, en conclusión las OR ajustadas suelen ser menores que sus contrapartes no ajustadas. Interpretar las OR también depende de si la variable independiente es continua o categórica. Para las variables continuas, primero se debe identificar una unidad de medida significativa que exprese mejor el grado de cambio en el resultado asociado con esa variable independiente. Finalmente, los intervalos de confianza (**IC**) al 95% se informan rutinariamente junto con las OR como una medida de precisión (es decir, si los hallazgos probablemente se mantendrán en la población no observada). Si el IC cruza 1.00, es posible que no haya una diferencia significativa en esa población.*

En problemas de clasificación con etiquetas binarias o etiquetas con una cantidad finita de opciones, la evaluación usualmente se realiza por medio de la matriz de confusión: el número de verdaderos/falsos positivos y negativos.

Resultado	*Positivo*	*Negativo*
Predecido Positivo	***TP***	***FP***
Predicido Negativo	***FN***	***TN***

Para problemas de regresión con etiquetas de valores continuos usualmente se calcula la raíz del error cuadrático medio

$$RMSE = \sqrt{\frac{1}{N}\sum_{i=1}^{N}(y_i - \hat{y})^2},$$
$$R^2 = 1 - \frac{\sum_{i=1}^{N}(y_i - \hat{y})^2}{\sum_{i=1}^{N}(y_i - \overline{y})^2}.$$

En cualquiera de los dos casos la evaluación final se lleva a cabo en el conjunto de prueba, el cuál es esencial dado que el último objetivo es obtener el predictor más general en los datos no utilizados para entrenar el algoritmo. Las siguientes métricas se utilizan para medir el rendimiento de un modelo en función de su capacidad para predecir correctamente las clases de un conjunto de datos.

***Recall** (Sensibilidad): Conocido como sensibilidad o tasa positiva real, mide la capacidad de un modelo para identificar correctamente todos los ejemplos positivos en un conjunto de datos. Se calcula como el número de verdaderos positivos dividido por la suma de verdaderos positivos y falsos negativos:*

$$Recall = \frac{Verdaderos\ Positivos}{Verdaderos\ Positivos + Falsos\ Negativos}. \tag{64}$$

Un recall alto significa que el modelo es bueno para detectar los casos positivos, minimizando los falsos negativos. Es importante en situaciones donde los falsos negativos son costosos o críticos.

***Precision** (Precisión): La precisión mide la capacidad de un modelo para predecir correctamente los casos positivos entre todas las predicciones positivas que realiza. Se calcula como el número de verdaderos positivos dividido por la suma de verdaderos positivos y falsos positivos:*

$$Precision = \frac{Verdaderos\ Positivos}{Verdaderos\ Positivos + Falsos\ Positivos}. \tag{65}$$

Una alta precisión significa que el modelo tiene una baja tasa de falsos positivos, es decir, que cuando predice una clase como positiva, es probable que sea correcta. La precisión es importante en situaciones en las que los falsos positivos son costosos o no deseados.

***Specificity** (Especificidad): La especificidad mide la capacidad de un modelo para predecir correctamente los casos negativos entre todas las predicciones negativas que realiza. También se conoce como tasa negativa real. Se calcula como el número de verdaderos negativos dividido por la suma de verdaderos negativos y falsos positivos:*

$$Specificity = \frac{Verdaderos\ Negativos}{Verdaderos\ Negativos + Falsos\ Positivos}. \tag{66}$$

Una alta especificidad indica que el modelo es bueno para identificar correctamente los casos negativos, minimizando los falsos positivos. Esto es importante en situaciones en las que los falsos positivos son costosos o problemáticos.

Estas métricas proporcionan una forma más completa de evaluar el rendimiento de un modelo de clasificación que simplemente mirar la precisión general.

