Notas de Clase Regresión Lineal

Capítulo 2: Regresión Lineal Simple

Nelfi González Alvarez
Profesora Asociada

Isabel Cristina Ramírez Guevara
Profesora Asociada

Escuela de Estadística

Universidad Nacional de Colombia, Sede Medellín

UNIVERSIDAD NACIONAL DE COLOMBIA

2021

Capítulo 2

Regresión lineal simple

2.1. Introducción

En muchas ocasiones es posible diseñar experimentos estadísticos controlados, en los cuáles es factible el estudio simultáneo de varios factores, aplicando procedimientos de aleatorización apropiados en lo que se conoce como análisis de varianza. Sin embargo en muchas ocasiones sólo se cuenta con un conjunto de datos sobre los cuáles es difícil esperar que hayan sido observados en condiciones estrictamente controladas, y de los cuáles también en pocas ocasiones se tienen réplicas para calcular el error experimental.
Cuando se enfrenta la situación anterior lo más apropiado es aplicar los métodos de mínimos cuadrados. Debe tenerse presente que los métodos de regresión permiten establecer asociaciones entre variables de interés entre las cuáles la relación usual no es necesariamente de causa - efecto.

2.2. Fundamentos

2.2.1. Nomenclatura

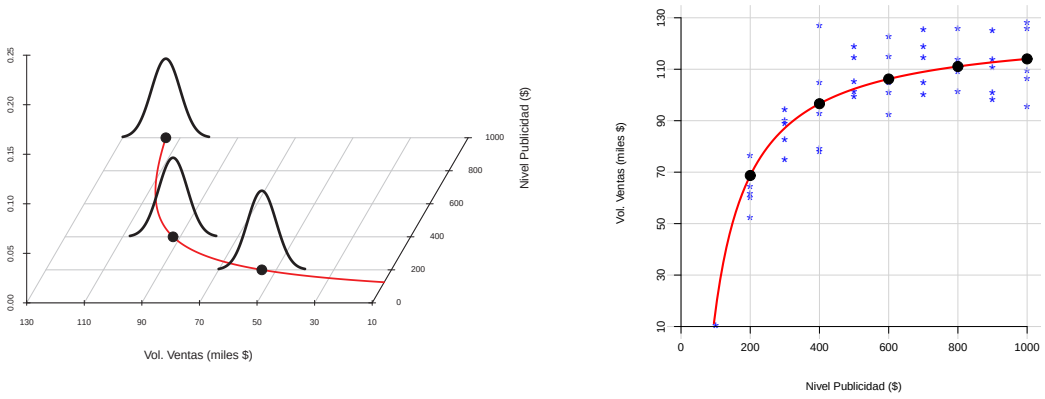
- Y: Variable respuesta o dependiente.
- X: Variable predictora, independiente o regresora.
- E: Error aleatorio
- β_0, β_1: Parámetros de la regresión. β_0 es el intercepto y β_1 la pendiente de la línea recta.
- $\widehat{\beta}_0$: Estimador del parámetro β_0
- $\widehat{\beta}_1$: Estimador del parámetro β_1.
- $\widehat{E}$: Residual, es una estimación del error aleatorio.
- $\widehat{Y}$: Es la estimación de $\mathrm{E}[Y|X]$

2.2.2. Significados de la regresión

La regresión tiene dos significados:

- Primero, podemos verla a partir de la distribución conjunta de las variables X e Y, en la cual podemos definir la distribución condicional de $Y|X$, $f(Y|X)$ y determinar $\mathrm{E}[Y|X]$. En este caso la regresión pretende ajustar la curva correspondiente a $\mathrm{E}[Y|X]$. Ver Figura 2.1(a).
- Segundo, dado un conjunto de pares de datos (x_i, y_i), $i = 1, 2, \ldots, n$, puede asumirse una forma funcional para la curva de regresión y tratar de ajustarla a los datos. Ver Figura 2.1(b).

El segundo caso es el que más se da en la práctica.



(a) (b)

Figura 2.1: Función de regresión: (a) Curva que pasa por las medias de Y dado cada valor de X, es decir, $E[Y|X] = f(X)$ (b) Curva con respecto a la cual es mínima la distancia vertical de las observaciones (x_i, y_i) $i = 1, 2, \ldots, n$.

2.2.3. Consideraciones

- La variable respuesta es una variable aleatoria cuyos valores se observan mediante la selección de los valores de la variable predictora en un intervalo de interés.
- Por lo anterior, las variables predictoras no son consideradas como variables aleatorias, sino como un conjunto de valores fijos que representan los puntos de observación, que se seleccionan con anticipación y se miden sin error. Sin embargo si esto último no se cumple, el método de mínimos cuadrados ordinarios puede seguir siendo válido si los errores en los valores de X son pequeños en comparación con los errores aleatorios.
- Los datos que se observan constituyen una muestra representativa de un medio acerca del cual se desea generalizar. Si no es así, no es apropiado realizar inferencias (extrapolaciones) en un rango de los datos por fuera del considerado.

- El modelo de regresión es lineal en los parámetros. Es decir, ningún parámetro de la regresión aparece como el exponente o es dividido o multiplicado por otro parámetro. Sin embargo, la línea de ajuste puede tener una curvatura (no ser lineal en X y/o en Y), caso en el cual mediante una transformación conveniente de las variables (X y/o Y), es posible aplicar las técnicas de regresión lineal sobre estas nuevas variables.

- Si la ecuación de regresión seleccionada es correcta, cualquier variabilidad en la variable respuesta que no puede ser explicada exactamente por dicha ecuación, es debida a un error aleatorio.

- Los valores observados de la variable respuesta no se encuentran estadísticamente correlacionados. Se supone que cada valor observado de Y está constituído por un valor real y una componente aleatoria.

