# Resumen en Extenso del Artículo:
# *Logistic Regression: a brief primer: Jill C. Stoltzfus*

Carlos

## Contents



## 1 Introducción

La regresión logística es una herramienta eficiente para analizar el efecto de un grupo de variables independientes sobre un resultado binario cuantificando la contribución única de cada variable independiente, por otra parte identifica iterativamente la combinación lineal más fuerte de variables con mayor probabilidad de identificar el resultado observado.

La regresión logística tiene sus raíces en el siglo XIX, con los trabajos de **Pierre François Verhulst**, quien introdujo la *curva logística* para modelar el crecimiento poblacional. Sin embargo, fue en el siglo XX cuando su aplicación estadística tomó forma. En 1944, **Joseph Berkson** introdujo el *modelo logit* en el contexto de bioestadística, proponiéndolo como alternativa al modelo probit. La Regresión Logística fue adoptada ampliamente en estudios biomédicos a partir de la década de 1960, gracias a su capacidad para manejar variables dicotómicas y ofrecer interpretaciones claras a través del las razones de momios. En décadas recientes, la regresión logística se ha convertido en una herramienta fundamental para el análisis de datos en epidemiología, medicina clínica, y ciencias sociales.

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).

**Nota 1.** *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.*

# 2 Regresión Logística

Existen diferentes tipos de regresión, dependiendo de los objetivos de investigación y del formato de las variables, siendo la regresión lineal una de las más utilizadas. La *regresión lineal* analiza resultados continuos y asume que la relación entre el resultado y las variables independientes sigue una forma funcional determinada. Generalmente es más deseable determinar la influencia de múltiples factores al mismo tiempo, ya que de este modo se pueden observar las contribuciones de cada variable. En este caso, la regresión lineal multivariada es la opción adecuada. La ecuación básica para la regresión lineal con variables independientes es:

$$\hat{Y} = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \ldots + \beta_i X_i. \tag{1}$$

Los componentes de esta ecuación son los siguientes: $\hat{Y}$ es el resultado continuo estimado; $\beta_0 + \beta_1 X_1 + \beta_2 X_2 + \ldots + \beta_i X_i$ es la ecuación de regresión lineal para las variables independientes del modelo, donde $\beta_0$ es la ordenada al origen o punto en el que la línea de regresión toca el eje vertical $Y$, se considera un valor constante; $\beta_1 X_1 + \beta_2 X_2 + \ldots + \beta_i X_i$ es el valor de cada variable independiente ($X_i$) ponderado por su respectivo coeficiente beta ($\beta$).

Los coeficientes beta determinan la pendiente de la línea de regresión, cuanto mayor sea el coeficiente beta, más fuerte es la contribución de dicha variable al resultado. Para una variable binaria, la regresión logística es el método usualmente elegido, la regresión logística puede incluir una o múltiples variables independientes, aunque examinar múltiples variables es generalmente más informativo, puesto que permite revelar la contribución única de cada variable ajustando por las demás. La regresión logística tiene ecuación:

$$P(\hat{Y}_i) = \frac{e^{\beta_0+\beta_1 X_1+\beta_2 X_2+\ldots+\beta_i X_i}}{1+e^{\beta_0+\beta_1 X_1+\beta_2 X_2+\ldots+\beta_i X_i}}. \tag{2}$$

Un aspecto importante de la regresión logística es que conserva muchas características de la regresión lineal en su análisis de resultados binarios. Sin embargo, existen diferencias clave entre las dos ecuaciones: $\hat{Y}_i$ representa la probabilidad estimada de pertenecer a una de las dos categorías binarias del resultado (categoría $i$) en lugar de representar un resultado continuo estimado; $e^{\beta_0+\beta_1 X_1+\beta_2 X_2+\ldots+\beta_i X_i}$ representa la ecuación de regresión lineal para las variables independientes expresadas en la escala *logit*.

La razón de esta transformación *logit* radica en los parámetros básicos del modelo de regresión logística, la escala logit resuelve este problema al transformar la ecuación de regresión lineal original para producir el logit (o logaritmo natural) de las razones de momios (odds) de estar en una categoría $(\hat{Y})$ frente a la otra categoría $(1-\hat{Y})$:

$$logit(\hat{Y}) = \ln\left(\frac{\hat{Y}}{1-\hat{Y}}\right) = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \ldots + \beta_i X_i. \tag{3}$$

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.

## 2.1 Variables independientes

**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].*

## 2.2 Estrategias de Construcción del Modelo

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.

**Nota 2.** *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.*

## 2.3 Validación Interna y Externa del Modelo

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].

