Iván García Mestiza, Jessica Rubí Lara Rosales,
Samuel Gurrola Viramontes

CIMAT

20 de mayo del 2025

Regresión no Lineal

Sección 1: **Introducción**

Una tarea importante en estadística es encontrar las relaciones, si existen, en un conjunto de variables cuando al menos una es aleatoria, estando sujeta a fluctuaciones aleatorias y posiblemente a errores de medición.
Usualmente se puede esperar tener una relación del estilo

$$\boldsymbol{y} \approx f(\boldsymbol{x}_1, \boldsymbol{x}_2, \ldots, \boldsymbol{x}_k).$$

Sin embargo, los datos frecuentemente contendrán ruido o errores de medición.

Los modelos de regresión lineal $y = x^T\beta + \varepsilon$ son una herramienta poderosa para cuantificar el impacto de covariables x, sobre una respuesta y.
Los modelos de regresión no-lineal extienden estos modelos a **casos en los que se cuenta con información acerca de la naturaleza del fenómeno**, buscando estimar $\boldsymbol{\theta}$ a partir de la relación

$$\boldsymbol{y} = f(\boldsymbol{x}; \boldsymbol{\theta}) + \varepsilon,$$

en donde f es una función no lineal en $\boldsymbol{\theta}$.

Modelo exponencial de crecimiento

Malthus (economista británico) propuso que la *tasa de cambio* de una población con respecto al tiempo es proporcional a la población.

$$\frac{dy}{dt} = ky.$$

Si la población inicial $y(0) = \alpha$, se tiene que la solución es

$$y = \alpha e^{kt}.$$

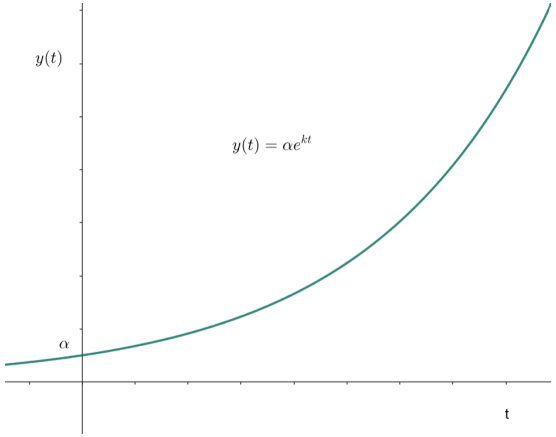


Figura: Crecimiento exponencial

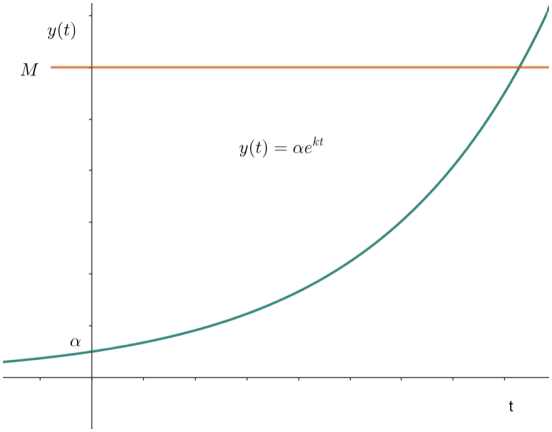


Figura: Crecimiento exponencial con una barrera

Modelo logístico de crecimiento

Verhulst consideró una variante en la que el ambiente no puede soportar más que cierta población máxima M, así que la población crece de manera acelerada y a partir de cierto punto el tamaño de la población se ralentiza.

$$\frac{dy}{dt} = ky\left(1 - \frac{y}{M}\right).$$

donde $y(t) \to M$ cuando $t \to \infty$. Suponiendo que $y(0) = \alpha$, usando el método de separación de variables obtenemos

$$y(t) = \frac{\alpha M}{(M-\alpha)e^{-kt} + \alpha} = \frac{M}{e^{-k(t-t_0)} + 1}$$

El modelo logístico, con su característico comportamiento sigmoidal es muy utilizado como modelo de crecimiento.