- El modelo de regresión con n pares de datos es:

$$Y|X_i = \beta_0 + \beta_1 X_i + E_i, i = 1, 2, \ldots, n \tag{2.1}$$

$$\mathrm{E}\left[Y|X_i\right] = \beta_0 + \beta_1 X_i \tag{2.2}$$

  Nota 2.1. Por simplicidad la ecuación (2.1) la escribimos como,

$$Y_i = \beta_0 + \beta_1 X_i + E_i, i = 1, 2, \ldots, n \tag{2.3}$$

- Los errores aleatorios $E_i \sim N(0, \sigma^2), i = 1, 2, ..., n$

- Los errores aleatorios E_i son estadísticamente independientes. Por tanto:

$$\mathrm{Cov}\left(E_i, E_j\right) = 0, \forall_{i,j}, i \neq j \tag{2.4}$$

$$\mathrm{Cov}\left(Y_i, Y_j\right) = 0, \forall_{i,j}, i \neq j \tag{2.5}$$

- La varianza de los errores aleatorios es $\sigma^2, \forall i, i = 1, 2, ..., n$ pero desconocida. Dado que se asume que las variables predictoras no son variables aleatorias, la varianza de los Y_i también es $\sigma^2, \forall i$ y por tanto es independiente del punto de observación, es decir, del valor de X. Pero en el caso que esta última suposición no pueda aplicarse, entonces el método de regresión empleado será el de mínimos cuadrados ponderados. Con estas consideraciones y las anteriores, podemos afirmar que:

$$Y|X_i \sim N(\beta_0 + \beta_1 X_i, \sigma^2) \tag{2.6}$$

  Ver ejemplo en Figura 2.2, en donde se ilustra la recta de regresión como la función de la media condicional de la respuesta dado la variable predictora x.

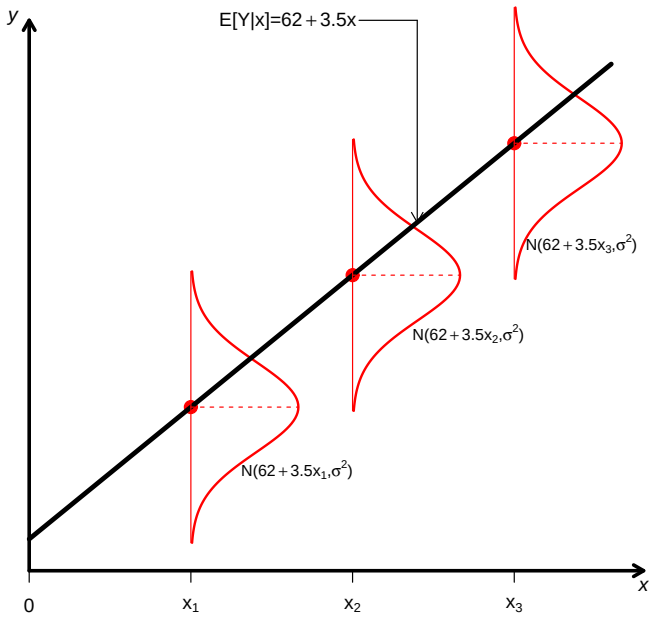


Figura 2.2: Recta en el modelo de regresión lineal simple como la función de la media condicional de $Y|x$. En este ejemplo se ha asumido que $Y|x \sim N\left(62+3.5x, \sigma^2\right)$, luego, la recta de regresión corresponde a $\mathrm{E}\left[Y|x\right] = 62+3.5x$ y en cada nivel de x se tiene la misma varianza para Y alrededor de la respectiva media condicional.

Nota 2.2. La interpretación de los coeficientes de regresión: β_1 representa el cambio en la media de Y dado un cambio unitario en X. Si el rango en que se observa X incluye a $x = 0$, entonces β_0 corresponde a la media de la distribución de Y cuando $x = 0$. Sin embargo, si $x = 0$ no ha sido observado en los datos, entonces β_0 no tiene interpretación práctica en el modelo de regresión.

2.3. Estimación por mínimos cuadrados ordinarios (MCO)

Para una selección preliminar de la variable predictora en un modelo de regresión simple (o sea que considera una sola variable predictora) es conveniente realizar el diagrama de dispersión y_i vs. x_i y mirar si existe una tendencia lineal en la nube de puntos. Si la nube de puntos parece mejor ajustada por una función no lineal, hay que buscar una transformación apropiada en X y/o Y que linealice; en este caso el modelo de regresión lineal a ajustar será: $Y^*|X_i^* = \beta_0 + \beta_1 X_i^* + E_i, i = 1, 2, ..., n$ donde Y^* y X^* son las variables Y y X transformadas.

Nota 2.3. Debe tenerse claro que el método de mínimos cuadrados es un método numérico, no estadístico; La estadística opera a partir de los supuestos distribucionales asignados en el modelo de regresión.

2.3.1. Objetivo

Obtener estimaciones de los parámetros de regresión, es decir hallar $\widehat{\beta}_0$ y $\widehat{\beta}_1$, tales que minimicen la suma de los cuadrados de los errores $S(\beta_0, \beta_1)$:

$$S(\beta_0, \beta_1) = \sum_{i=1}^{n} E_i^2 = \sum_{i=1}^{n} [Y_i - (\beta_0 + \beta_1 X_i)]^2 \tag{2.7}$$