## 2.4 Interpretación de los Resultados del Modelo

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 $\text{OR} = e^{\beta_i}$. Con base en esta fórmula, un cambio de una unidad en la variable independiente multiplica la probabilidad del resultado por la cantidad contenida en $e^{\beta_i}$.

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.

**Nota 3.** *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.*

**Nota 4.** *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:*

$$Recall = \frac{Verdaderos\ Positivos}{Verdaderos\ Positivos + Falsos\ Negativos}. \tag{4}$$

*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:*

$$Precision = \frac{Verdaderos\ Positivos}{Verdaderos\ Positivos + Falsos\ Positivos}. \tag{5}$$

*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:*

$$Specificity = \frac{Verdaderos\ Negativos}{Verdaderos\ Negativos + Falsos\ Positivos}. \tag{6}$$

*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.*

**Nota 5.** *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.*

# 3 Fundamentos

Para resolver el problema de regresión lineal se requiere encontrar $\beta_0$ y $\beta_1$ tal que:

$$\hat{y} = \beta_0 + \beta_1 x$$

se aproxime al mínimo de todos los posibles coeficientes de regresión $\beta_0$ y $\beta_1$:

$$(\beta_0, \beta_1) = \arg \min_{(\beta_0,\beta_1)} \sum \left[y_i - (\beta_0 + \beta_1 x_i)\right]^2$$

Que se obtiene resolviendo

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

Utilizando un modelo lineal centralizado:

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

donde:

$$\beta_0 = \beta_0^* + \beta_1 \bar{x}.$$

se requiere resolver para $\beta_0^*$. Entonces

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

Sabemos que:

$$\beta_0 = \beta_0^* + \beta_1 \bar{x} \quad \Rightarrow \quad \beta_0^* = \beta_0 - \beta_1 \bar{x}$$

por lo tanto

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

por lo tanto

$$2\sum\left[y_i-(\beta_0^*+(x_i-\bar{x})\beta_1)\right](-1)=0 \qquad \Leftrightarrow \qquad -2\sum\left[y_i-\beta_0^*-\beta_1(x_i-\bar{x})\right]=0$$
$$\Leftrightarrow \sum y_i-n\beta_0^*-\beta_1\sum(x_i-\bar{x})=0 \qquad \Leftrightarrow \qquad n\beta_0^*=\sum y_i$$
$$\text{ya que } \sum(x_i-\bar{x})=0, \text{ por lo tanto } \beta_0^*=\bar{y}.$$

Por otra parte la derivada respecto a $\beta_1$

$$\frac{\partial}{\partial\beta_1}\sum\left[y_i-(\beta_0^*+\beta_1x_i)\right]^2=0 \;\Leftrightarrow\; \frac{\partial}{\partial\beta_1}\sum\left[y_i-(\beta_0^*-\beta_1\bar{x}+\beta_1x_i)\right]^2=0$$
$$\frac{\partial}{\partial\beta_1}\sum\left[y_i-(\beta_0^*+\beta_1(x_i-\bar{x}))\right]^2=0 \;\Leftrightarrow\; 2\sum\left[y_i-(\beta_0^*+\beta_1(x_i-\bar{x}))\right](x_i-\bar{x})(-1)=0$$

de aquí

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

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

$$\sum y_i(x_i-\bar{x})-\bar{y}\sum(x_i-\bar{x})-\beta_1\sum(x_i-\bar{x})^2=0$$
$$\Leftrightarrow \sum(y_i-\bar{y})(x_i-\bar{x})-\beta_1\sum(x_i-\bar{x})^2=0$$
$$\Rightarrow \beta_1=\frac{\sum(y_i-\bar{y})(x_i-\bar{x})}{\sum(x_i-\bar{x})^2}=\frac{S_{xy}}{S_{xx}}$$

por lo tanto

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

**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 $\rho_i$. 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 la intersección está incluida en el vector $\beta$, es decir, $X_i = [1, x_i^\top]$ y $\beta = [\beta_0, \boldsymbol{\beta}^\top]$. Como $y$ es una variable aleatoria Bernoulli con probabilidad $\rho_i$, se tiene:*

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

*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*

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

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

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

y varianza:

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

**Nota 6.** *Este error no se distribuye normalmente, por lo tanto, la aproximación por mínimos cuadrados no se puede aplicar. Además, dado que $y_i \in \{0, 1\}$, el modelo reglinal puede arrojar valores mayores que 1 y menores que 0. Por tanto, es mejor el modelo logístico.*

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

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

**Nota 7.** *La función logit implícitamente coloca un hiperplano que separa:*

$$
\begin{aligned}
\beta^\top x > 0 \quad &\iff \quad \text{instancia positiva,}\\
\beta^\top x < 0 \quad &\iff \quad \text{instancia negativa.}
\end{aligned}
$$