*La **subestimación** ocurre cuando un predictor falla en encontrar patrones incluso en los datos de entrenamiento (cuando un modelo lineal simple se utiliza para explicar dependencia dependencias no lineales en los datos).*

*El **sobreajuste** ocurre cuando el desempeño de un predictor disminuye notablemente en los datos de prueba en comparación con los datos de prueba, debido al aprendizaje de demasiado detalle y ruido, en lugar de identificar patrones generales.*

Tanto el subajuste como el sobreajuste pueden ser debido a la insuficiente calidad de los datos: ruido excesivo, características faltantes o irrelevantes, sesgo en los datos, o datos dispersos. También pueden ocurrir como consecuencia de una pobre aplicación del algoritmo: excesiva o insuficiente flexibilidad en la selección de los parámetros, protocolo de entrenamiento inapropiado, o contaminación de los datos de entrenamiento con el conjunto de datos de prueba.

References

[1] Darlington RB. *Regression and Linear Models*. Columbus, OH: McGraw-Hill Publishing Company, 1990.

[2] Tabachnick BG, Fidell LS. *Using Multivariate Statistics*. 5th ed. Boston, MA: Pearson Education, Inc., 2007.

[3] Hosmer DW, Lemeshow SL. *Applied Logistic Regression*. 2nd ed. Hoboken, NJ: Wiley-Interscience, 2000.

[4] Campbell DT, Stanley JC. *Experimental and Quasi-experimental Designs for Research*. Boston, MA: Houghton Mifflin Co., 1963.

[5] Stokes ME, Davis CS, Koch GG. *Categorical Data Analysis Using the SAS System*. 2nd ed. Cary, NC: SAS Institute, Inc., 2000.

[6] Newgard CD, Hedges JR, Arthur M, Mullins RJ. Advanced statistics: the propensity score—a method for estimating treatment effect in observational research. *Acad Emerg Med*. 2004; **11**:953–961.

[7] Newgard CD, Haukoos JS. Advanced statistics: missing data in clinical research—part 2: multiple imputation. *Acad Emerg Med*. 2007; **14**:669–678.

[8] Allison PD. *Logistic Regression Using the SAS System: Theory and Application*. Cary, NC: SAS Institute, Inc., 1999.

[9] Peduzzi P, Concato J, Kemper E, Holford TR, Feinstein AR. A simulation study of the number of events per variable in logistic regression analysis. *J Clin Epidemiol*. 1996; **49**:1373–1379.

[10] Agresti A. *An Introduction to Categorical Data Analysis*. Hoboken, NJ: Wiley, 2007.

[11] Feinstein AR. *Multivariable Analysis: An Introduction*. New Haven, CT: Yale University Press, 1996.

[12] Altman DG, Royston P. What Do We Mean by Validating a Prognostic Model? *Stats Med*. 2000; **19**:453–473.

[13] Kohavi R. A study of cross-validation and bootstrap for accuracy estimation and model selection. In: *Proceedings of the 14th International Joint Conference on Artificial Intelligence (IJCAI)*. Montreal, Quebec, Canada, August 20–25, 1995. 1995:1137–1143.

[14] Efron B, Tibshirani R. *An Introduction to the Bootstrap*. New York: Chapman & Hall, 1993.

[15] Miller ME, Hiu SL, Tierney WM. Validation techniques for logistic regression models. *Stat Med*. 1991; **10**:1213–1226.

[16] Hosmer DW, Hosmer T, Le Cessie S, Lemeshow S. A comparison of goodness-of-fit tests for the logistic regression model. *Stat Med*. 1997; **16**:965–980.

[17] Kuss O. Global goodness-of-fit tests in logistic regression with sparse data. *Stat Med*. 2002; **21**:3789–3801.

[18] Zou KH, O'Malley AJ, Mauri L. Receiver-operating characteristic analysis for evaluating diagnostic tests and predictive models. *Circulation*. 2007; **115**:654–657.

[19] Mazurenko, S., Prokop, Z., and Damborsky, J. (2019). Machine learning in enzyme engineering. ACS Catalysis, 10(2), 1210-1223.

