Curvas ROC y Regresión Lineal

Julio Deride Silva

Área de Matemática
Facultad de Ciencias Químicas y Farmcéuticas
Universidad de Chile

4 de junio de 2010

Curvas ROC y Regresión Lineal

Julio Deride Silva

Área de Matemática
Facultad de Ciencias Químicas y Farmcéuticas
Universidad de Chile

4 de junio de 2010

Tabla de Contenidos

Tabla de Contenidos

Outline

Ejemplo.

Se quiere diagnosticar si un paciente presenta cierta enfermedad o no. Para ello, se controla alguna variable relacionada con la enfermedad y se desea determinar si el paciente está enfermo a través del nivel de dicha variable. Es así como, si queremos saber si un paciente presenta un colesterol alto (enfermo) o no (sano), se mide la variable X correspondiente al % de ácidos grasos en la sangre. Finalmente, el test se construye de la siguiente forma:

- Si $X > c$, entonces el paciente presenta colesterol alto;
- si $X \leq c$, entonces el paciente presenta colesterol bajo.

Para determinar el valor óptimo de c emplearemos criterios extras, los cuales se discutirán más adelante.

Notación

De manera general, denotaremos T^+ y T^- al resultado del test positivo y negativo, respectivamente. A su vez, denotaremos R^+ y R^- a los casos en que el individuo presenta y no presenta la condición en estudio, respectivamente. Luego, la clasificación se puede resumir en el siguiente cuadro

	Estado real	
Resultado Test	Cond.Ausente(R^-)	Cond.Presente(R^+)
Cond. Encontrada (T^+)	Falso Positivo FP	No hay error
Cond. No Encontrada (T^-)	No hay error	Falso Negativo FN

Observaciones.

- El test ideal es aquel que minimice los errores.
- Dada la clasificación, podemos definir dos tipos de errores: *falsos positivos*, individuos cuyo test arrojó un resultado positivo, sin tener presenta la condición; y *falsos negativos*, correspondientes a aquellos individuos cuyo test arrojó un resultado negativo, a pesar de presentar la condición.

Outline

Definiciones.

Sea X una variable aleatoria sobre la cual queremos determinar el test.

Definición (Sensibilidad)

Es la probabilidad de clasificar correctamente a un individuo cuyo estado real es definido como positivo, respecto a la condición de prueba. Esto es

$$\textit{Sensibilidad} = \mathbb{P}[T^+|R^+].$$

Definiciones.

Definición (Especificidad)

Es la probabilidad de clasificar correctamente a un individuo cuyo estado real es definido como negativo, respecto a la condición de prueba. Esto es

$$\textit{Especificidad} = \mathbb{P}[T^- | R^-].$$

Estimaciones

Dada una muestra, podemos estimar las probabilidades anteriores de la siguiente forma:

$$\begin{aligned} \text{Sensibilidad} &= \frac{\text{número de verdaderos positivos}}{\text{número de positivos reales}} \\ &= FVP \\ \text{Especificidad} &= \frac{\text{número de verdaderos negativos}}{\text{número de negativos reales}} \\ &= FVN, \end{aligned}$$

(FVP: fracción de verdaderos positivos y FVN: fracción de verdaderos negativos).

Outline

Ejemplo.

Consideremos un test cuyos resultados de clasificación se resumen en la siguiente tabla

	R^+	R^-	Total
T^+	12	9	21
T^-	4	27	31
Total	16	36	52

Ejemplo.

Dados los resultados de este test en esta muestra, podemos hacer las siguientes estimaciones

$$\begin{aligned} \text{Sensibilidad} &= \mathbb{P}[T^+|R^+] \\ &\approx \text{FVP} \\ &= \frac{VP}{VP+FN} \\ &= \frac{12}{16} \\ &= \frac{3}{4} \end{aligned}$$

Ejemplo

Por otra parte,

$$\begin{aligned} \text{Especificidad} &= \mathbb{P}[T^- | R^-] \\ &\approx \text{FVN} \\ &= \frac{VN}{VN + FP} \\ &= \frac{27}{36} \\ &= \frac{3}{4} \end{aligned}$$

Outline

Índices Predictivos.