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

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

Calculemos 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:

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

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:

$$
g(\beta) = \nabla_\beta \ln \mathcal{L}(\beta) = X^\top (y-\rho) = 0. \tag{7}
$$

Ahora calculemos la segunda derivada de $\beta$

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

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 $\mathbb{V} = \operatorname{diag}(v_1, v_2, \ldots, v_n)$, entonces

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

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

$$
\mathcal{I}(\beta) = -\mathbb{E}[\mathcal{H}(\beta)] = X^\top \mathbb{V} X, \tag{8}
$$

con Varianza:

$$\mathbb{V}(\hat{\beta}) = \mathcal{I}^{-1}(\beta) = (X^{\top}\mathbb{V}X)^{-1}. \tag{9}$$

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 $\lambda$ es el *parámetro de regularización.*

**Nota 8.**

- *Para resultados binarios, la* ***función pérdida o desviación (DEV)*** *útil para medir la bondad de ajuste del modelo es la log-verosimilitud negativa, y está dada por:*

$$\mathrm{DEV}(\hat{\beta}) = -2\ln \mathcal{L}(\hat{\beta})$$

*para mayor información ver [3].*

- *Minimizar DEV es equivalente a maximizar la log-verosimilitud;*
- *el* ***gradiente conjugado*** *aplicado a IRLS proporciona mejores resultados para estimar $\beta$ que cualquier otro método (Khalouf, 2002; Munka, 2003).*

Retomando,

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

Se actualiza la fórmula para Newton-Raphson en la iteración $(CH)$, dada por:

$$\beta^{(CH)} = \beta + \left(X^{\top}\mathbb{V}X + \lambda I\right)^{-1} X^{\top}\left(y-p\right).$$

Caso:

$$\beta^{(C)} = \left(X^{\top}\mathbb{V}X + \lambda I\right)^{-1}\left(X^{\top}\mathbb{V}Z^{(C)}\right) \tag{10}$$

por lo tanto

$$\beta^{(CH)} = \left(X^{\top}\mathbb{V}X + \lambda I\right)^{-1} X^{\top}\left(\mathbb{V}Z^{(C)} + (y-p)\right) = \left(X^{\top}\mathbb{V}X + \lambda I\right)^{-1} X^{\top}\mathbb{V}Z^{(C)} \tag{11}$$

Donde:

$$Z^{(C)} = X\beta^{(C)} + \mathbb{V}^{-1}(y-p) \quad \text{(Adjusted response, Hastie et al., 2009)} \tag{12}$$

**Nota 9.** *Si* $\mathcal{I}\left(X^\top \mathbb{V} X + \lambda I\right)$ *es densa, el cálculo iterativo puede ser extremadamente lento (Komarek, 2004). El problema de minimos cuadrados ponderados sería*

$$\left(X^\top \mathbb{V} X + \lambda I\right) \beta^{(CH)} = X^\top \mathbb{V}^{2^{(C)}}, \tag{13}$$

*que consiste en un sistema lineal de ecuaciones y variables, y resolverlo es equivalente a minimizar la función cuadrática:*

$$\frac{1}{2}\beta^\top \left(X^\top \mathbb{V} X + \lambda I\right) \beta - \beta^\top \left(X^\top \mathbb{V}^{2^{(C)}}\right). \tag{14}$$

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 $\theta_0$ y $\theta_1$ 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 $\theta_0$

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

y el gradiente de la función de costo respecto a $\theta_1$

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

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 $\alpha$ 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

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

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

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

y

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

Por lo tanto

$$\hat{\mathbf{y}} \;\; =\mathbf{X}\theta \quad \Leftrightarrow \quad \begin{bmatrix}\hat{y}^{(1)}\\ \hat{y}^{(2)}\\ \vdots\\ \hat{y}^{(N)}\end{bmatrix}=\begin{bmatrix}h_\theta\mathbf{x}^{(1)}\\ h_\theta\mathbf{x}^{(2)}\\ \vdots\\ h_\theta\mathbf{x}^{(N)}\end{bmatrix}=\begin{bmatrix}x_0^{(1)} & x_1^{(1)} & \cdots & x_d^{(1)}\\ x_0^{(2)} & x_1^{(2)} & \cdots & x_d^{(2)}\\ \vdots & \vdots & \ddots & \vdots\\ x_0^{(N)} & x_1^{(N)} & \cdots & x_d^{(N)}\end{bmatrix}\begin{bmatrix}\theta_0\\ \theta_1\\ \vdots\\ \theta_d\end{bmatrix}, \tag{15}$$