[20] Yang, Y.; Niroula, A.; Shen, B.; Vihinen, M. PON-Sol: Prediction of Effects of Amino Acid Substitutions on Protein Solubility. Bioinformatics 2016, 32, 2032!2034.

[21] Folkman, L.; Stantic, B.; Sattar, A.; Zhou, Y. EASE-MM: Sequence-Based Prediction of Mutation-Induced Stability Changes with Feature-Based Multiple Models. J. Mol. Biol. 2016, 428, 1394! 1405.

[22] Teng, S.; Srivastava, A. K.; Wang, L. Sequence Feature-Based Prediction of Protein Stability Changes upon Amino Acid Substitutions. BMC Genomics 2010, 11, S5.

[23] Huang, L.; Gromiha, M. M.; Ho, S. iPTREE-STAB: Interpretable Decision Tree Based Method for Predicting Protein Stability Changes Upon Mutations. Bioinformatics 2007, 23, 1292! 1293.

[24] Koskinen, P.; Toronen, P.; Nokso-Koivisto, J.; Holm, L. PANNZER: High-Throughput Functional Annotation of Uncharac- terized Proteins in an Error-Prone Environment. Bioinformatics 2015, 31, 1544!1552.

[25] De Ferrari, L.; Mitchell, J. B. From Sequence to Enzyme Mechanism Using Multi-Label Machine Learning. BMC Bioinf. 2014, 15, 150.

[26] Falda, M.; Toppo, S.; Pescarolo, A.; Lavezzo, E.; Di Camillo, B.; Facchinetti, A.; Cilia, E.; Velasco, R.; Fontana, P. Argot2: A Large Scale Function Prediction Tool Relying on Semantic Similarity of Weighted Gene Ontology Terms. BMC Bioinf. 2012, 13, S14.

[27] Cozzetto, D.; Buchan, D. W.; Bryson, K.; Jones, D. T. Protein Function Prediction by Massive Integration of Evolutionary Analyses and Multiple Data Sources. BMC Bioinf. 2013, 14, S1.

[28] Kulski, J. Next Generation Sequencing: Advances, Applications and Challenges; InTechOpen: London, 2016.

[29] Straiton, J.; Free, T.; Sawyer, A.; Martin, J. From Sanger Sequencing to Genome Databases and Beyond. BioTechniques 2019, 66, 60-63.

[30] Ardui, S.; Ameur, A.; Vermeesch, J. R.; Hestand, M. S. Single Molecule Real-Time (SMRT) Sequencing Comes of Age: Applications and Utilities for Medical Diagnostics. Nucleic Acids Res. 2018, 46, 2159-2168.

[31] Kono, N., and Arakawa, K. (2019). Nanopore sequencing: Review of potential applications in functional genomics. Development, growth and differentiation, 61(5), 316-326.

[32] Bunzel, H. A., Garrabou, X., Pott, M., and Hilvert, D. (2018). Speeding up enzyme discovery and engineering with ultrahigh-throughput methods. Current opinion in structural biology, 48, 149-156.

[33] Wrenbeck, E. E., Faber, M. S., and Whitehead, T. A. (2017). Deep sequencing methods for protein engineering and design. Current opinion in structural biology, 45, 36-44.

[34] Fowler, D. M., and Fields, S. (2014). Deep mutational scanning: a new style of protein science. Nature methods, 11(8), 801-807.

[35] Gupta, K., and Varadarajan, R. (2018). Insights into protein structure, stability and function from saturation mutagenesis. Current opinion in structural biology, 50, 117-125.

[36] UniProt Consortium. UniProt: A Worldwide Hub of Protein Knowledge. Nucleic Acids Res. 2018, 47, D506-D515.

[37] Evans, R.; Jumper, J.; Kirkpatrick, J.; Sifre, L.; Green, T.; Qin, C.; Zidek, A.; Nelson, A.; Bridgland, A.; Penedones, H.; Petersen, S.; Simonyan, K.; Jones, D. T.; Silver, D.; Kavukcuoglu, K.; Hassabis, D.; Senior, A. W. De Novo Structure Prediction with Deeplearning Based Scoring. In Thirteenth Critical Assessment of Techniques for Protein Structure Prediction Abstracts; 2018; pp 11-12.