Haciendo uso valores muestrales y el teorema de Bayes, se construyen los siguientes indicadores predictivos:

Definición (Índice Predictivo de Verdaderos Positivos)

$$\mathbb{P}[R^+|T^+] = \frac{\mathbb{P}[T^+|R^+] \cdot \mathbb{P}[R^+]}{\mathbb{P}[T^+|R^+] \cdot \mathbb{P}[R^+] + \mathbb{P}[T^+|R^-] \cdot \mathbb{P}[R^-]}$$

Definición (Índice Predictivo de Verdaderos Negativos)

$$\mathbb{P}[R^-|T^-] = \frac{\mathbb{P}[T^-|R^-] \cdot \mathbb{P}[R^-]}{\mathbb{P}[T^-|R^-] \cdot \mathbb{P}[R^-] + \mathbb{P}[T^-|R^+] \cdot \mathbb{P}[R^+]}$$

Outline

Diseño del Test.

Supongamos que se mide una variable aleatoria X que es continua. Luego, el test se construirá de la siguiente forma:

- Si $X > c$, entonces el paciente es diagnosticado positivo (T^+);
- si $X \leq c$, entonces el paciente es diagnosticado negativo (T^-).

Además, supondremos que la función de densidad de probabilidad de la variable X dependerá si presenta la condición en estudio o no. Esto es:

- Si $X \sim f_+$ si la persona presenta la condición (R^+).
- si $X \sim f_-$ si la persona no presenta la condición (R^-).

Sensibilidad y Especificidad.

Dado un valor de c fijo, se puede calcular directamente los valores de sensibilidad y especificidad del test de la siguiente forma:

$$\begin{aligned}
\text{Sensibilidad} &= \mathbb{P}[T^+|R^+] \\
&= \mathbb{P}[X > c|X \sim f_+] \\
&= \int_c^\infty f_+(t)\text{dt}. \\
\text{Especificidad} &= \mathbb{P}[T^-|R^-] \\
&= \mathbb{P}[X \leq c|X \sim f_-] \\
&= \int_{-\infty}^c f_-(t)\text{dt}. \\
1 - \text{Especificidad} &= \int_c^\infty f_-(t)\text{dt}.
\end{aligned}$$

Curva ROC

Para cada valor de c se encuentra un par $(Sensibilidad(c), 1 - Especificidad(c))$, los cuales definen una curva. Esta curva es la llamada **Curva ROC**. Para estimaciones muestrales, la curva ROC se construirá variando el nivel de corte de la variable en estudio y estimando con los pares (FVP, FFP). Esta se ilustra en la figura (1).

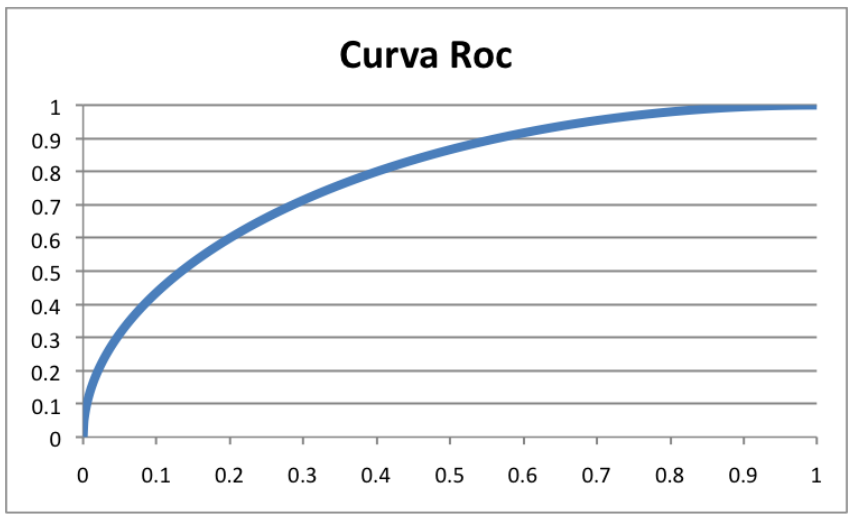


Figura: Curva ROC

Observaciones.

Consideremos lo siguiente:

- Los puntos $(0,0)$ y $(1,1)$ siempre pertenecen a la curva ROC.
- La forma de la curva dependerá de cuán diferentes sean las densidades. Es así como, en el caso extremo, $f_+ = f_-$, se concluye que, independiente el valor de corte c, $FVP \approx FFP$. El test es malo y no permite discriminar de buena manera.

Observación.

Si las densidades se presentan como la figura (2), la curva ROC será como en (3), donde se nota claramente que el test es mejor.