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

$$J(\theta)=\frac{1}{2N}(\mathbf{X}\theta-\mathbf{y})^\top(\mathbf{X}\theta-\mathbf{y}). \tag{16}$$

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:

$$J(\theta) = \frac{1}{2N}\left(\theta^\top\left(\mathbf{X}^\top\mathbf{X}\right)\theta - 2\left(\mathbf{X}^\top y\right)^\top\theta + \mathbf{y}^\top\mathbf{y}\right). \tag{17}$$

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 10.**
- *Resolver por gradiente descentiente, y calcular $(X^\top X)^{-1}$ puede ser difícil.*
- $(X^\top X)^\top = X^\top(X^\top)^\top = X^\top X.$

**Ejemplo 2.** *Ahora calculemos la log-verosimilitud considerando que cada variable se distribuye Bernoulli: $p\,(x|\theta) = \theta\ x\,(1-\theta)^{1-x}$ para $x = 0, 1$. Sea $x_1, x_2, \ldots, x_n$, $\quad \theta = \frac{1}{n}\sum_{i=1}^n x_i$, entonces $\theta = \sum_{i=1}^n \frac{x_i}{n}$ es el valor más probable para estimar $\theta$. La función de verosimilitud conjunta*

$$\Psi(x_1, x_2, \ldots, x_n; \theta) = \prod_{i=1}^{n} f(x_i; \theta) = \theta^{\sum_{i=1}^n x_i}(1-\theta)^{n-\sum_{i=1}^n x_i}, \tag{18}$$

*entonces la función de Log-verosimilitud está dada por:*

$$\log L(\theta) = \sum_{i=1}^{n} x_i \log\theta + (n - \sum_{i=1}^{n} x_i)\log(1-\theta), \tag{19}$$

*calculando la derivada de la log-verosimilitud*

$$\begin{aligned}
\frac{\partial \log L(\theta)}{\partial\theta} &= \frac{\sum_{i=1}^n x_i}{\theta} - \frac{n - \sum_{i=1}^n x_i}{1-\theta} && (20) \\
&= \frac{(1-\theta)\sum_{i=1}^n x_i - (n - \sum_{i=1}^n x_i)\theta}{\theta(1-\theta)} = \frac{\sum_{i=1}^n x_i - n\theta}{\theta(1-\theta)}, && (21)
\end{aligned}$$

*igualando a cero y resolviendo*

$$\frac{\sum_{i=1}^n x_i - n\theta}{\theta(1-\theta)} = 0 \Leftrightarrow \theta = \frac{\sum_{i=1}^n x_i}{n}. \tag{22}$$

**Ejemplo 3.** *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.*

$$\mathcal{L}\left(\theta\right) = \prod_{i=1}^{N} p\left(u^{(i)}; \theta\right) \tag{23}$$

*donde*

$$\theta_{ML} = \arg\max_{\theta} \mathcal{L}(\theta) = \arg\max_{\theta} \sum_{i=1}^{N} \log p\left(u^{(i)}; \theta\right) \tag{24}$$

*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* $\sigma^2$*, entonces*

$$y^{(i)} = h_\theta\left(x^{(i)}\right) + \epsilon^{(i)} = \theta^\top \mathbf{X}^{(i)} + \epsilon^{(i)}, \tag{25}$$

*por lo tanto*

$$y^{(i)} \sim N\left(\theta^\top \mathbf{X}^{(i)}, \sigma^2\right), \tag{26}$$

*entonces*

$$\begin{aligned} p\left(y|\mathbf{X}, \theta, \sigma^2\right) &= \prod_{i=1}^{N} p\left(y|\mathbf{x}^{(i)}, \theta, \sigma^2\right) = \prod_{i=1}^{N} \left(2\pi\sigma^2\right) e^{-\frac{1}{2\sigma^2}\left(y^{(i)} - \theta^\top \mathbf{x}^{(i)}\right)^2} & (27) \\ &= \left(2\pi\sigma^2\right)^{-\frac{N}{2}} e^{-\frac{1}{2\sigma^2}\sum_{i=1}^{2}\left(y^{(i)} - \theta^\top \mathbf{x}^{(i)}\right)^2} & (28) \\ &= \left(2\pi\sigma^2\right)^{-\frac{N}{2}} e^{-\frac{1}{2\sigma^2}(y - \mathbf{X}\theta)^\top (y - \mathbf{X}\theta)} & (29) \end{aligned}$$

*entonces la verosimilitud es*

$$p\left(y|\mathbf{X}, \theta, \sigma^2\right) = \left(2\pi\sigma^2\right)^{-\frac{N}{2}} e^{-\frac{1}{2\sigma^2}(y - \mathbf{X}\theta)^\top (y - \mathbf{X}\theta)} \tag{30}$$