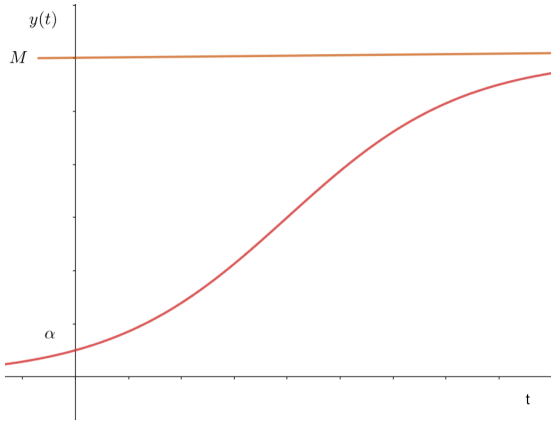
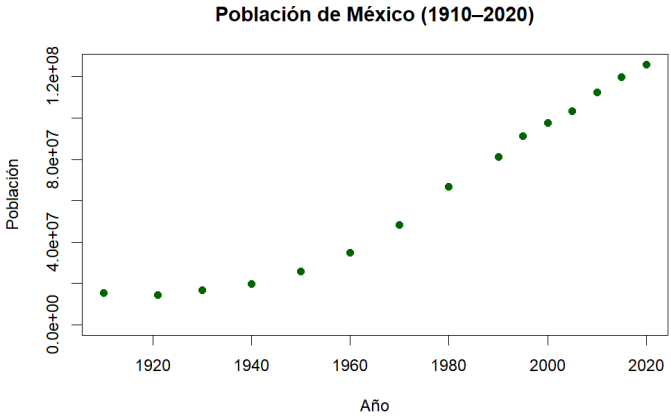


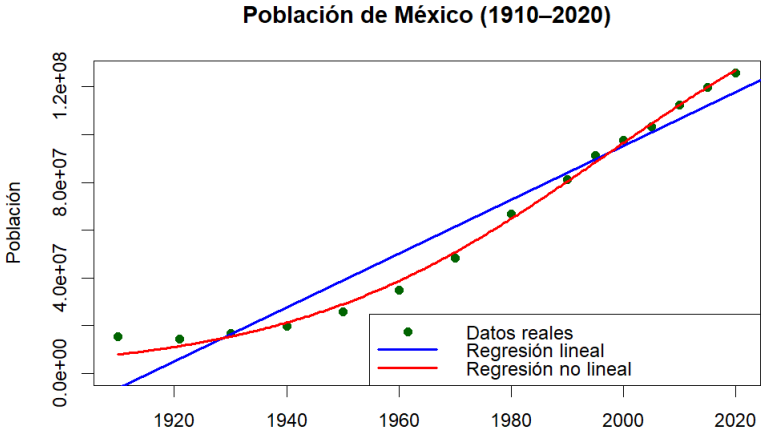
Figura: Crecimiento logístico

Base de datos del INEGI de las personas (nacionales o extranjeros) que residen en México.

Año	Población en México
1910	15160369
1921	14334780
1930	16552722
1940	19653552
1950	25791017
1960	34923129
1970	48225238
1980	66846833
1990	81249645
1995	91158290
2000	97483412
2005	103263388
2010	112336538
2015	119938473
2020	126014024



$$y(t) = \frac{M}{1 + e^{-k(t-t_0)}} \qquad M = 181,900,000,\ k = 0.03583 \text{ y } t_0 = 1997$$



En general, escribiremos el modelo de regresión no lineal como

$$y = f(\boldsymbol{x}, \boldsymbol{\theta}) + \varepsilon$$

- y variable de respuesta o dependientes.
- $\boldsymbol{x}$ variable explicativa o regresora.
- $\boldsymbol{\theta} = (\theta_1, \ldots, \theta_p)^\top$ es un vector de tamaño $p \times 1$ de parámetros desconocidos.
- f es una función no lineal de los parámetros.
- ε es un término de error aleatorio no correlacionado con $\varepsilon \sim N(0, \sigma^2)$.

Dado que

$$\begin{aligned} \mathbb{E}(\boldsymbol{y}) &= \mathbb{E}[f(\boldsymbol{x}, \boldsymbol{\theta}) + \varepsilon] \\ &= f(\boldsymbol{x}, \boldsymbol{\theta}) \end{aligned}$$

llamamos a $f(\boldsymbol{x}, \boldsymbol{\theta})$ la **función de respuesta esperada**.

Lineal

$$\boldsymbol{y} = \beta_0 + \beta_1 \boldsymbol{z}_1 + \cdots + \beta_{p-1} \boldsymbol{z}_{p-1} + \boldsymbol{\varepsilon},$$

se admiten transformaciones como $x_i x_j$, $\exp(x_i)$, $\sqrt{x_i}$ y $\sin(x_i)$. Son lineales en los parámetros desconocidos β_i, $i = 1, 2, \ldots, p$.