[38] Kinch, L. N.; Shi, S.; Cheng, H.; Cong, Q.; Pei, J.; Mariani, V.; Schwede, T.; Grishin, N. V. CASP9 Target Classification. Proteins: Struct., Funct., Genet. 2011, 79, 21-36.

[39] Shehu, A.; Barbará, D.; Molloy, K. A Survey of ComputationalMethods for Protein Function Prediction. In Big Data Analytics in Genomics; Wong, K. C., Ed.; Springer: Cham, 2016; pp 225-298.

[40] Zhang, C.; Freddolino, P. L.; Zhang, Y. COFACTOR: Improved Protein Function Prediction by Combining Structure, Sequence and Protein-Protein Interaction Information. Nucleic Acids Res. 2017, 45, W291-W299.

[41] Kumar, N.; Skolnick, J. EFICAz2. 5: Application of a High-Precision Enzyme Function Predictor to 396 Proteomes. Bioinformatics 2012, 28, 2687-2688.

[42] Li, Y.; Wang, S.; Umarov, R.; Xie, B.; Fan, M.; Li, L.; Gao, X. DEEPre: Sequence-Based Enzyme EC Number Prediction by Deep Learning. Bioinformatics 2018, 34, 760-769.

[43] Yang, M.; Fehl, C.; Lees, K. V.; Lim, E. K.; Offen, W. A.; Davies, G. J.; Bowles, D. J.; Davidson, M. G.; Roberts, S. J.; Davis, B. G. Functional and Informatics Analysis Enables Glycosyltransferase Activity Prediction. Nat. Chem. Biol. 2018, 14, 1109-1117.

[44] Niwa, T.; Ying, B. W.; Saito, K.; Jin, W.; Takada, S.; Ueda, T.; Taguchi, H. Bimodal Protein Solubility Distribution Revealed by an Aggregation Analysis of the Entire Ensemble of Escherichia Coli Proteins. Proc. Natl. Acad. Sci. U. S. A. 2009, 106, 4201-4206.

[45] Klesmith, J. R.; Bacik, J. P.; Wrenbeck, E. E.; Michalczyk, R.; Whitehead, T. A. Trade-Offs Between Enzyme Fitness and Solubility Illuminated by Deep Mutational Scanning. Proc. Natl. Acad. Sci. U. S. A. 2017, 114, 2265-2270.

[46] Ruiz-Blanco, Y. B.; Paz, W.; Green, J.; Marrero-Ponce, Y. ProtDCal: A Program to Compute General-Purpose-Numerical Descriptors for Sequences and 3D-Structures of Proteins. BMC Bioinf. 2015, 16, 162.

[47] Han, X.; Wang, X.; Zhou, K. Develop Machine Learning-Based Regression Predictive Models for Engineering Protein Solubility. Bioinformatics 2019, 35, 4640-4646.

[48] Musil, M.; Konegger, H.; Hon, J.; Bednar, D.; Damborsky, J. Computational Design of Stable and Soluble Biocatalysts. ACS Catal. 2019, 9, 1033-1054.

[49] Li, G.; Dong, Y.; Reetz, M. T. Can Machine Learning Revolutionize Directed Evolution of Selective Enzymes? Adv. Synth. Catal. 2019, 361, 2377-2386.

[50] Wu, Z.; Kan, S. B. J.; Lewis, R. D.; Wittmann, B. J.; Arnold, F. H. Machine Learning-Assisted Directed Protein Evolution with Combinatorial Libraries. Proc. Natl. Acad. Sci. U. S. A. 2019, 116, 8852-8858.

[51] Wolpert, D. H.; Macready, W. G. No Free Lunch Theorems for Optimization. IEEE Trans. Evol. Comput. 1997, 1, 67-82.