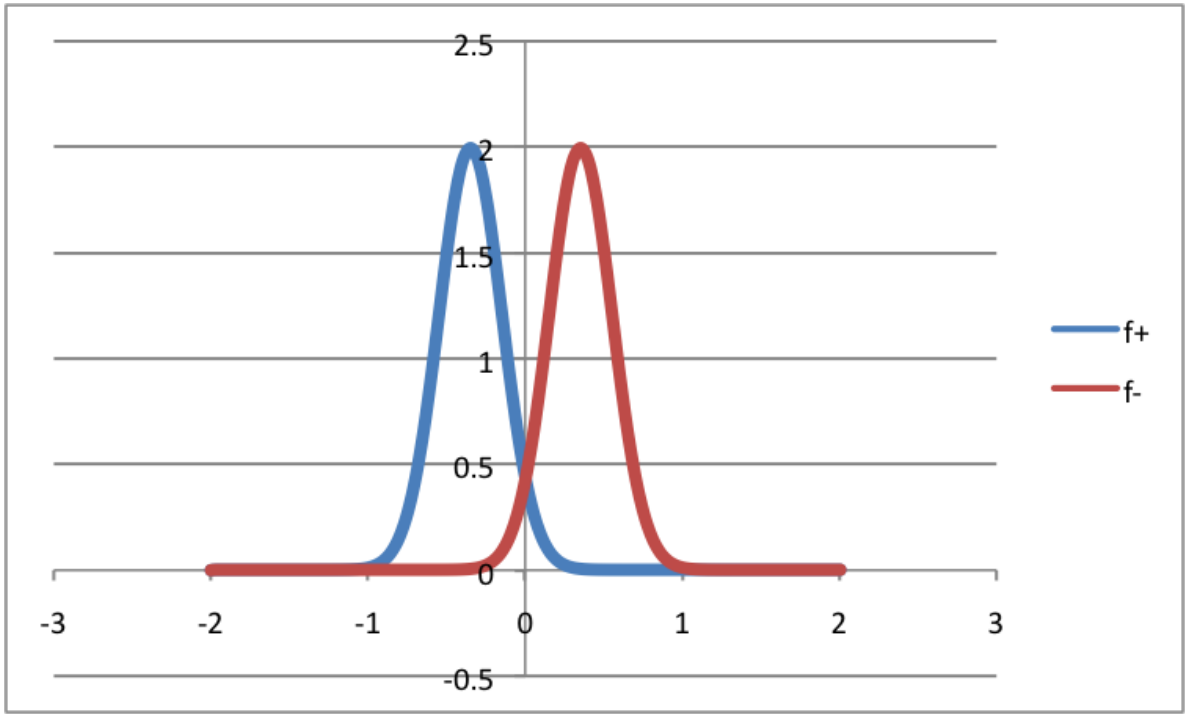


Figura: Densidades muy diferentes

Curva ROC

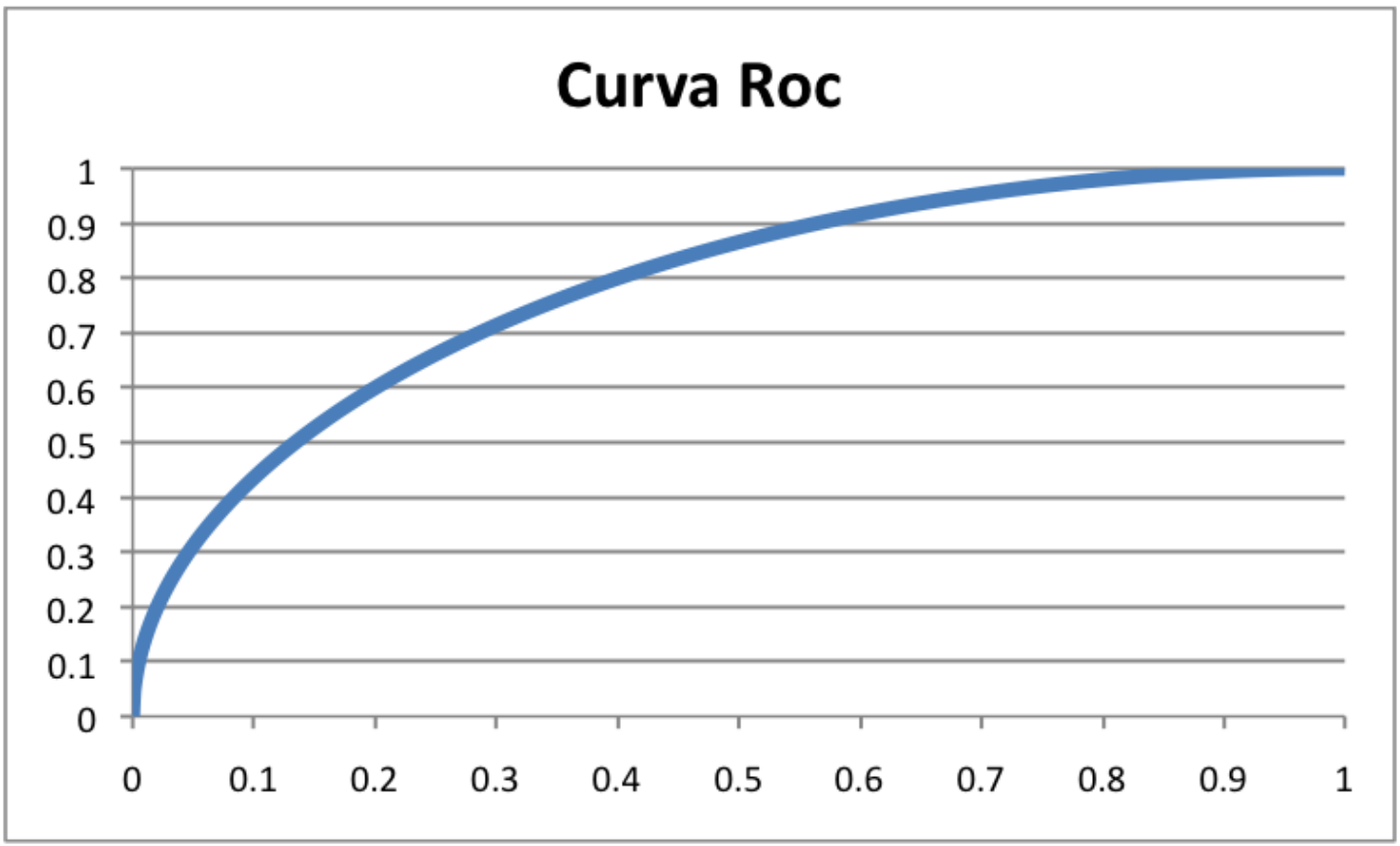


Figura: Curva ROC

Defínción.

Definición (Índice de separación de poblaciones)

$$\mathcal{I}_{sep} = 2\left(\textit{Área bajo la curva ROC} - \frac{1}{2}\right)$$

Se tiene

1. $\mathcal{I}_{sep} \in [0, 1]$.
2. $\mathcal{I}_{sep} = 0$, entonces las poblaciones son indistinguibles.
3. $\mathcal{I}_{sep} = 1$, la población que presenta la condición y la población que no la presenta se pueden distinguir perfectamente.

Outline

Ejemplo

Considere un test de diagnóstico que puede tomar tres valores: Alto, Medio y Bajo. Los pacientes pueden ser normales (E^-) o enfermos (E^+). Se realizó un experimento sobre 100 pacientes y los resultados se resumen en el cuadro (1):

	Estado Real	
Valor del Test	Normal(E^-)	Enfermo(E^+)
Alto	30	8
Medio	14	12
Bajo	6	30