No Lineal

$$y = \theta_1 e^{\theta_2 x} + \varepsilon, \qquad y = \frac{\theta_1 \theta_3}{(\theta_3 - \theta_1) e^{-\theta_2 x} + \theta_1} + \varepsilon$$

No lineal en los parámetros

$$\boldsymbol{y} = f(\boldsymbol{x}, \boldsymbol{\beta}) + \boldsymbol{\varepsilon} = \beta_0 + \sum_{j=1}^{k} \beta_j x_j + \boldsymbol{\varepsilon}.$$

Ahora,

$$\frac{\partial f(\boldsymbol{x}, \boldsymbol{\beta})}{\partial \beta_j} = x_j, \quad j = 0, 1, \ldots, k,$$

donde $\boldsymbol{x}_0 \equiv 1$. Nótese que en el caso lineal las derivadas son constantes con respecto a $\boldsymbol{\beta}$. Por otro lado, consideremos el modelo no lineal

$$y = f(x, \boldsymbol{\theta}) + \varepsilon = \theta_1 e^{\theta_2 x} + \varepsilon.$$

Las derivadas de la función expectativa con respecto a θ_1 y θ_2 son:

$$\frac{\partial f(x, \boldsymbol{\theta})}{\partial \theta_1} = e^{\theta_2 x} \quad \text{y} \quad \frac{\partial f(x, \boldsymbol{\theta})}{\partial \theta_2} = \theta_1 x e^{\theta_2 x}.$$

Dado que las derivadas son función de los parámetros desconocidos θ_1 y θ_2, el modelo es no lineal.

Sección 2: **Métodos de Estimación**

A veces es útil considerar una **transformación** que induzca linealidad en la función de expectativa del modelo. Por ejemplo, en el modelo

$$y = f(x, \boldsymbol{\theta}) + \varepsilon = \theta_1 e^{\theta_2 x} + \varepsilon \tag{1}$$

se tiene que $\mathbb{E}(y) = f(x, \theta) = \theta_1 e^{\theta_2 x}$, por lo que podemos linealizar la función de expectativa tomando logaritmos:

$$\ln \mathbb{E}(y) = \ln \theta_1 + \theta_2 x.$$

Luego, podemos pensar en ajustar el modelo

$$\ln y = \ln \theta_1 + \theta_2 x + \varepsilon = \beta_0 + \beta_1 x + \varepsilon \tag{2}$$

y usar regresión lineal simple para estimar β_0 y β_1.
Sin embargo, en general las estimaciones de los parámetros bajo transformaciones no son equivalentes a las estimaciones del modelo general.
Además, en la ecuación (1) la estructura del error es aditiva, por lo que tomar logaritmos no puede producir el modelo en la ecuación (2).

Si la estructura del error es multiplicativa, por ejemplo

$$y = \theta_1 e^{\theta_2 x} \varepsilon$$

entonces tomar logaritmos será apropiado, ya que

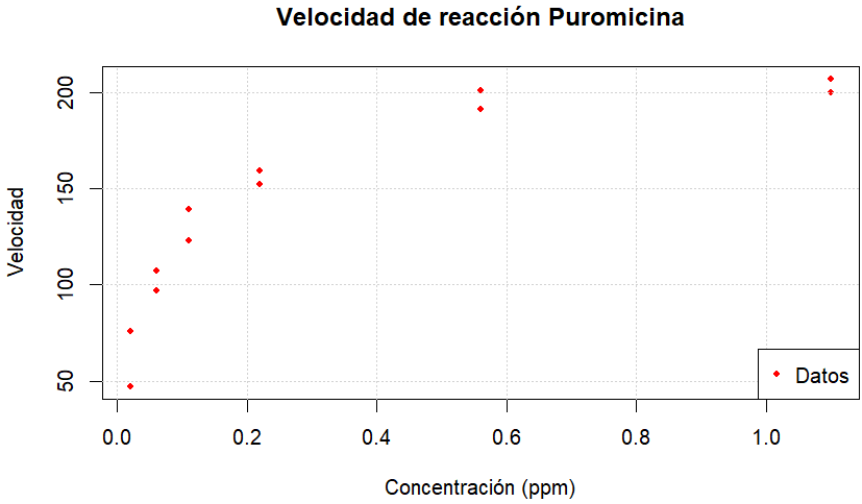
$$\ln y = \ln \theta_1 + \theta_2 x + \ln \varepsilon = \beta_0 + \beta_1 x + \varepsilon^*$$

y si ε^* sigue una distribución normal, todas las propiedades estándar del modelo de regresión lineal y la inferencia asociada serán aplicables.
El problema a menudo gira en torno a la estructura del error: ¿se aplican los supuestos estándar sobre los errores al modelo no lineal original o al linealizado?