*y la log-verosimilitud es*

$$\mathcal{L}\left(\theta, \sigma^2\right) = -\frac{N}{2} \log\left(2\pi\sigma^2\right) \left[-\frac{1}{2\sigma^2} (y - \mathbf{X}\theta)^\top (y - \mathbf{X}\theta)\right]. \tag{31}$$

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

Se define la función sigmoide

$$\sigma(u) = \frac{1}{1 + e^{-u}} \quad \Rightarrow \text{logistic regression classifier} \tag{32}$$

donde la regla de decisión para $y$

$$y = \sigma(h_{\boldsymbol{\theta}}(x)) = \sigma(\boldsymbol{\theta}^\top x) \tag{33}$$

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

$$\begin{aligned} p\left(y^{(i)} = 1 \mid x^{(i)}; \boldsymbol{\theta}\right) &= \sigma(\boldsymbol{\theta}^\top x^{(i)}) & (34) \\ p\left(y^{(i)} = 0 \mid x^{(i)}; \boldsymbol{\theta}\right) &= 1 - \sigma(\boldsymbol{\theta}^\top x^{(i)}) & (35) \end{aligned}$$

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

$$p\left(y^{(i)} \mid x^{(i)}; \boldsymbol{\theta}\right) = \sigma(\boldsymbol{\theta}^\top x^{(i)})^{y^{(i)}} \cdot \left[1 - \sigma(\boldsymbol{\theta}^\top x^{(i)})\right]^{1 - y^{(i)}} \tag{36}$$

mientras que la probabilidad conjunta de todas las etiquetas

$$\prod_{i=1}^{N} \sigma(\boldsymbol{\theta}^\top x^{(i)})^{y^{(i)}} \left(1 - \sigma(\boldsymbol{\theta}^\top x^{(i)})\right)^{(1-y^{(i)})} \tag{37}$$

La log-verosimilitud para regresión logística está dada por:

$$\ell(\boldsymbol{\theta}) = \sum_{i=1}^{N} y^{(i)} \log\left(\sigma(\boldsymbol{\theta}^\top x^{(i)})\right) + (1 - y^{(i)}) \log\left(1 - \sigma(\boldsymbol{\theta}^\top x^{(i)})\right) \tag{38}$$

Antes de calcular la derivada, recordemos:

$$\sigma(z) = \frac{1}{1 + e^{-z}} \tag{39}$$

con derivada

$$\frac{d}{dz}\sigma(z) = \frac{d}{dz}\left(1 + e^{-z}\right)^{-1} = -(1 + e^{-z})^{-2} \cdot (-e^{-z}) = \frac{e^{-z}}{(1 + e^{-z})^2} \tag{40}$$

además:

$$\begin{aligned} \sigma(z) &= \frac{1}{1 + e^{-z}} \Rightarrow 1 - \sigma(z) = \frac{e^{-z}}{1 + e^{-z}} && (41) \\ &\Rightarrow \sigma(z)(1 - \sigma(z)) = \frac{e^{-z}}{(1 + e^{-z})^2} && (42) \\ &\therefore \frac{d}{dz}\sigma(z) = \sigma(z)(1 - \sigma(z)) && (43) \end{aligned}$$

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

$$\begin{aligned} \frac{\partial \ell(\boldsymbol{\theta})}{\partial \theta_j} &= \sum_{i=1}^{N} \left[y^{(i)} \log \sigma(\boldsymbol{\theta}^\top x^{(i)}) + (1 - y^{(i)}) \log(1 - \sigma(\boldsymbol{\theta}^\top x^{(i)}))\right] && (44) \\ &= \sum_{i=1}^{N} \left[\frac{y^{(i)}}{\sigma(\boldsymbol{\theta}^\top x^{(i)})} - \frac{1 - y^{(i)}}{1 - \sigma(\boldsymbol{\theta}^\top x^{(i)})}\right] \cdot \frac{d}{d\theta_j}\sigma(\boldsymbol{\theta}^\top x^{(i)}) && (45) \\ &= \sum_{i=1}^{N} \left[\frac{y^{(i)}}{\sigma(\boldsymbol{\theta}^\top x^{(i)})} - \frac{1 - y^{(i)}}{1 - \sigma(\boldsymbol{\theta}^\top x^{(i)})}\right] \cdot \sigma(\boldsymbol{\theta}^\top x^{(i)})\sigma(1 - \boldsymbol{\theta}^\top x^{(i)})x_j^{(i)} && (46) \\ &= \sum_{i=1}^{N} \left[\frac{y^{(i)} - \sigma(\boldsymbol{\theta}^\top x^{(i)})}{\sigma(\boldsymbol{\theta}^\top x^{(i)})(1 - \sigma(\boldsymbol{\theta}^\top x^{(i)}))}\right] \cdot \sigma(\boldsymbol{\theta}^\top x^{(i)})\sigma(1 - \boldsymbol{\theta}^\top x^{(i)})x_j^{(i)} && (47) \\ &= \sum_{i=1}^{N} \left[y^{(i)} - \sigma(\boldsymbol{\theta}^\top x^{(i)})\right] x_j^{(i)} && (48) \end{aligned}$$

la cual es la función recursiva para calcular el gradiente.