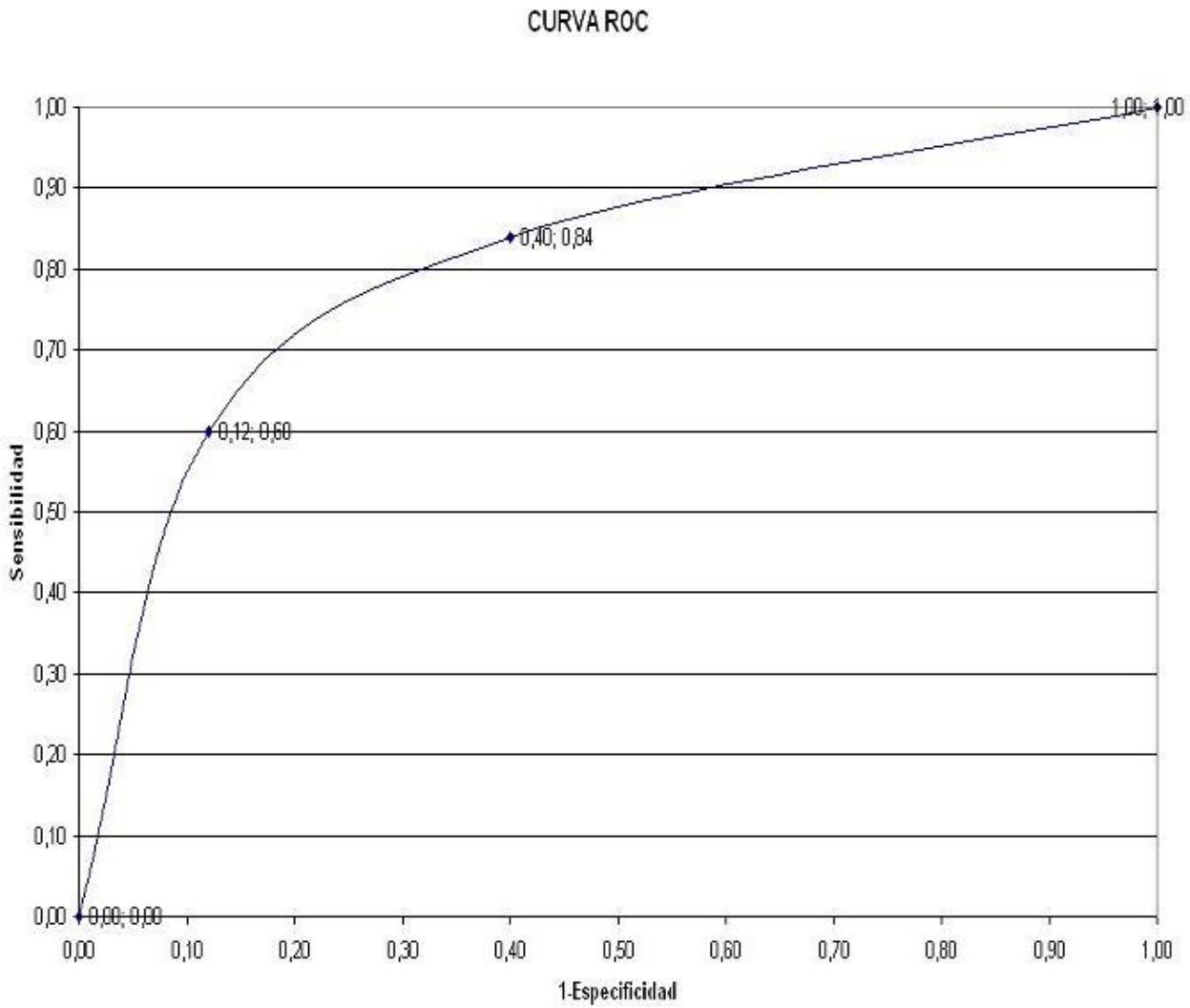
Cuadro: Valores Muestrales.

Sensibilidad y Especificidad

Valor del Test	**Sens. (FVP)**	**(1-Especificidad) (F**
Siempre Normal (T^-)	0	0
Alto y Medio (T^-) - Bajo (T^+)	0,60	0,12
Alto (T^-) - Medio y Bajo (T^+)	0,84	0,40
Nunca Normal (T^+)	1	1

Cuadro: Sensibilidad y (1-Especificidad)

Curva ROC


CURVA ROC
Sensibilidad
1-Especificidad
1,00
0,90
0,80
0,70
0,60
0,50
0,40
0,30
0,20
0,10
0,00
0,00
0,10
0,20
0,30
0,40
0,50
0,60
0,70
0,80
0,90
1,00
1,00; 1,00
0,40; 0,84
0,12; 0,60
0,00; 0,00

Outline

Introducción y Motivación.

Se espera establecer una relación funcional entre dos variables. A partir de los valores muestrales se construye una relación del tipo lineal que permita explicar el comportamiento de una variable en función de otra.
La pregunta en este caso es determinar si existe una relación entre dos variables y para responderla se plantean modelos e hipótesis y se estudia la validez de estas.
La relación más simple que podemos suponer entre dos variables es una relación del tipo lineal, para la cual el modelo ofrece la ventaja de ser simple y fácil de usar.

Ejemplos.

1. Un automovilista preocupado por la alza en los combustibles, desea estudiar la relación entre el gasto de combustible (y) y la velocidad media de viaje (x).
2. Un estudiante de química desea comprobar la primera ley de la termodinámica. Para ello, realiza experimentos sobre un gas en un ambiente isobárico y toma mediciones de su temperatura y volumen. Desea establecer si existe una relación lineal entre ambas variables.

Outline

Modelo.

Se plantea un modelo de la forma

$$Y = a + b \cdot x + \varepsilon, \quad \varepsilon \sim \mathcal{N}(0, \sigma^2),$$

donde x será una variable exógena, independiente y que denominaremos *explicativa*, Y será una variable endógena dependiente, y ε será una variable aleatoria distribuida de forma normal, con esperanza cero y varianza σ^2 (puede ser desconocida).
Este modelo dice que las realizaciones de la variable Y dependen linealmente del resultado que toma la variable x, más un error aleatorio. Por lo tanto, x contiene información para *estimar* el resultado de Y.