El modelo de **Michaelis–Menten** es un modelo de cinética química que relaciona la velocidad inicial de una reacción enzimática con la concentración de sustrato x. Dicho modelo es

$$y = \frac{\theta_1 x}{x + \theta_2} + \varepsilon. \tag{3}$$

Se tienen datos de la velocidad inicial de una reacción para una enzima tratada con puromicina, y se desean estimar los coeficientes θ_1 y θ_2.


Velocidad de reacción Puromicina
Velocidad
50
100
150
200
0.0
0.2
0.4
0.6
0.8
1.0
Concentración (ppm)
Datos

La función de respuesta esperada puede ser linealizada fácilmente como sigue:

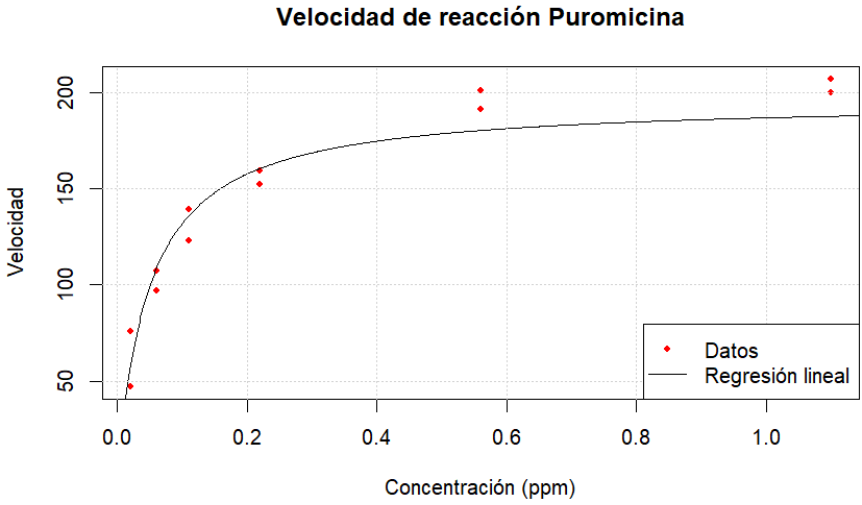
$$\frac{1}{f(x,\boldsymbol{\theta})} = \frac{x+\theta_2}{\theta_1 x} = \frac{1}{\theta_1} + \frac{\theta_2}{\theta_1} x.$$

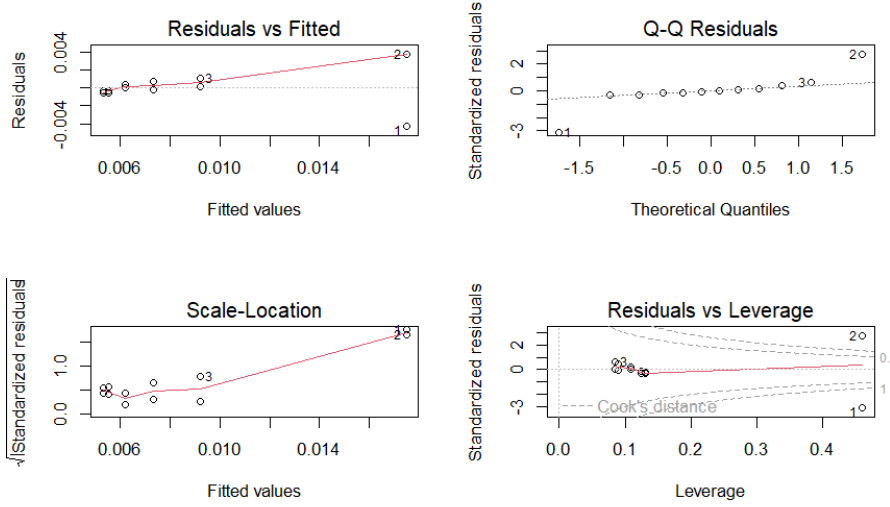
Por lo tanto, un primer acercamiento es ajustar el modelo lineal

$$y^\star = \beta_0 + \beta_1 u + \varepsilon,$$

en donde $y^\star = 1/y$ y $u = 1/x$. El modelo lineal ajustado resulta ser:

$$y^\star = 0.0051072 + 0.0002472u.$$


Velocidad de reacción Puromicina
Velocidad
50
100
150
200
0.0
0.2
0.4
0.6
0.8
1.0
Concentración (ppm)
Datos
Regresión lineal