2.3.2. Valor de los estimadores

Dados los pares de observaciones $(x_1, y_1), ..., (x_n, y_n)$, hallar β_0 y β_1 que minimicen a $S(\beta_0, \beta_1)$. Sean $\widehat{\beta}_0$ y $\widehat{\beta}_1$ los respectivos valores con los cuales se obtiene tal valor mínimo. Este problema de optimización implica resolver el siguiente sistema de ecuaciones:

$$\frac{\partial S(\beta_0, \beta_1)}{\partial \beta_0} = 0, \tag{2.8}$$

$$\frac{\partial S(\beta_0, \beta_1)}{\partial \beta_1} = 0, \tag{2.9}$$

de lo cual surgen las denominadas ecuaciones normales:

$$\sum_{i=1}^{n} y_i = n\beta_0 + \beta_1 \sum_{i=1}^{n} x_i, \tag{2.10}$$

$$\sum_{i=1}^{n} x_i y_i = \beta_0 \sum_{i=1}^{n} x_i + \beta_1 \sum_{i=1}^{n} x_i^2, \tag{2.11}$$

y de éstas tenemos que las estimaciones por mínimos cuadrados de los parámetros son:

$$\widehat{\beta}_0 = \bar{y} - \widehat{\beta}_1 \bar{x}, \tag{2.12}$$

$$\widehat{\beta}_1 = \frac{n \sum_{i=1}^{n} x_i y_i - \left(\sum_{i=1}^{n} x_i \right) \left(\sum_{i=1}^{n} y_i \right)}{n \sum_{i=1}^{n} x_i^2 - \left(\sum_{i=1}^{n} x_i \right)^2}, \tag{2.13}$$

o bien:

$$\widehat{\beta}_1 = \frac{\sum_{i=1}^{n} (x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^{n} (x_i - \bar{x})^2}, \tag{2.14}$$

o bien:

$$\widehat{\beta}_1 = \frac{\sum_{i=1}^{n} (x_i - \bar{x}) y_i}{\sum_{i=1}^{n} (x_i - \bar{x})^2}. \tag{2.15}$$

Por tanto, una estimación de la respuesta media (o respuesta ajustada), en $X = x_i$, es:

$$\hat{y}_i = \widehat{\beta}_0 + \widehat{\beta}_1 x_i \tag{2.16}$$

o bien:

$$\hat{y}_i = \bar{y} + (x_i - \bar{x})\widehat{\beta}_1 \tag{2.17}$$

2.3.3. Un ejemplo sencillo

Considere los $n = 11$ pares de datos (x_i, y_i), presentados en la siguiente tabla, obtenidos por simulación del modelo $Y_i = 62 + 3.5 * x_i + E_i$, $E_i \overset{\text{iid}}{\sim} N(0, \sigma^2)$, con $\sigma = 50$. Los cálculos para la estimación por MCO también son ilustrados.

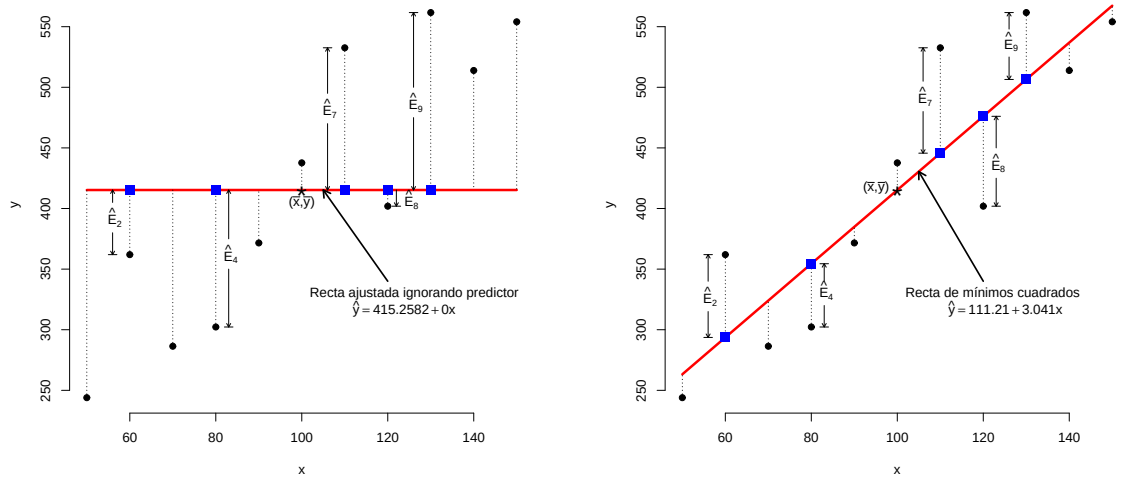
Tabla 2.1: Datos simulados y cálculos para ajuste del modelo de regresión lineal simple por mínimos cuadrados ordinarios.

i	x_i	y_i	$x_i - \bar{x}$	$(x_i - \bar{x})^2$	$y_i - \bar{y}$	$(y_i - \bar{y})^2$	$(x_i - \bar{x})(y_i - \bar{y})$	$\widehat{y}_i$	$\widehat{E}_i$
1	50	243.93	-50	2500	-171.328182	29353.3459	8566.40909	263.235	-19.305
2	60	361.95	-40	1600	-53.3081818	2841.76225	2132.32727	293.639	68.311
3	70	286.37	-30	900	-128.888182	16612.1634	3866.64545	324.044	-37.674
4	80	302.25	-20	400	-113.008182	12770.8492	2260.16364	354.449	-52.199
5	90	371.58	-10	100	-43.6781818	1907.78357	436.781818	384.853	-13.273
6	100	437.67	0	0	22.4118182	502.289594	0	415.258	22.412
7	110	532.59	10	100	117.331818	13766.7556	1173.31818	445.663	86.927
8	120	401.91	20	400	-13.3481818	178.173958	-266.963636	476.068	-74.158
9	130	561.66	30	900	146.401818	21433.4924	4392.05455	506.472	55.188
10	140	513.88	40	1600	98.6218182	9726.26302	3944.87273	536.877	-22.997
11	150	554.05	50	2500	138.791818	19263.1688	6939.59091	567.282	-13.232
n	$\bar{x}$	$\bar{y}$	$\sum_{i=1}^{n}(x_i - \bar{x})$	$\sum_{i=1}^{n}(x_i - \bar{x})^2$	$\sum_{i=1}^{n}(y_i - \bar{y})$	$\sum_{i=1}^{n}(y_i - \bar{y})^2$	$\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})$	Media de $\widehat{y}$	Media de $\widehat{E}$
11	100	415.258	0	11000	0	128356.048	33445.2	415.258	0
$\widehat{\beta}_1 = \sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y}) / \sum_{i=1}^{n}(x_i - \bar{x})^2 = 3.041$; $\widehat{\beta}_0 = \bar{y} - \widehat{\beta}_1 \bar{x} = 111.211$; SSE $= \sum_{i=1}^{n} \widehat{E}_i^2 = 26666.83$									

