Notas sobre Regresión Logística
Breve introducción con aplicaciones

Carlos E Martinez-Rodriguez

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.*

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 β_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 (β).

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.

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 [***?***, ***?***].*

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 [**?**, **?**]. 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[**?**].*

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.

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; [**?**, **?**]

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

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;[**?**] y

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

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 [**?**, **?**].

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 [**?**, **?**, **?**].

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 [**?**, **?**].

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 [**?**] y Kuss [**?**]. 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* [**?**].

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^{β_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.

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.

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

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

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

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

Sabemos que:

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

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

entonces

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

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

$$\beta_0^*=\bar{y}. \tag{7}$$

Por otra parte, la derivada respecto a β_1:

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

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

$$\beta_1=\frac{\sum(y_i-\bar{y})(x_i-\bar{x})}{\sum(x_i-\bar{x})^2}=\frac{S_{xy}}{S_{xx}}, \tag{8}$$

por lo tanto

$$\boxed{\beta_1=\frac{S_{xy}}{S_{xx}},} \quad \text{y} \quad \boxed{\beta_0=\bar{y}-\beta_1\bar{x}.} \tag{9}$$

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

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

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:

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

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

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

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

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

donde

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

por lo tanto

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

luego, se tiene la respuesta ajustada (Hastie et al., 2009)

$$Z^{(C)} = X\beta^{(C)} + \mathbb{V}^{-1}(y - p).$$

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

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

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

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

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*

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

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 6. $(X^\top X)^\top = X^\top(X^\top)^\top = X^\top X$.