Residuals vs Fitted
Residuals
Fitted values
Q-Q Residuals
Standardized residuals
Theoretical Quantiles
Scale-Location
√|Standardized residuals|
Fitted values
Residuals vs Leverage
Standardized residuals
Cook's distance
Leverage

Al hacer una prueba de Anderson-Darling, obtenemos un p-valor de 0.006107, y como se puede observar en las gráficas anteriores, el ajuste no es muy bueno.
El modelo lineal tiene como parámetros ajustados $\theta_1 = 195.80270885$ y $\theta_2 = 0.04840653$.
Posteriormente se verá su relación con los parámetros obtenidos al realizar regresión no lineal.

El modelo no lineal es

$$y_i = f(\boldsymbol{x}_i, \boldsymbol{\theta}) + \varepsilon_i, \quad i = 1, 2, \ldots, n,$$

donde ahora $\boldsymbol{x}_i = (x_{i1}, x_{i2}, \ldots, x_{ik})$ para $i = 1, 2, \ldots, n$. La función de mínimos cuadrados es

$$S(\boldsymbol{\theta}) = \sum_{i=1}^{n} \left(y_i - f(\boldsymbol{x}_i, \boldsymbol{\theta})\right)^2.$$

Diferenciamos la ecuación anterior con respecto a cada elemento de $\boldsymbol{\theta}$, lo cual proporcionará un conjunto de p ecuaciones normales para la situación de regresión no lineal, que son:

$$\sum_{i=1}^{n} \left(y_i - f(\boldsymbol{x}_i, \boldsymbol{\theta})\right) \left[\frac{\partial f(\boldsymbol{x}_i, \boldsymbol{\theta})}{\partial \theta_j}\right]_{\boldsymbol{\theta}=\hat{\boldsymbol{\theta}}} = 0 \quad \text{para } j = 1, 2, \ldots, p.$$

Consideremos el modelo de regresión no lineal

$$y = \theta_1 e^{\theta_2 x} + \varepsilon.$$

Las ecuaciones normales de mínimos cuadrados para este modelo son

$$\sum_{i=1}^{n} \left(y_i - \hat{\theta}_1 e^{\hat{\theta}_2 x_i} \right) e^{\hat{\theta}_2 x_i} = 0, \qquad \sum_{i=1}^{n} \left(y_i - \hat{\theta}_1 e^{\hat{\theta}_2 x_i} \right) \hat{\theta}_1 x_i e^{\hat{\theta}_2 x_i} = 0.$$

Después de simplificar, las ecuaciones normales quedan:

$$\sum_{i=1}^{n} y_i e^{\hat{\theta}_2 x_i} - \hat{\theta}_1 \sum_{i=1}^{n} e^{2\hat{\theta}_2 x_i} = 0, \qquad \sum_{i=1}^{n} y_i x_i e^{\hat{\theta}_2 x_i} - \hat{\theta}_1 \sum_{i=1}^{n} x_i e^{2\hat{\theta}_2 x_i} = 0.$$

Estas ecuaciones no son lineales en $\hat{\theta}_1$ y $\hat{\theta}_2$, y no existe una solución cerrada simple.

Nota:

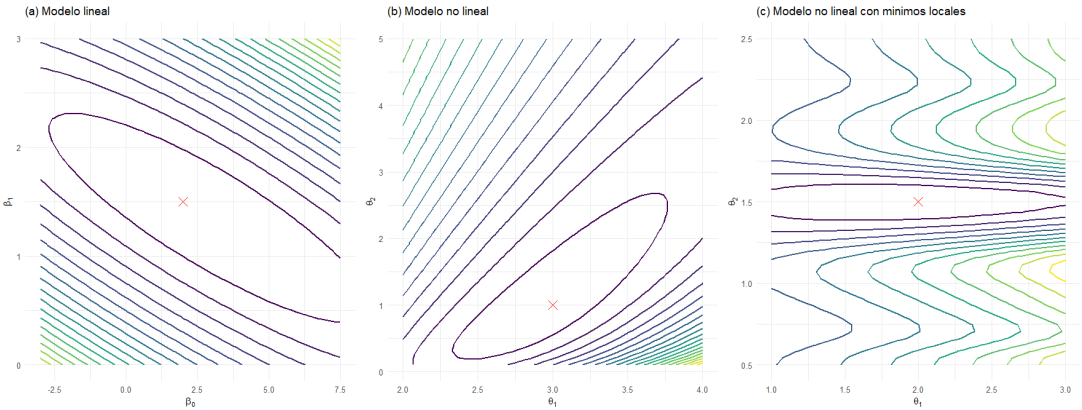
- Las derivadas de la función de respuesta esperada serán funciones de los parámetros desconocidos.
- La función f también es una función no lineal, por lo que las ecuaciones normales pueden ser muy difíciles de resolver.
- Al igual que en el modelo de regresión lineal, se espera que el número de datos n sea mayor que el número de parámetros a estimar p.