La ecuación de la recta ajustada corresponde a $\widehat{y_i} = 111.211 + 3.041 x_i$ donde la suma de los cuadrados de los residuos es denotada por SSE y es del orden de 26666.83. La recta ajustada se muestra en la Figura 2.3(b). Dada la magnitud de la dispersión y la cantidad de observaciones, las estimaciones de los parámetros distan mucho de los correspondientes valores para los parámetros del modelo generador de los datos. En la Figura 2.3(a) se muestra la recta ajustada ignorando la relación con X, es decir, bajo el modelo $Y_i = \beta_0 + E_i$, $E_i \overset{\text{iid}}{\sim} N(0, \sigma^2)$. Puede mostrarse que en este último modelo se cumple que $\widehat{\beta}_0 = \bar{y}$ y la variabilidad total observada en la variable respuesta, denotada por SST, donde SST $= \sum_{i=1}^{n}(y_i - \bar{y})^2$, es también la suma de cuadrados de residuos, y para los datos observados vale $128356,0476$. Al comparar los valores SSE de ambos modelos, es claro que el modelo de regresión lineal conduce a menor valor de SSE, como se espera de acuerdo a la definición de los estimadores mínimos cuadráticos. Adicionalmente, observe que el centro de los datos, es decir $(\bar{x}, \bar{y})$, cae sobre la recta de ajuste.

2.3.4. Sumas de cuadrados y de productos cruzados

En la Tabla 2.2 se resumen los principales tipos de sumas que se originan en el procedimiento de ajuste por mínimos cuadrados.



(a) (b)

Figura 2.3: Ilustración del criterio de mínimos cuadrados. (a) Recta ajustada asumiendo como modelo a $Y_i = \beta_0 + E_i$, $E_i \overset{\text{iid}}{\sim} N\left(0, \sigma^2\right)$, (b) Recta ajustada mediante mínimos cuadrados ordinarios con el modelo $Y_i = \beta_0 + \beta_1 x_i + E_i$, $E_i \overset{\text{iid}}{\sim} N\left(0, \sigma^2\right)$. La suma de cuadrados de residuos, SSE $= \sum_{i=1}^{n} \widehat{E}_i^2 = 26666.83$, del ajuste en (b) es menor que en (a); para este último se tiene que SSE = SST $= \sum_{i=1}^{n} \left(y_i - \bar{y}\right)^2 = 128356$, en cambio en (b), SSE < SST.

Tabla 2.2: Principales sumas en el ajuste por mínimos cuadrados

Tipo de sumas	Expresión
Suma de cuadrados corregidos en x:	$S_{xx} = \sum_{i=1}^{n}(x_i - \bar{x})^2 = \sum_{i=1}^{n} x_i^2 - n\bar{x}^2 = \sum_{i=1}^{n}(x_i - \bar{x})x_i$
Suma de cuadrados corregidos en y. También es conocida como suma de cuadrados totales o SST:	$S_{yy} = \sum_{i=1}^{n}(y_i - \bar{y})^2 = \sum_{i=1}^{n} y_i^2 - n\bar{y}^2 = \sum_{i=1}^{n}(y_i - \bar{y})y_i$
Suma de productos cruzados:	$S_{xy} = \sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y}) = \sum_{i=1}^{n}(x_i - \bar{x})y_i$
Suma de los cuadrados de los residuos o SSE. Es la estimación de $S(\beta_o, \beta_1)$. Sea $\widehat{E}_i = y_i - \widehat{y}_i$ el i-ésimo residuo, entonces:	$\text{SSE} = \sum_{i=1}^{n} \widehat{E}_i^2 = \sum_{i=1}^{n}(y_i - \widehat{y}_i)^2 = S_{yy} - \widehat{\beta}_1 S_{xy}$
Suma de cuadrados de regresión SSR:	$\text{SSR} = \sum_{i=1}^{n}(\widehat{y}_i - \bar{y})^2 = \widehat{\beta}_1^2 S_{xx} = \widehat{\beta}_1 S_{xy}$

Nota 2.4. $\widehat{\beta}_1$ puede ser expresado en función de S_{xy} y de S_{xx} así:

$$\widehat{\beta}_1 = \frac{S_{xy}}{S_{xx}} \tag{2.18}$$

2.4. Estimación por máxima verosimilitud (Estimadores MLE)

El método de mínimos cuadrados produce los mejores estimadores lineales insesgados para los parámetros de la recta y puede ser usado para la estimación de parámetros de un modelo de regresión lineal sin consideraciones distribucionales sobre los errores. Sin embargo, para poder aplicar testes de hipótesis y construir intervalos de confianza, es necesario realizar y validar tales supuestos. Considerando para el modelo de regresión lineal simple los supuestos de normalidad, independencia y varianza constante para los errores, podemos usar el método de estimación de máxima verosimilitud (MLE). Sea $(x_1, y_1), (x_2, y_2), \ldots, (x_n, y_n)$ los n pares de datos observados, donde, $Y|X_i = \beta_0 + \beta_1 X_i + E_i$, con $\mathrm{E}[Y|X_i] = \beta_0 + \beta_1 X_i$, y los errores E_i satisfaciendo: $E_i \overset{\text{iid}}{\sim} N(0, \sigma^2)$. Asumiendo fijos los niveles o valores en que X es observada, vimos que $Y|X_i \sim N(\beta_0 + \beta_1 X_i, \sigma^2)$.