# 4 Apendice

## 4.1 Fundamentos de la Regresión Logística

Dado un conjunto de datos $(x_i, y_i)$ para $i = 1, \ldots, n$, queremos encontrar los coeficientes $\beta_0$ y $\beta_1$ que minimicen el error cuadrático:

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

Es más sencillo resolver este problema usando la forma *centralizada*:

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

donde:

$$\beta_0 = \beta_0^* - \beta_1 \bar{x},$$

se quiere minimizar la función de versosimilitud:

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

por ello debemos de calcular la derivada parcial:

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

Igualando a cero y resolviendo:

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

dado que $\sum_{i=1}^{n}(x_i - \bar{x}) = 0$, se tiene

$$n\beta_0^* = \sum_{i=1}^{n} y_i \Rightarrow \beta_0^* = \bar{y}.$$

Ahora derivamos con respecto a $\beta_1$:

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

Igualando a cero y resolviendo

$$\begin{aligned}
\sum_{i=1}^{n}[y_i - \beta_0^* - \beta_1(x_i - \bar{x})](x_i - \bar{x}) &= 0 \text{ recordemos que } \beta_0^* = \bar{y}, \\
\sum_{i=1}^{n}[y_i - \bar{y} - \beta_1(x_i - \bar{x})](x_i - \bar{x}) &= 0 \\
\sum_{i=1}^{n}(y_i - \bar{y})(x_i - \bar{x}) - \beta_1 \sum_{i=1}^{n}(x_i - \bar{x})^2 &= 0
\end{aligned}$$

Despejamos $\beta_1$:

$$\beta_1 = \frac{\sum_{i=1}^{n}(y_i - \bar{y})(x_i - \bar{x})}{\sum_{i=1}^{n}(x_i - \bar{x})^2} = \frac{S_{xy}}{S_{xx}}.$$

Por lo tanto

$$\beta_0 = \beta_0^* - \beta_1\bar{x} = \bar{y} - \beta_1\bar{x}.$$

**Ejemplo 4.** *Para una variable binaria $y_i \in \{0,1\}$, el modelo de regresión logística predice la probabilidad:*

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

*con función de verosimilitud*

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

*y Log-verosimilitud*

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

*Usando que $p_i = \frac{1}{1+e^{-x_i\beta}}$, entonces:*

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

*con Gradiente y Hessiano*
*Partimos de:*

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

*Derivando con respecto a $\beta_j$:*

$$\frac{\partial \ell(\beta)}{\partial \beta_j} = \sum_{i=1}^{n} \left[y_i x_{ij} - \frac{e^{x_i\beta}}{1 + e^{x_i\beta}} x_{ij}\right] = \sum_{i=1}^{n} x_{ij}(y_i - p_i)$$

*donde la forma vectorial del gradiente:*

$$\nabla_\beta \ell(\beta) = X^\top(\mathbf{y} - \mathbf{p})$$

*La derivada del gradiente es:*

$$\frac{\partial^2 \ell(\beta)}{\partial \beta_j \partial \beta_k} = -\sum_{i=1}^{n} x_{ij}x_{ik}p_i(1 - p_i)$$

*Forma matricial:*

$$\nabla_\beta^2 \ell(\beta) = -X^\top V X, \quad \textit{donde } V = \textit{diag}(p_i(1 - p_i))$$

*Penalización L2 añadida a la log-verosimilitud:*

$$\ell_\lambda(\beta) = \ell(\beta) - \frac{\lambda}{2}\|\beta\|^2$$

*Gradiente regularizado:*

$$\nabla_\beta \ell_\lambda(\beta) = X^\top(\mathbf{y} - \mathbf{p}) - \lambda\beta$$