Como los errores $\varepsilon_i \sim N(0, \sigma^2)$ son iid, la función de verosimilitud es:

$$L(\boldsymbol{\theta}, \sigma^2) = \frac{1}{(2\pi\sigma^2)^{n/2}} \exp\left[-\frac{1}{2\sigma^2}\sum_{i=1}^{n}[y_i - f(\boldsymbol{x}_i; \boldsymbol{\theta})]^2\right].$$

Maximizar esta función de verosimilitud es equivalente a minimizar la suma de cuadrados residuales. Por lo tanto, las estimaciones por mínimos cuadrados son iguales a las estimaciones de máxima verosimilitud.

Geometría de Mínimos Cuadrados Lineales y No Lineales



Un método ampliamente utilizado en algoritmos computacionales para regresión no lineal es la **linealización** de la función no lineal seguida de algún método para estimación de parámetros.
Se logra mediante una expansión en **series de Taylor** de $f(\boldsymbol{x}_i, \boldsymbol{\theta})$ alrededor del punto $\boldsymbol{\theta}_0 = (\theta_1^0, \theta_2^0, \ldots, \theta_p^0)$, conservando sólo los términos lineales. De aquí se obtiene el modelo aproximado

$$f(\boldsymbol{x}_i, \boldsymbol{\theta}) \approx f(\boldsymbol{x}_i, \boldsymbol{\theta}_0) + \sum_{j=1}^{p} \left[\frac{\partial f(\boldsymbol{x}_i, \boldsymbol{\theta})}{\partial \theta_j} \right]_{\boldsymbol{\theta}=\boldsymbol{\theta}_0} (\theta_j - \theta_j^0). \tag{4}$$

Si definimos

$$
\begin{aligned}
f_i^0 &= f(\boldsymbol{x}_i, \boldsymbol{\theta}_0), \\
\beta_j^0 &= \theta_j - \theta_j^0, \\
X_{ij}^0 &= \left[\frac{\partial f(\boldsymbol{x}_i, \boldsymbol{\theta})}{\partial \theta_j} \right]_{\boldsymbol{\theta}=\boldsymbol{\theta}_0},
\end{aligned}
$$

observamos que el modelo de regresión no lineal puede escribirse como:

$$
y_i - f_i^0 = \sum_{j=1}^{p} \beta_j^0 X_{ij}^0 + \varepsilon_i, \quad i = 1, 2, \ldots, n. \tag{5}
$$

Es decir, ahora tenemos un modelo de regresión lineal.

Podemos escribir la ecuación (5) como

$$\boldsymbol{y}_0 = X_0\boldsymbol{\beta}_0 + \boldsymbol{\varepsilon},$$

por lo que la estimación de $\boldsymbol{\beta}_0$ es

$$\hat{\boldsymbol{\beta}}_0 = (X_0^\top X_0)^{-1} X_0^\top \boldsymbol{y}_0 = (X_0^\top X_0)^{-1} X_0^\top (\boldsymbol{y} - \boldsymbol{f}_0) \tag{6}$$

Dado que $\boldsymbol{\beta}^0 = \boldsymbol{\theta} - \boldsymbol{\theta}_0$, definimos

$$\hat{\boldsymbol{\theta}}_1 = \hat{\boldsymbol{\beta}}_0 + \boldsymbol{\theta}_0$$

como las nuevas estimaciones de $\boldsymbol{\theta}$. A $\hat{\boldsymbol{\beta}}_0$ también se le conoce como **vector de incrementos**.