Sean $\mathbf{x} = (x_1, \ldots, x_n)$ y $\mathbf{y} = (y_1, \ldots, y_n)$, los valores de X y Y, respectivamente, observados en la muestra de tamaño n. La función de verosimilitud para los parámetros del modelo: $(\beta_0, \beta_1, \sigma^2)$, es denotada por $L(\beta_0, \beta_1, \sigma^2|\mathbf{x}, \mathbf{y})$, y es hallada a partir de la distribución conjunta de las variables Y_i, evaluada en las observaciones $y_1, \ldots, y_n$: $f(y_1, \ldots, y_n|\beta_0, \beta_1, \sigma^2)$, que por la condición de independencia es igual al producto de las densidades de probabilidad marginales, por tanto, bajo los supuestos del modelo normal, podemos escribir

$$\begin{aligned} L\left(\beta_0, \beta_1, \sigma^2|\mathbf{x}, \mathbf{y}\right) =& f\left(y_1, \ldots, y_n|\beta_0, \beta_1, \sigma^2\right) = \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left\{-\frac{1}{2\sigma^2}\left(y_i - \beta_0 - \beta_1 x_i\right)^2\right\} \\ &= \left(2\pi\sigma^2\right)^{-n/2} \exp\left\{-\frac{1}{2\sigma^2}\sum_{i=1}^{n}\left(y_i - \beta_0 - \beta_1 x_i\right)^2\right\}. \end{aligned} \tag{2.19}$$

El objetivo es hallar el valor de los parámetros desconocidos $\beta_0, \beta_1, \sigma^2$, que maximicen L, o equivalentemente, que maximicen el logaritmo natural de la verosimilitud, que denotaremos por $\log L$,

$$\log L = -\frac{n}{2}\log(2\pi) - \frac{n}{2}\log\left(\sigma^2\right) - \frac{1}{2\sigma^2}\sum_{i=1}^{n}\left(y_i - \beta_0 - \beta_1 x_i\right)^2. \tag{2.20}$$

Observe que para cualquier valor fijo de σ^2, $\log L$ puede ser maximizada como una función de β_0 y β_1. Sean $\widetilde{\beta}_0$, $\widetilde{\beta}_1$ los valores de los parámetros que maximizan a $\log L$ y por tanto a L, con σ^2 fijo. En la ecuación (2.20), vemos que para maximizar $\log L$ como función de β_0 y β_1, es necesario minimizar $S(\beta_0, \beta_1) = \sum_{i=1}^{n}\left(y_i - \beta_0 - \beta_1 x_i\right)^2$. Por tanto, bajo el modelo normal, los estimadores de máxima verosimilitud son iguales a los respectivos estimadores de mínimos cuadrados, $\widehat{\beta}_0$, $\widehat{\beta}_1$.

Para hallar el estimador MLE para σ^2, reemplazamos los estimadores MLE de los paráme-

tros de regresión, $\widetilde{\beta}_0$, $\widetilde{\beta}_1$, en (2.20),

$$\log L = -\frac{n}{2}\log(2\pi) - \frac{n}{2}\log\left(\sigma^2\right) - \frac{1}{2\sigma^2}\sum_{i=1}^{n}\left(y_i - \widetilde{\beta}_0 - \widetilde{\beta}_1 x_i\right)^2,$$

ecuación sobre la que se aplica el procedimiento de maximización y se encuentra que el estimador MLE para σ^2, que denotaremos por $\widetilde{\sigma^2}$, es igual a

$$\widetilde{\sigma}^2 = \frac{1}{n}\sum_{i=1}^{n}\left(y_i - \widetilde{\beta}_0 - \widetilde{\beta}_1 x_i\right)^2 = \frac{1}{n}\sum_{i=1}^{n}\left(y_i - \widehat{\beta}_0 - \widehat{\beta}_1 x_i\right)^2 = \frac{1}{n}\sum_{i=1}^{n}\left(y_i - \widehat{y}_i\right)^2 = \frac{1}{n}\text{SSE} \tag{2.21}$$

2.5. Estimación insesgada de la varianza σ^2

Puede demostrarse que bajo los supuestos del modelo en relación a los errores, un estimador insesgado de la varianza es:

$$\widehat{\sigma}^2 = \frac{\text{SSE}}{n-2}, \tag{2.22}$$

y es tal que, $\text{E}\left[\widehat{\sigma}^2\right] = \sigma^2$. Esta estimador también recibe el nombre de error cuadrático medio de la regresión y denotado por MSE.

Observe que, podemos escribir al estimador MLE, $\widetilde{\sigma}^2$, de la siguiente forma

$$\widetilde{\sigma}^2 = \left(\frac{n-2}{n}\right)\widehat{\sigma}^2, \tag{2.23}$$

por tanto $\text{E}\left[\widetilde{\sigma}^2\right] = \left(\frac{n-2}{n}\right)\sigma^2$, es decir, el estimador MLE, $\widetilde{\sigma}^2$, es un estimador sesgado de la varianza, aunque asintóticamente es insesgado, pues

$$\lim_{n\to\infty}\text{E}\left[\widetilde{\sigma}^2\right] = \lim_{n\to\infty}\left(\frac{n-2}{n}\right)\sigma^2 = \sigma^2. \tag{2.24}$$

Nota 2.5. También puede demostrarse que los estimadores MLE son de mínima varianza cuando son comparados a todos los posibles estimadores insesgados y además son consistentes, es decir, a medida que aumenta el tamaño de la muestra, la diferencia entre estos y el respectivo parámetro va para cero.