[52] Wolpert, D. H. The Lack of a Priori Distinctions between Learning Algorithms. Neural Comput. 1996, 8, 1341-1390.

[53] Walsh, I.; Pollastri, G.; Tosatto, S. C. Correct Machine Learning on Protein Sequences: A Peer-Reviewing Perspective. Briefings Bioinf. 2016, 17, 831-840.

[54] Rao, R.; Bhattacharya, N.; Thomas, N.; Duan, Y.; Chen, X.; Canny, J.; Abbeel, P.; Song, Y. S. Evaluating Protein Transfer Learning with TAPE. arXiv preprint arXiv:1906.08230, 2019.

[55] Romero, P. A.; Krause, A.; Arnold, F. H. Navigating the Protein Fitness Landscape with Gaussian Processes. Proc. Natl. Acad. Sci. U. S. A. 2013, 110, E193-E201

[56] Eraslan, G.; Avsec, Z; Gagneur, J.; Theis, F. J. Deep Learning: New Computational Modelling Techniques for Genomics. Nat. Rev. Genet. 2019, 20, 389-403.

[57] Repecka, D.; Jauniskis, V.; Karpus, L.; Rembeza, E.; Zrimec, J.; Poviloniene, S.; Rokaitis, I.; Laurynenas, A.; Abuajwa, W.; Savolainen, O.; Meskys, R.; Engqvist, M. K. M.; Zelezniak, A. Expanding Functional Protein Sequence Space Using Generative Adversarial Networks. bioRxiv 2019, DOI: 10.1101/789719.

[58] Riesselman, A. J.; Ingraham, J. B.; Marks, D. S. Deep Generative Models of Genetic Variation Capture the Effects of Mutations. Nat. Methods 2018, 15, 816-822.

[59] Thornton, C.; Hutter, F.; Hoos, H. H.; Leyton-Brown, K. Auto- WEKA: Combined Selection and Hyperparameter Optimization of Classiffication Algorithms. In Proceedings of the 19th ACM SIGKDD international conference on Knowledge discovery and data mining; 2013; pp 847-855.

[60] Polikar, R. Ensemble based systems in decision making. IEEE Circuits and systems magazine 2006, 6, 21-45.

[61] Gammerman, A.; Vovk, V. Hedging Predictions in Machine Learning. Comput. J. 2007, 50, 151-163.

[62] Samek, W.; Wiegand, T.; Müller, K. Explainable Artificial Intelligence: Understanding, Visualizing and Interpreting Deep Learning Models. ITU Journal: ICT Discoveries 2017, 39-48.

[63] Shrikumar, A.; Greenside, P.; Kundaje, A. Learning Important Features through Propagating Activation differences. In Proceedings of the 34th International Conference on Machine Learning; 2017; Vol. 70, pp 3145-3153.

[64] Simonyan, K.; Vedaldi, A.; Zisserman, A. Deep Inside Convolutional Networks: Visualising Image Classiffication Models and Saliency Maps. arXiv preprint arXiv:1312.6034 2013.

[65] Brookes, D. H.; Park, H.; Listgarten, J. Conditioning by Adaptive Sampling for Robust Design. In Proceedings of the 36th International Conference on Machine Learning; 2019; Vol. 97, pp 773-782.

[66] Ribeiro, M. T.; Singh, S.; Guestrin, C. "Why Should I Trust You?" Explaining the Predictions of Any Classiffier. In Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining; 2016; pp 1135-1144.

[67] Szegedy, C.; Zaremba, W.; Sutskever, I.; Bruna, J.; Erhan, D.; Goodfellow, I.; Fergus, R. Intriguing Properties of Neural Networks. arXiv preprint arXiv:1312.6199 201

[68] Yu, M. K.; Ma, J.; Fisher, J.; Kreisberg, J. F.; Raphael, B. J.; Ideker, T. Visible Machine Learning for Biomedicine. Cell 2018, 173, 1562-1565.