Podemos usar las nuevas estimaciones $\hat{\boldsymbol{\theta}}_1$ en la ecuación (4) para producir otro conjunto de estimaciones, digamos $\hat{\boldsymbol{\theta}}_2$, y así sucesivamente.
En general, en la k-ésima iteración tenemos:

$$\hat{\boldsymbol{\theta}}_{k+1} = \hat{\boldsymbol{\theta}}_k + \hat{\boldsymbol{\beta}}_k = \hat{\boldsymbol{\theta}}_k + (X_k^\top X_k)^{-1} X_k^\top (\boldsymbol{y} - \boldsymbol{f}_k), \tag{7}$$

donde

$$\begin{aligned} X_k &= [X_{ij}^k], \\ \boldsymbol{f}_k &= [f_1^k, f_2^k, \ldots, f_n^k]^\top, \\ \hat{\boldsymbol{\theta}}_k &= [\theta_1^k, \theta_2^k, \ldots, \theta_p^k]^\top. \end{aligned}$$

Este proceso iterativo continúa hasta alcanzar la convergencia, es decir, hasta que

$$\frac{\hat{\theta}_j^{k+1} - \hat{\theta}_j^k}{\hat{\theta}_j^k} < \delta, \quad j = 1, 2, \ldots, p,$$

donde δ es un número pequeño, por ejemplo 1.0×10^{-6}.
En cada iteración se debe evaluar la suma de cuadrados residual $S(\hat{\boldsymbol{\theta}}_k)$ para asegurar que se ha obtenido una reducción en su valor.

Podemos hacer los cálculos anteriores para el ejemplo de Puromicina como sigue: Tomando los valores iniciales $\theta_{10} = 205$ y $\theta_{20} = 0.08$, se tiene la suma de cuadrados residual $S(\boldsymbol{\theta}_0) = 3155$.

$$\frac{\partial f(x, \theta_1, \theta_2)}{\partial \theta_1} = \frac{x}{\theta_2 + x} \quad \text{y} \quad \frac{\partial f(x, \theta_1, \theta_2)}{\partial \theta_2} = \frac{-\theta_1 x}{(\theta_2 + x)^2}$$

Como la primera observación de x es $x_1 = 0.02$, tenemos:

$$X_{11}^0 = \left.\frac{x_1}{\theta_2 + x_1}\right|_{\theta_2 = 0.08} = \frac{0.02}{0.08 + 0.02} = 0.2,$$

y también,

$$X_{12}^0 = \left.\frac{-\theta_1 x_1}{(\theta_2 + x_1)^2}\right|_{\substack{\theta_1 = 205 \\ \theta_2 = 0.08}} = \frac{(-205)(0.02)}{(0.08 + 0.02)^2} = -410.$$

Las derivadas X_{ij}^0 se recopilan en la matriz $\boldsymbol{X}_0$ y el vector de incrementos se calcula a partir de la ecuación (6) como:

$$\hat{\boldsymbol{\beta}}_0 = \begin{bmatrix} 8.03 \\ -0.017 \end{bmatrix}.$$

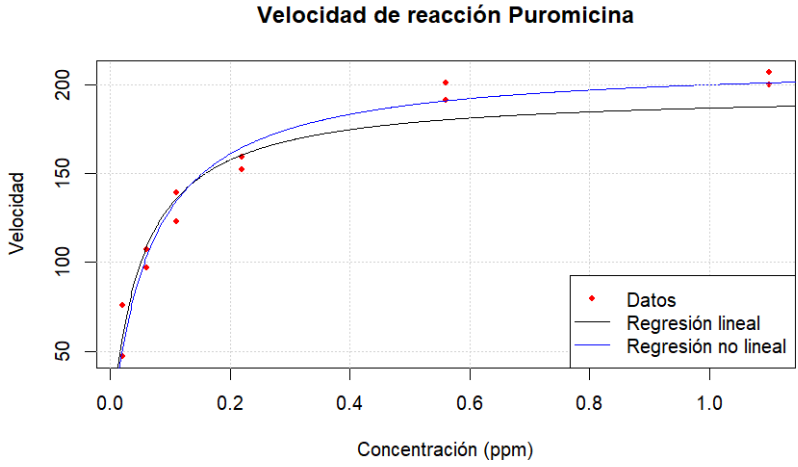
La nueva estimación de $\hat{\boldsymbol{\theta}}_1$ a partir de la ecuación (7) es:

$$\hat{\boldsymbol{\theta}}_1 = \hat{\boldsymbol{\beta}}_0 + \boldsymbol{\theta}_0 = \begin{bmatrix} 8.03 \\ -0.017 \end{bmatrix} + \begin{bmatrix} 205.00 \\ 0.08 \end{bmatrix} = \begin{bmatrix} 213.03 \\ 0.063 \end{bmatrix}.$$