2.6. Propiedades de los estimadores de mínimos cuadrados bajo el modelo normal

Bajo la validez de los supuestos considerados sobre los errores, tenemos que:

1. Los estimadores de mínimos cuadrados, $\widehat{\beta}_0$ y $\widehat{\beta}_1$, son los mejores estimadores lineales **insesgados** de β_0 y β_1, respectivamente, y son iguales a los estimadores de máxima verosimilitud bajo los supuestos estadísticos del modelo lineal normal. Por tanto, $\widehat{Y} = \widehat{\beta}_0 + \widehat{\beta}_1 X$, es un estimador insesgado para $\text{E}\left[Y|X\right] = \beta_0 + \beta_1 X$.

2. $\widehat{\beta}_0$ y $\widehat{\beta}_1$ son combinaciones lineales de las variables aleatorias $Y_1, ..., Y_n$, pues estos pueden escribirse en la forma que muestra la Tabla 2.3, por tanto, como $Y_1, ..., Y_n$ son variables normales e incorrelacionadas, entonces $\widehat{\beta}_0$ y $\widehat{\beta}_1$ son variables aleatorias normales.

Tabla 2.3: Estimadores como combinaciones lineales de la respuesta

Cantidad estimada	Estimador como combinación lineal de los Y_i	Peso de Y_i en la combinación
Intercepto	$\widehat{\beta}_0 = \sum_{i=1}^{n} m_i Y_i$	$m_i = \frac{1}{n} - \bar{x}c_i$
Pendiente	$\widehat{\beta}_1 = \sum_{i=1}^{n} c_i Y_i$	$c_i = \frac{x_i - \bar{x}}{S_{xx}}$
Respuesta para $X = x_i$	$\widehat{Y}_i = \sum_{j=1}^{n} h_{ij} Y_j$	$h_{ij} = m_j + c_j x_i = \frac{1}{n} + \frac{(x_i-\bar{x})(x_j-\bar{x})}{S_{xx}}$, para $j = i:\ h_{ii} = \frac{1}{n} + \frac{(x_i-\bar{x})^2}{S_{xx}}$

3. Como consecuencia de la propiedad anterior, se puede mostrar que la respuesta estimada $\widehat{Y}_i$ también es una combinación lineal de las variables aleatorias $Y_1, \ldots, Y_n$, como muestra la Tabla 2.3, y es una variable normal.

4. La varianza de los estimadores es según se describe en la Tabla 2.4. Para la obtención de los resultados debe tenerse en cuenta que las variables $Y_1, \ldots, Y_n$ son independientes y tienen la misma varianza, así como las expresiones de sumas de cuadrados dadas previamente.

Tabla 2.4: Varianza de los estimadores. Tenga en cuenta que bajo los supuestos del modelo, se tiene la independencia (ind) entre las variables aleatorias Y_i y la homogeneidad de varianza (hv).

Para $\widehat{\beta}_0$	$\widehat{\beta}_1$	Para $\widehat{Y}_i$
$\text{Var}\left[\widehat{\beta}_0\right] = \text{Var}\left[\sum_{i=1}^{n} m_i Y_i\right]$	$\text{Var}\left[\widehat{\beta}_1\right] = \text{Var}\left[\sum_{i=1}^{n} c_i Y_i\right]$	$\text{Var}\left[\widehat{Y}_i\right] = \text{Var}\left[\sum_{j=1}^{n} h_{ij} Y_j\right]$
entonces,	entonces,	entonces,
$\text{Var}\left[\widehat{\beta}_0\right] \overset{\text{ind, hv}}{=} \sum_{i=1}^{n} m_i^2 \sigma^2$	$\text{Var}\left[\widehat{\beta}_1\right] \overset{\text{ind, hv}}{=} \sum_{i=1}^{n} c_i^2 \sigma^2$	$\text{Var}\left[\widehat{Y}_i\right] \overset{\text{ind, hv}}{=} \sum_{j=1}^{n} h_{ij}^2 \sigma^2$
de donde,	de donde,	de donde,
$\text{Var}\left[\widehat{\beta}_0\right] = \frac{\sum_{i=1}^{n} x_i^2}{nS_{xx}}\ \sigma^2$	$\text{Var}\left[\widehat{\beta}_1\right] = \frac{\sigma^2}{S_{xx}}$	$\text{Var}\left[\widehat{Y}_i\right] = \left[\frac{1}{n} + \frac{(x_i-\bar{x})^2}{S_{xx}}\right]\sigma^2$
o bien,		
$\text{Var}\left[\widehat{\beta}_0\right] = \left[\frac{1}{n} + \frac{\bar{x}^2}{S_{xx}}\right]\sigma^2$		

5. La covarianza entre los estimadores de los parámetros es:

$$\begin{aligned}
\text{Cov}\left(\widehat{\beta}_0, \widehat{\beta}_1\right) &= \text{Cov}\left(\sum_{i=1}^{n} m_i Y_i, \sum_{i=1}^{n} c_i Y_i\right) \\
&= \sum_{i=1}^{n} m_i c_i \text{Cov}\left(Y_i, Y_i\right) + \sum_{i=1}^{n}\sum_{j\neq i}^{n} m_i c_j \text{Cov}\left(Y_i, Y_j\right) \\
&= \sum_{i=1}^{n} m_i c_i \text{Var}\left[Y_i\right] \\
&= \sigma^2 \sum_{i=1}^{n} m_i c_i \\
&= -\frac{\bar{x}}{S_{xx}}\sigma^2 \qquad (2.25)
\end{aligned}$$