*Hessiano regularizado:*

$$\nabla^2_\beta \ell_\lambda(\beta) = -X^\top V X - \lambda I$$

*Actualización de Newton truncado:*

$$\beta^{(k+1)} = \beta^{(k)} + s$$

*Donde $s$ resuelve:*

$$(X^\top V X + \lambda I)s = X^\top V z - \lambda\beta$$

*Ajuste del intercepto (King & Zeng)*

$$\tilde{\beta}_0 = \hat{\beta}_0 - \ln\left[\left(\frac{1-\tau}{\tau}\right)\left(\frac{\hat{y}}{1-\hat{y}}\right)\right]$$

*Peso para observación $i$:*

$$w_i = \frac{Q_i}{H_i}$$

*Verosimilitud ponderada:*

$$\ell_w(\beta) = \sum_{i=1}^{n} w_i \ln\left(\frac{e^{x_i\beta}}{1+e^{x_i\beta}}\right)$$

*Log-verosimilitud penalizada (ajuste de Jeffreys):*

$$\ell^*(\beta) = \ell(\beta) + \frac{1}{2}\log|I(\beta)|$$

*Regla de decisión*

$$\hat{y}_i = \begin{cases} 1 & si\ p_i \geq c \\ 0 & si\ p_i < c \end{cases}, \quad con\ c = 0.5$$

La RL modela la probabilidad de un evento binario $y_i \in \{0,1\}$ en función de un vector de predictores $x_i$ mediante:

$$p_i = \frac{1}{1+e^{-x_i\beta}} \tag{49}$$

La verosimilitud del modelo es:

$$L(\beta) = \prod_{i=1}^{n} p_i^{y_i}(1-p_i)^{1-y_i} \tag{50}$$

Y su log-verosimilitud:

$$\ell(\beta) = \sum_{i=1}^{n} \left[y_i \ln(p_i) + (1-y_i)\ln(1-p_i)\right] \tag{51}$$

- Gradiente:
$$\nabla_\beta \ell(\beta) = X^T(\mathbf{y} - \mathbf{p}) \tag{52}$$

- Hessiano:
$$\nabla^2_\beta \ell(\beta) = -X^T V X, \quad \text{donde } V = \text{diag}(p_i(1-p_i)) \tag{53}$$

Para evitar el sobreajuste, se añade un término de penalización L2 (ridge):

$$\ell_\lambda(\beta) = \ell(\beta) - \frac{\lambda}{2}\|\beta\|^2 \tag{54}$$

- Gradiente regularizado:
$$\nabla_\beta \ell_\lambda(\beta) = X^T(\mathbf{y} - \mathbf{p}) - \lambda\beta \tag{55}$$

- Hessiano regularizado:
$$\nabla^2_\beta \ell_\lambda(\beta) = -X^T V X - \lambda I \tag{56}$$

Una técnica común para estimar los parámetros del modelo es IRLS, que utiliza pesos $v_i = p_i(1-p_i)$ y variables ajustadas $z_i$:

$$z_i = x_i\hat{\beta} + \frac{y_i - p_i}{v_i} \tag{57}$$

En cada iteración, se resuelve:

$$(X^T V X + \lambda I)\hat{\beta}^{(c+1)} = X^T V z^{(c)} \tag{58}$$

Este método es eficiente para bases de datos de tamaño moderado.
En problemas a gran escala, se recomienda el método del gradiente conjugado:

- Se inicializa el residuo $r^{(0)} = b - A\beta^{(0)}$.
- Se actualizan las direcciones de búsqueda y pasos óptimos iterativamente.
- Permite resolver sistemas lineales sin invertir matrices.

Es especialmente útil cuando $X^T V X$ es grande o disperso.

- **Ajuste del intercepto:** basado en la tasa real de eventos:
$$\tilde{\beta}_0 = \hat{\beta}_0 - \ln\left(\frac{1-\tau}{\tau} \cdot \frac{y}{1-y}\right) \tag{59}$$

- **Ponderación:** modifica la verosimilitud con pesos:
$$\ell(\beta|y, X) = \sum_{i=1}^{n} w_i \ln\left(\frac{e^{x_i\beta}}{1+e^{x_i\beta}}\right) \tag{60}$$

- La regresión logística es robusta y se adapta bien a diferentes contextos de datos.
- Las técnicas de regularización y los métodos numéricos como IRLS y CG la hacen escalable.
- Las correcciones para eventos raros mejoran la inferencia en muestras sesgadas.

- Es una herramienta base para modelos más complejos como regresión multinomial o clasificación ordinal.

- La función logística transforma la probabilidad de un evento en **odds**, y posteriormente en **log-odds** (*logit*), acotando los valores entre 0 y 1. Esto asegura interpretaciones coherentes para eventos dicotómicos. El modelo toma la forma:

$$\text{logit}(p) = \log\left(\frac{p}{1-p}\right) = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \ldots + \beta_k X_k$$

- **Evaluación general** Se emplean dos pruebas principales:
  - **Razón de verosimilitudes (likelihood ratio test)**: compara un modelo completo con uno nulo, evaluando si los predictores mejoran significativamente la predicción.
  - **Prueba de Hosmer-Lemeshow**: mide el ajuste entre valores observados y esperados por deciles de riesgo. Un valor $p > 0.05$ indica buen ajuste.