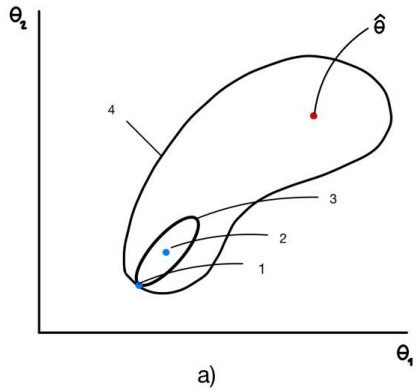
La suma de cuadrados residual en este punto es $S(\hat{\boldsymbol{\theta}}_1) = 1206$, que es considerablemente menor que $S(\boldsymbol{\theta}_0)$. Por lo tanto, $\hat{\boldsymbol{\theta}}_1$ se adopta como la nueva estimación de $\boldsymbol{\theta}$, y se realizaría otra iteración.
El algoritmo Gauss-Newton converge en $\hat{\boldsymbol{\theta}}^\top = [212.7,\ 0.0641]^\top$ con $S(\hat{\boldsymbol{\theta}}) = 1195$. Así, el modelo ajustado obtenido por linealización es:

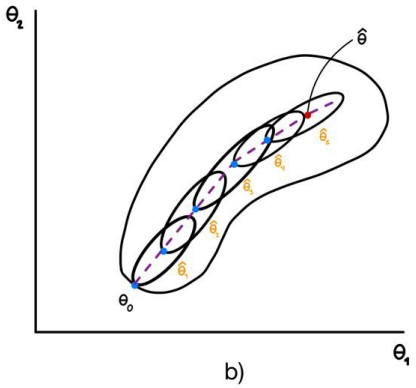
$$\hat{y} = \frac{\hat{\theta}_1 x}{x + \hat{\theta}_2} = \frac{212.7x}{x + 0.0641}.$$

Estos cálculos pueden realizarse automáticamente con la función nls() de R.


Velocidad de reacción Puromicina
Velocidad
50
100
150
200
0.0
0.2
0.4
0.6
0.8
1.0
Concentración (ppm)
Datos
Regresión lineal
Regresión no lineal

Perspectiva gráfica de linealización





- Valor inicial $\boldsymbol{\theta}_0$
- Primera iteración de la solución linealizada $\hat{\boldsymbol{\theta}}_1$
- Suma residual del contorno de secuencias en el problema linealizado
- $S(\boldsymbol{\theta})$ contorno pasando a través de $\boldsymbol{\theta}$.

Cuando el procedimiento de estimación converge a un vector final de estimaciones de parámetros $\hat{\theta}$, podemos obtener una estimación de la varianza del error σ^2 mediante el error cuadrático medio

$$\hat{\sigma}^2 = MS_{\text{Res}} = \frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{n-p} = \frac{\sum_{i=1}^{n}\left[y_i - f(\boldsymbol{x}_i, \hat{\boldsymbol{\theta}})\right]^2}{n-p} = \frac{S(\hat{\boldsymbol{\theta}})}{n-p},$$

También podemos estimar la **matriz de covarianzas asintótica** (para muestras grandes) del vector de parámetros $\hat{\boldsymbol{\theta}}$ como

$$\text{Var}(\hat{\boldsymbol{\theta}}) = \sigma^2\left(\hat{X}^\top \hat{X}\right)^{-1}, \tag{8}$$

en donde $\hat{X}$ es la matriz de derivadas parciales definida previamente, evaluada en la última iteración del estimador de mínimos cuadrados $\hat{\boldsymbol{\theta}}$.

Este método se basa en una aproximación cuadrática a la función objetivo, en este caso $S(\boldsymbol{\theta})$. Supongamos que $\boldsymbol{\theta}_0$ es un valor inicial cercano al valor óptimo. Entonces, podemos aproximar $S(\boldsymbol{\theta})$ mediante una expansión de Taylor de segundo orden alrededor de $\boldsymbol{\theta}_0$:

$$S(\boldsymbol{\theta}) \approx S(\boldsymbol{\theta}_0) + (\boldsymbol{\theta} - \boldsymbol{\theta}_0)^\top \nabla S(\boldsymbol{\theta}_0) + \frac{1}{2}(\boldsymbol{\theta} - \boldsymbol{\theta}_0)^\top H_S(\boldsymbol{\theta}_0)(\boldsymbol{\theta} - \boldsymbol{\theta}_0),$$

Gráficas de convergencia de métodos iterativos **SAMUEL** (comparar descenso máximo, incrementos fraccionarios, Marquardt)

Sección 3: **Inferencia estadística en regresión no lineal**

Intro al tema

Validez de la inferencia aproximada

Sección 4: **Validación del modelo**

Sección 5: **Conclusiones**

Ventajas y desventajas, incluyendo no identificabilidad

Referencias