6. La covarianza entre la variable respuesta y su correspondiente estimador, en un valor dado x_i, es:

$$\begin{aligned}
\text{Cov}\left(Y_i, \widehat{Y}_i\right) &= \text{Cov}\left(Y_i, \sum_{j=1}^{n} h_{ij} Y_j\right) \\
&= h_{ii}\text{Cov}\left(Y_i, Y_i\right) + \sum_{j\neq i}^{n} h_{ij}\text{Cov}\left(Y_i, Y_j\right) \\
&= h_{ii}\sigma^2 \\
&= \left[\frac{1}{n} + \frac{(x_i - \bar{x})^2}{S_{xx}}\right]\sigma^2 \qquad (2.26)
\end{aligned}$$

7. La covarianza entre la variable respuesta ajustada en x_i y la ajustada en x_k, con $i, k \in \{1, 2, \ldots, n\}$, $i \neq k$, es,

$$\begin{aligned}
\text{Cov}\left(\widehat{Y}_i, \widehat{Y}_k\right) &= \text{Cov}\left(\widehat{\beta}_0 + \widehat{\beta}_1 x_i, \widehat{\beta}_0 + \widehat{\beta}_1 x_k\right) \\
&= \text{Cov}\left(\widehat{\beta}_0, \widehat{\beta}_0\right) + x_i\text{Cov}\left(\widehat{\beta}_0, \widehat{\beta}_1\right) + x_k\text{Cov}\left(\widehat{\beta}_0, \widehat{\beta}_1\right) + x_i x_k\text{Cov}\left(\widehat{\beta}_1, \widehat{\beta}_1\right) \\
&= \text{Var}\left[\widehat{\beta}_0\right] + (x_i + x_k)\,\text{Cov}\left(\widehat{\beta}_0, \widehat{\beta}_1\right) + x_i x_k\text{Var}\left[\widehat{\beta}_1\right] \\
&= \left[\frac{1}{n} + \frac{\bar{x}^2}{S_{xx}}\right]\sigma^2 - \frac{\bar{x}\,(x_i + x_k)}{S_{xx}}\sigma^2 + \frac{x_i x_k}{S_{xx}}\sigma^2 \\
&= \left[\frac{1}{n} + \frac{(x_i - \bar{x})\,(x_k - \bar{x})}{S_{xx}}\right]\sigma^2 \qquad (2.27)
\end{aligned}$$

8. La suma de los residuales del modelo de regresión con intercepto es siempre cero:

$$\sum_{i=1}^{n} \widehat{E}_i = 0 \tag{2.28}$$

9. La suma de los valores observados y_i es igual a la suma de los valores ajustados $\widehat{y}_i$:

$$\sum_{i=1}^{n} y_i = \sum_{i=1}^{n} \widehat{y}_i \tag{2.29}$$

10. La recta de regresión siempre pasa a través del centroide de los datos $(\bar{x}, \bar{y})$.

11. La suma de los residuales ponderados por el correspondiente valor de la variable predictora es cero:

$$\sum_{i=1}^{n} x_i \widehat{E}_i = 0 \tag{2.30}$$

12. La suma de los residuales ponderados por el correspondiente valor ajustado es siempre igual a cero:

$$\sum_{i=1}^{n} \widehat{y}_i \widehat{E}_i = 0 \tag{2.31}$$

2.7. Inferencias sobre los parámetros del modelo de regresión

Sujetos a los supuestos del modelo de regresión, podemos realizar pruebas de hipótesis sobre los parámetros, donde los estadístico de prueba bajo la respectiva hipótesis nula, se distribuyen como una variable aleatoria t-Student cuyos grados de libertad dependen de los grados de libertad del estimador de la varianza según el modelo (o sea, el MSE). En particular, interesa realizar los test de significancia para los parámetros, y la estimación por intervalo de confianza. La Tabla 2.5 muestra los elementos de las pruebas de significancia individuales y las expresiones de los intervalos de confianza de nivel $(1-\alpha)\,100\,\%$.

Nota 2.6. Tenga en cuenta lo siguiente:

- t_{n-2} es la *v.a* t-Student con $n-2$ grados de libertad, en tanto que $t_{\alpha/2,n-2}$ es un percentil de la distribución t-Student con $n-2$ grados de libertad, tal que, $P\left(t_{n-2} > t_{\alpha/2,n-2}\right) = \alpha/2$.
- Si la pendiente de la recta de regresión es significativa, entonces el modelo de regresión lineal simple también lo es, es decir, la variabilidad en la variable respuesta explicada por la regresión en X es significativa respecto a la variabilidad total observada.
- Para pruebas sobre los parámetros, diferentes a los tests de significancia presentados en la Tabla 2.5, en la construcción de los estadísticos T_0, β_0 y β_1 toman los valores especificados en la correspondiente hipótesis H_0, mientras que los criterios de rechazo se deben establecer según la desigualdad planteada en la hipótesis alternativa H_1.

Tabla 2.5: Pruebas de significancia en intervalos de confianza (I.C) sobre los coeficientes de regresión.