Procedimiento

1. Se dispone de una muestra de tamaño n con observaciones pareadas de ambas variables, es decir, $\{(x_i, y_i)\}_{i=1}^{n}$,
2. se busca una recta que mejor se ajuste a los datos. El criterio para determinar la mejor recta será considerar el mínimo error total que se comente al aproximar por una recta. Para ello, consideraremos que cada observación entrega un error ε_i, y el error total será la suma de cada error al cuadrado, para evitar problemas de signo.

Con esto, el problema consiste en encontrar el coeficiente de posición a y la pendiente b de la recta, cuyo error total sea mínimo.

Outline

Estimación

Encontrar $\hat{a}$, $\hat{b}$ coeficientes de la recta que son solución del problema

$$\min_{\{a,b\}} \sum_{i=1}^{n} \varepsilon_i^2 \Leftrightarrow \min_{\{a,b\}} \sum_{i=1}^{n} (y_i - a - bx_i)^2.$$

Si definimos $F(a,b) = \sum_{i=1}^{n} (y_i - a - bx_i)^2$, se tienen las siguientes condiciones de primer orden:

Condiciones de primer orden

$$\frac{\partial F}{\partial a}(\hat{a},\hat{b}) = 0 \quad \Leftrightarrow \quad \sum_{i=1}^{n}(y_i - \hat{a} - \hat{b}x_i) = 0$$

$$\frac{\partial F}{\partial b}(\hat{a},\hat{b}) = 0 \quad \Leftrightarrow \quad \sum_{i=1}^{n} x_i(y_i - \hat{a} - \hat{b}x_i) = 0$$

De acá, resolviendo el sistema, se obtiene que los coeficientes vienen dados por

Coeficientes.

$$
\begin{aligned}
\hat{b} &= \frac{\left(\frac{1}{n}\sum_{i=1}^{n} x_i y_i\right) - \overline{xy}}{\left(\frac{1}{n}\sum_{i=1}^{n} x_i^2\right) - \overline{x}^2}. \\
&= \frac{cov^{\mathrm{m}}(x,y)}{var^{\mathrm{m}}(x)}. \\
\hat{a} &= \overline{y} - \hat{b}\overline{x}
\end{aligned}
$$

Observación.

Los coeficientes $\hat{a}$ y $\hat{b}$ dependen de los valores muestrales. Con esto, para diferentes muestras, obtendremos **diferentes estimaciones de rectas** a través de los mínimos cuadrados.

Estimación.

Para coeficientes estimados con mínimos cuadrados, definimos $\hat{y}_i = \hat{a} + \hat{b}x_i$, la cual llamaremos *estimación de* y_i, $e_i = y_i - \hat{y}_i$, el cual llamaremos residuo. Finalmente, si se dispone de una nueva observación para la variable x, x_{i+1}, llamaremos predicción del modelo a $y_{i+1} = \hat{a} + \hat{b}x_{i+1}$.

Outline

Bondad del modelo.

Para determinar cuán bueno es el ajuste lineal en una muestra, definimos

Definición (Coeficiente de correlación lineal.)

$$R = \frac{cov^m(x,y)}{\sqrt{var^m(x)}\sqrt{var^m(y)}}$$

Definición (Coeficiente de determinación.)

$$R^2 = \left(\frac{cov^m(x,y)}{\sqrt{var^m(x)}\sqrt{var^m(y)}}\right)^2$$

Observaciones

Para establecer la bondad del modelo, notemos que:

- $R \in [-1, 1]$.
- $R^2 \in [0, 1]$.
- Cuando $R^2 \approx 1$, se tiene que existe una correlación lineal fuerte entre las variables.
- Cuando $R^2 \approx 0$, se tiene que no existe una correlación **lineal** entre las variables.
- El coeficiente de determinación se interpreta como el porcentaje de variabilidad de la variable Y explicada por el modelo lineal.

Outline

Aplicaciones.

- **Modelo exponencial.**

$$Y = ae^{a+bx}$$

- **Modelo de tiempos de reacción.**

$$Y = Ae^{-ct}$$

- **Modelos en economía.**

$$Y = AK^{\alpha}L^{1-\alpha}.$$

Outline

Ejemplo.

Se tienen los siguientes datos respecto de la concentración de una solución en función del tiempo, agrupados en el cuadro 3

t	0,50	1,00	1,50	2,00	2,50	3,00	3,50	4,00
C	3,03	2,31	1,30	1,67	1,43	0,63	0,71	0,84
t	4,50	5,00	5,50	6,00	6,50	7,00	7,50	8,00
C	0,38	0,25	0,24	0,27	0,18	0,10	0,12	0,07

Cuadro: Datos.