[69] Copley, S. D. Shining a light on enzyme promiscuity. Curr. Opin. Struct. Biol. 47, 167–175 (2017).

[70] Nobeli, I., Favia, A. D. and Thornton, J. M. Protein promiscuity and its implications for biotechnology. Nat. Biotechnol. 27, 157–167 (2009)

[71] Adrio, J. L. and Demain, A. L. Microbial enzymes: tools for biotechnological processes. Biomolecules 4, 117–139 (2014).

[72] Wang, S. et al. Engineering a synthetic pathway for gentisate in pseudomonas chlororaphis p3. Front. Bioeng. Biotechnol. 8, 1588 (2021).

[73] Wu, M.-C., Law, B., Wilkinson, B. and micklefied, J. Bioengineering natural product biosynthetic pathways for therapeutic applications. Curr. Opin. Biotechnol. 23, 931–940 (2012)

[74] Rembeza, E., Boverio, A., Fraaije, M. W. and Engqvist, M. K. Discovery of two novel oxidases using a high-throughput activity screen. ChemBioChem 23, e202100510 (2022).

[75] Longwell, C. K., Labanieh, L. and Cochran, J. R. High-throughput screening technologies for enzyme engineering. Curr. Opin. Biotechnol. 48, 196–202 (2017).

[76] Black, G. W. et al. A high-throughput screening method for determining the substrate scope of nitrilases. Chem. Commun. 51, 2660–2662 (2015).

[77] Pertusi, D. A. et al. Predicting novel substrates for enzymes with minimal experimental effort with active learning. Metab. Eng. 44,171-181 (2017).

[78] Mou, Z. et al. Machine learning-based prediction of enzyme substrate scope: Application to bacterial nitrilases. Proteins Struct. Funct. Bioinf. 89, 336-347 (2021).

[79] Yang, M. et al. Functional and informatics analysis enables glycosyltransferase activity prediction. Nat. Chem. Biol. 14, 1109–1117 (2018).

[80] Rottig, M., Rausch, C. and Kohlbacher, O. Combining structure and sequence information allows automated prediction of substrate specificities within enzyme families. PLoS Comput. Biol. 6, e1000636 (2010).

[81] Chevrette, M. G., Aicheler, F., Kohlbacher, O., Currie, C. R. and Medema, M. H. Sandpuma: ensemble predictions of nonribosomal peptide chemistry reveal biosynthetic diversity across actinobacteria. Bioinformatics 33, 3202-3210 (2017).

[82] Goldman, S., Das, R., Yang, K. K. and Coley, C. W. Machine learning modeling of family wide enzyme-substrate specificity screens. PLoS Comput. Biol. 18, e1009853 (2022).

[83] Visani, G. M., Hughes, M. C. and Hassoun, S. Enzyme promiscuity prediction using hierarchy-informed multi-label classiffication Bioinformatics 37, 2017-2024 (2021).

[84] Ryu, J. Y., Kim, H. U. and Lee, S. Y. Deep learning enables high-quality and high-throughput prediction of enzyme commission numbers. PNAS 116, 13996-14001 (2019).

[85] Li, Y. et al. DEEPre: sequence-based enzyme EC number prediction by deep learning. Bioinformatics 34, 760-769 (2017).

[86] Sanderson, T., Bileschi, M. L., Belanger, D. and Colwell, L. J. Proteinfer, deep neural networks for protein functional inference. eLife 12, e80942 (2023).

[87] Bileschi, M. L. et al. Using deep learning to annotate the protein universe. Nat. Biotechnol.https://doi.org/10.1038/s41587-021-01179-w (2022).

[88] Rembeza, E. and Engqvist, M. K. Experimental and computational investigation of enzyme functional annotations uncovers misannotation in the ec 1.1. 3.15 enzyme class. PLoS Comput. Biol. 17, e1009446 (2021).