- **Evaluación de predictores** La significancia individual de cada predictor se evalúa con el **estadístico Wald**, basado en la relación entre el coeficiente estimado y su error estándar. También se puede usar la razón de verosimilitudes para cada predictor.

- **Exactitud Predictiva y Discriminación**
  - **Tabla de clasificación**: compara predicciones contra observaciones reales, generando métricas como sensibilidad, especificidad, precisión y valor predictivo.
  - **Curva ROC (Receiver Operating Characteristic)**: representa la sensibilidad frente a 1 - especificidad. El área bajo la curva (AUC) cuantifica la capacidad discriminativa del modelo. Un AUC de 0.5 indica clasificación aleatoria, mientras que 1.0 representa clasificación perfecta.

- **Validación del Modelo**

  Se destaca la importancia de validar los modelos, ya sea de manera **interna** (con subconjuntos del mismo conjunto de datos) o **externa** (con nuevos datos). Se discuten métodos como *bootstrap*, *jackknife* y validación cruzada.

  También se mencionan medidas como:
  - $R^2$ **de Cox & Snell**
  - $R^2$ **de Nagelkerke**

  Estas proporcionan información sobre el poder explicativo del modelo, aunque no son equivalentes al $R^2$ clásico de regresión lineal.

Para una variable binaria $y_i \in \{0, 1\}$, el modelo de regresión logística predice la probabilidad:

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

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

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

**Gradiente**:

$$\nabla_\beta \ell(\beta) = X^\top (y - p)$$

**Hessiano (segunda derivada)**:

$$\nabla^2_\beta \ell(\beta) = -X^\top V X, \quad V = \mathrm{diag}(p_i(1 - p_i))$$

Se añade un término penalizado:

$$\ell_\lambda(\beta) = \ell(\beta) - \frac{\lambda}{2} \|\beta\|^2$$

Gradiente regularizado:

$$\nabla_\beta \ell_\lambda(\beta) = X^\top (y - p) - \lambda \beta$$

Hessiano regularizado:

$$\nabla^2_\beta \ell_\lambda(\beta) = -X^\top V X - \lambda I$$

Actualización generalizada del paso de Newton:

$$\beta^{(k+1)} = \beta^{(k)} + s$$

Donde $s$ es solución de:

$$(X^\top V X + \lambda I) s = X^\top V z - \lambda \beta$$

5

## Ajuste del intercepto (King & Zeng)

$$\tilde{\beta}_0 = \hat{\beta}_0 - \ln \left[ \left( \frac{1 - \tau}{\tau} \right) \left( \frac{\hat{y}}{1 - \hat{y}} \right) \right]$$

Peso para observación $i$:

$$w_i = \frac{Q_i}{H_i}$$

Verosimilitud ponderada:

$$\ell_w(\beta) = \sum_{i=1}^{n} w_i \ln \left( \frac{e^{x_i \beta}}{1 + e^{x_i \beta}} \right)$$

Se basa en el ajuste de penalización de tipo Jeffreys:

$$\ell^*(\beta) = \ell(\beta) + \frac{1}{2} \log |I(\beta)|$$

$$\hat{y}_i = \begin{cases} 1 & \text{si } p_i \geq c \\ \\ 0 & \text{si } p_i < c \end{cases} \qquad (\text{usualmente } c = 0.5)$$

**Nota 11.** *La prueba* $\boldsymbol{F}$ *es más estable que otras pruebas.*

$$F = \frac{\sum(\hat{y}_i - \bar{y}_i)^2/(m-1)}{\sum(y_i - \hat{y}_i)^2/(n-m)} \sim F(m-1, n-m) \tag{61}$$

*Donde si* $F > F_\alpha(m-1, n-m)$*, entonces una relación lineal significativa entre* $y$ *y* $x_1, x_2, \ldots, x_m$ *es razonable desde la perspectiva de la prueba de hipótesis con prioridad* $\alpha$*.*

*mientras que la prueba* $t$ *está dada por*

$$t_j = \frac{\beta_j}{\sqrt{c_{jj}}\sqrt{\frac{\sum(y_i - \hat{y}_i)^2}{n-m}}}, \tag{62}$$

*donde* $c_{jj}$ *es la componente* $j$*-ésima de la matriz inversa* $(X^\top X)^{-1}$*.*

# 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.