Diagrama de dispersión.

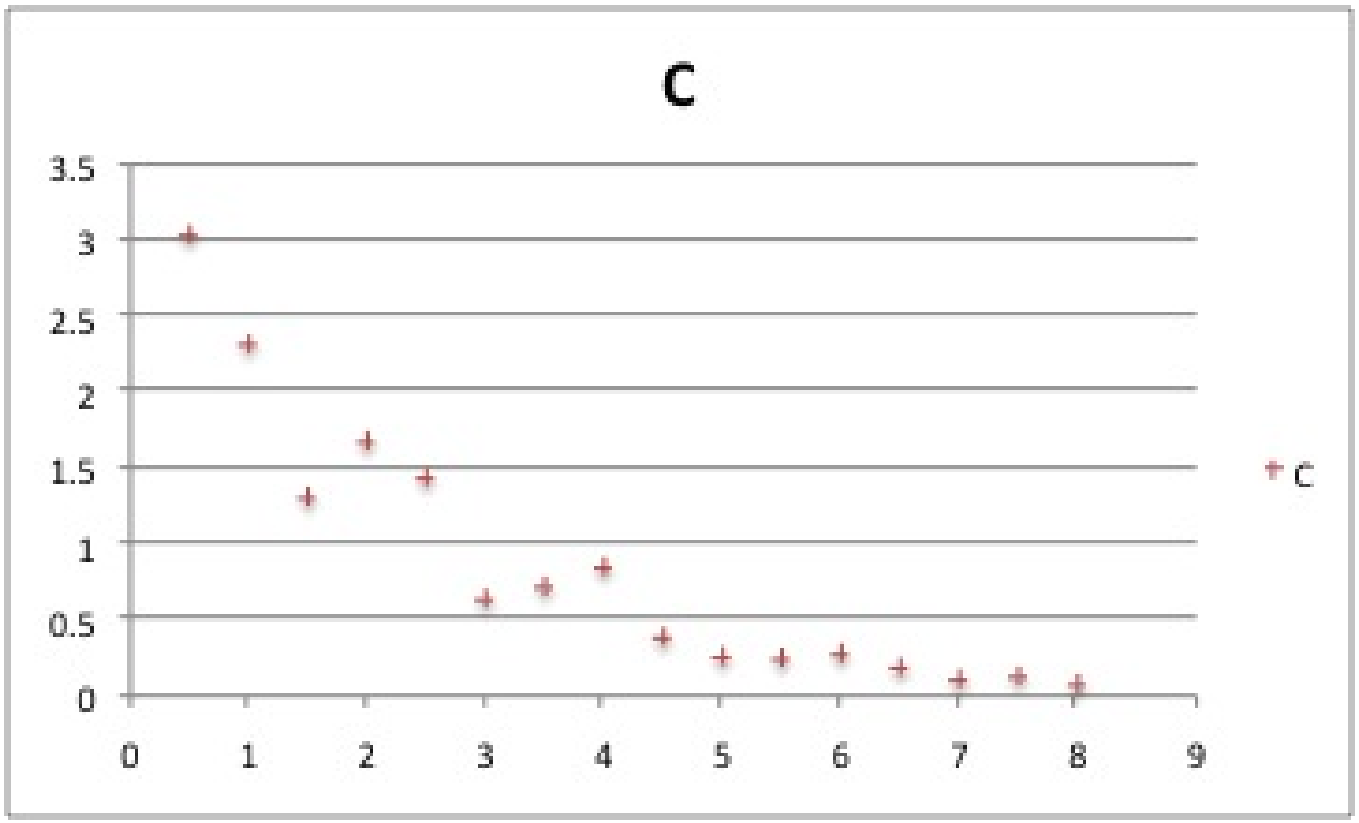


Figura: Diagrama de dispersión.

Ejemplo.

1. Es razonable suponer que existe una relación lineal entre X e Y que permita predecir Y en función de X, ya que se puede apreciar en el diagrama 6 que ambas variables siguen un crecimiento inverso que podría ser lineal.
2. Estimando los coeficientes del modelo $y = n + mx$.

$$\begin{aligned} \hat{m} &= -0,33 \\ \hat{n} &= 2,24 \end{aligned}$$

3. Se calcula el coeficiente de determinación:

$$R^2 = 0,79,$$

el cual es un valor alto, lo que sugiere que el modelo de regresión lineal es apropiado y explica un 79 % de la variación de C.

Ejemplo.