[89] Ozturk, H., Ozgur, A. and Ozkirimli, E. Deepdta: deep drugtarget binding affinity prediction. Bioinformatics 34, i821-i829 (2018).

[90] Feng, Q., Dueva, E., Cherkasov, A. and Ester, M. Padme: A deep learning-based framework for drug-target interaction prediction. Preprint at https://doi.org/10.48550/arXiv.1807.09741 (2018).

[91] Karimi, M., Wu, D., Wang, Z. and Shen, Y. Deep affinity: interpretable deep learning of compound–protein affinity through UNIFIED recurrent and convolutional neural networks. Bioinformatics 35, 3329-3338 (2019).

[92] Kroll, A., Engqvist, M. K., Heckmann, D. and Lercher, M. J. Deep learning allows genome-scale prediction of michaelis constants from structural features. PLoS Biol. 19, e3001402 (2021).

[93] Li, F. et al. Deep learning-based k cat prediction enables improved enzyme-constrained model reconstruction. Nat. Catal. 5, 662-672 (2022).

[94] Weininger, D. SMILES, a chemical language and information system. 1. introduction to methodology and encoding rules. J. Chem. Inf. Comput. Sci. 28, 31-36 (1988).

[95] Rogers, D. and Hahn, M. Extended-connectivity fingerprints. J. Chem. Inf. Model. 50, 742-754 (2010).

[96] Zhou, J. et al. Graph neural networks: A review of methods and applications. AI Open 1, 57-81 (2020).

[97] Yang, K. et al. Analyzing learned molecular representations for property prediction. J. Chem. Inf. Model. 59, 3370-3388 (2019).

[98] Rives, A. et al. Biological structure and function emerge from scaling unsupervised learning to 250 million protein sequences. PNAS 118, e2016239118 (2021).

[99] Alley, E. C., Khimulya, G., Biswas, S., AlQuraishi, M. and Church, G. M. Unified rational protein engineering with sequence-based deep representation learning. Nat. Methods. 16, 1315-1322 (2019).

[100] Xu, Y. et al. Deep dive into machine learning models for protein engineering. J. Chem. Inf. Model. 60, 2773–2790 (2020).

[101] Bekker, J. and Davis, J. Learning from positive and unlabeled data: A survey. Mach. Learn. 109, 719-760 (2020)

[102] Kearnes, S., McCloskey, K., Berndl, M., Pande, V. and Riley, P. Mole- cular graph convolutions: moving beyond !ngerprints. J. Comput. -Aided Mol. Des. 30, 595–608 (2016).

[103] Duvenaud, D. K. et al. Convolutional networks on graphs for learning molecular fingerprints. In Advances in Neural Information Processing Systems, 2224-2232 (2015).

[104] Zhou, J. et al. Graph neural networks: A review of methods and applications. AI Open 1, 57–81 (2020).

[105] Hu, W. et al. Strategies for pre-training graph neural networks. Preprint at https://doi.org/10.48550/arXiv.1905.12265 (2019).

[106] Capela, F., Nouchi, V., Van Deursen, R., Tetko, I. V. and Godin, G. Multitask learning on graph neural networks applied to molecular property predictions. Preprint at https://doi.org/10.48550/arXiv.1910.13124 (2019).

[107] Vaswani, A. et al. Attention is all you need. In Advances in neural information processing systems, 5998–6008 (2017).

[108] Suzek, B. E. et al. Uniref clusters: a comprehensive and scalable alternative for improving sequence similarity searches. Bioinfor- matics 31, 926–932 (2015).

[109] Elnaggar, A. et al. Prottrans: Towards cracking the language of lifes code through self-supervised deep learning and high performance computing. IEEE Trans. Pattern Anal. Mach. Intell. PP https://doi.org/10.1109/TPAMI.2021.3095381 (2021).

[110] Wittmann, B. J., Johnston, K. E., Wu, Z., and Arnold, F. H. (2021). Advances in machine learning for directed evolution. Current opinion in structural biology, 69, 11-18.