Se pretende buscar otro tipo de relación, en base al diagrama de los datos. Para ello, se plantea el modelo

$$y = ae^{bx}. \tag{1}$$

Aplicando el logaritmo, el modelo (1) se transforma en uno lineal.

$$\begin{aligned} y &= ae^{bx} \\ \ln(y) &= \ln(a) + bx \end{aligned}$$

Estimando los coeficientes y calculando el coeficiente de determinación R^2.

$$\begin{aligned} \hat{b} &= -0,48 \\ \ln(\hat{a}) &= 1,31 \\ R^2 &= 0,96. \end{aligned}$$

Diagrama de dispersión.

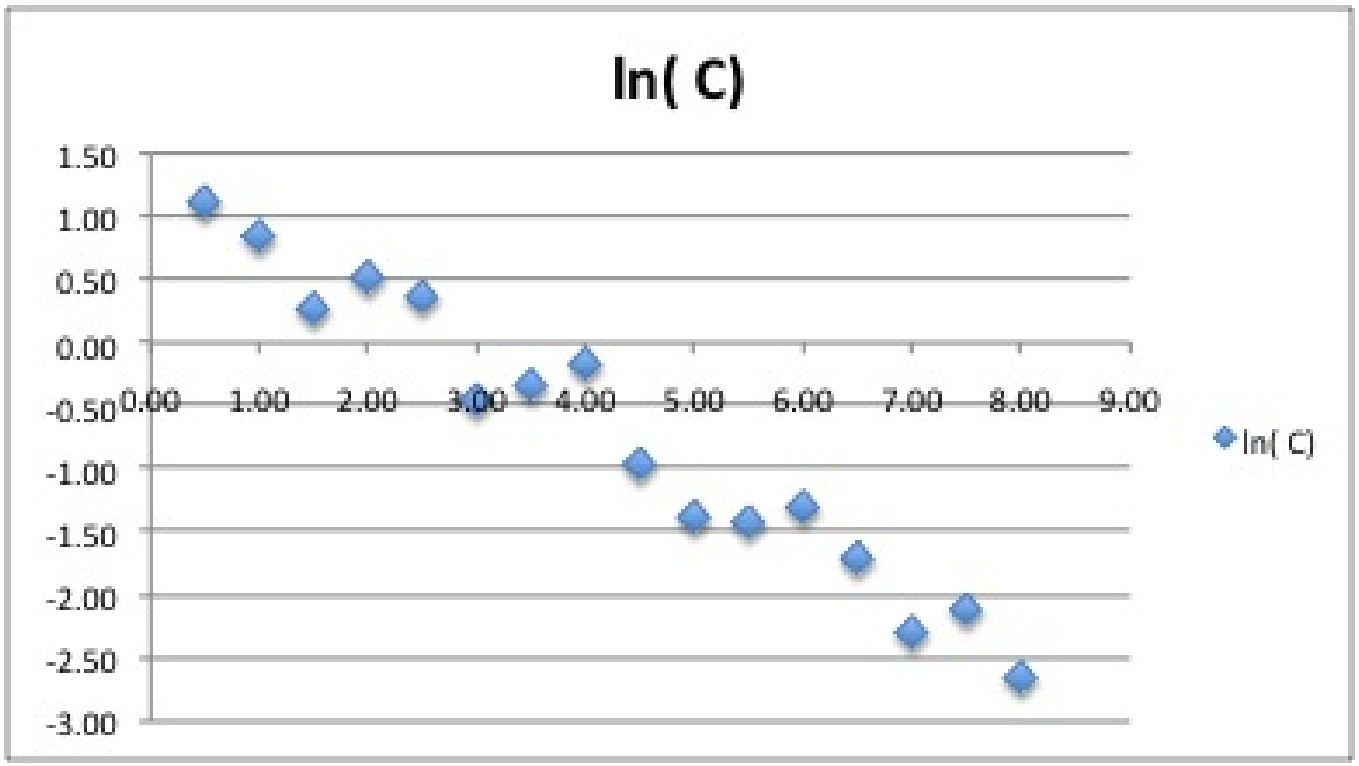


Figura: Diagrama de dispersión.

Conclusiones

En este caso, se tiene que la relación es fuertemente lineal, con un coeficiente de determinación cercano a 1. Con este modelo se explica el 96 % de la variación de C.
Finalmente, en ambos casos se obtienen buenos resultados de regresión. Sin embargo, el segundo modelo posee un coeficiente de determinación más cercano a 1, y por lo tanto, la relación que explica la variabilidad de C de la mejor fora es $C = ae^{bx}$.