Parámetro	Test de significancia	Estadístico de prueba	Criterio de rechazo	I.C del $(1-\alpha)\,100\,\%$
β_0	$H_0:\ \beta_0 = 0$ $H_1:\ \beta_0 \neq 0$	$T_0 = \dfrac{\widehat{\beta}_0 - \beta_0}{\sqrt{\dfrac{\text{MSE}\sum_{i=1}^{n} x_i^2}{nS_{xx}}}} \sim t_{n-2}$ con $\beta_0 = 0$ en el test de significancia.	con nivel α : si $\lvert T_0\rvert > t_{\alpha/2,n-2}$ con valor P, si: $P\left(\lvert t_{n-2}\rvert > \lvert T_0\rvert\right)$ es pequeño.	$\widehat{\beta}_0 \pm t_{\alpha/2,n-2}\sqrt{\dfrac{\text{MSE}\sum_{i=1}^{n} x_i^2}{nS_{xx}}}$
β_1	$H_0:\ \beta_1 = 0$ $H_1:\ \beta_1 \neq 0$	$T_0 = \dfrac{\widehat{\beta}_1 - \beta_1}{\sqrt{\dfrac{\text{MSE}}{S_{xx}}}} \sim t_{n-2}$ con $\beta_1 = 0$ en el test de significancia.	con nivel α : si $\lvert T_0\rvert > t_{\alpha/2,n-2}$ con valor P, si: $P\left(\lvert t_{n-2}\rvert > \lvert T_0\rvert\right)$ es pequeño.	$\widehat{\beta}_1 \pm t_{\alpha/2,n-2}\sqrt{\dfrac{\text{MSE}}{S_{xx}}}$

2.8. Inferencias respecto a la respuesta media $\mu_{y|x_0}$ y valores futuros

Desde que los valores ajustados de la variable respuesta también son combinaciones lineales de las variables aleatorias $Y_1, \ldots, Y_n$, como fue mostrado previamente en la Tabla 2.3, bajo los supuestos de normalidad e independencia entre los errores, podemos afirmar que las variables $\widehat{Y}_i$, $i = 1, 2, \ldots, n$, son variables aleatorias normales, aunque no son independientes. Vea de nuevo la ecuación (2.27). Recuerde también que $\widehat{Y}_i$ estima a $\mu_{y|x_i} = \mathrm{E}\left[Y|X = x_i\right]$. Podemos hacer inferencias sobre esta media y además predecir un valor futuro Y_0 de la respuesta en un valor fijo de $X = x_0$. Así, bajo los supuestos del modelo obtenemos los resultados exhibidos en la Tabla 2.6.

Mientras que el intervalo de confianza para $\mu_{y|x_0}$ proporciona un rango en el cual pudiera estar la media de la respuesta para $X = x_0$, con el nivel de confianza dado, el intervalo de predicción en un valor $X = x_0$, estima, con el nivel de confianza dado, el rango de los posibles valores en el cual podría ser observado el valor de la variable respuesta. Asumimos que en este valor particular x_0 obtenemos un valor futuro de la variable aleatoria Y_0, y por tanto, éste no ha sido utilizado en el ajuste del modelo de regresión. La predicción de Y_0 con base en el modelo ajustado con la muestra de pares (x_i, y_i), $i = 1, 2, \ldots, n$, corresponde a $\widehat{Y}_0 = \widehat{\beta}_0 + \widehat{\beta}_1 x_0$, y dado que Y_0 no hizo parte de la muestra de ajuste, las variables aleatorias Y_0 y $\widehat{Y}_0$ son estadísticamente independientes. De esto se desprende que el error del pronóstico: $\widehat{E}_0 = Y_0 - \widehat{Y}_0$, es una variable aleatoria normal cuya media es cero y la varianza es igual a

$$\mathrm{Var}\left[\widehat{E}_0\right] = \mathrm{Var}\left[Y_0\right] + \mathrm{Var}\left[\widehat{Y}_0\right] = \left[1 + \frac{1}{n} + \frac{\left(x_0 - \bar{x}\right)^2}{S_{xx}}\right]\sigma^2. \tag{2.32}$$

Tabla 2.6: Inferencias sobre la respuesta media y respuesta futura

Para la respuesta media en $X = x_0$			
Cantidad	Hipótesis H_0	Estadístico de prueba	Intervalo de confianza del $(1-\alpha)\,100\,\%$
$\mu_{y\|x_0}$	$\mu_{y\|x_0} = c$	$T_0 = \dfrac{\widehat{Y}_0 - \mu_{y\|x_0}}{\sqrt{\text{MSE}\left[\frac{1}{n} + \frac{(x_0-\bar{x})^2}{S_{xx}}\right]}} \sim t_{n-2}$ con $\mu_{y\|x_0} = c,\ \widehat{Y}_0 = \widehat{\beta}_0 + \widehat{\beta}_1 x_0$	$\widehat{Y}_0 \pm t_{\alpha/2,n-2}\sqrt{\text{MSE}\left[\frac{1}{n} + \frac{(x_0-\bar{x})^2}{S_{xx}}\right]}$ con $\widehat{Y}_0 = \widehat{\beta}_0 + \widehat{\beta}_1 x_0$
Para una respuesta futura en $X = x_0$			
Cantidad	Pronóstico	Estadístico	Intervalos de predicción del $(1-\alpha)\,100\,\%$
Y_0	$\widehat{Y}_0$	$T_0 = \dfrac{\widehat{Y}_0 - Y_0}{\sqrt{\text{MSE}\left[1 + \frac{1}{n} + \frac{(x_0-\bar{x})^2}{S_{xx}}\right]}} \sim t_{n-2}$ con $\widehat{Y}_0 = \widehat{\beta}_0 + \widehat{\beta}_1 x_0$	$\widehat{Y}_0 \pm t_{\alpha/2,n-2}\sqrt{\text{MSE}\left[1 + \frac{1}{n} + \frac{(x_0-\bar{x})^2}{S_{xx}}\right]}$ con $\widehat{Y}_0 = \widehat{\beta}_0 + \widehat{\beta}_1 x_0$

Nota 2.7. En general, no se recomienda realizar extrapolaciones por fuera del rango de variación observado en el conjunto de datos sobre la variable explicatoria. Por ello es importante que en los datos muestrales se haya cubierto el rango en el que naturalmente pudiera tomar valores la variable explicatoria. De lo contrario, es posible que en un rango fuera del observado, la relación estadística formulada no resulte válida. Considere la situación planteada en la Figura 2.4. Allí, si sólo se consideran las observaciones para el intervalo $[x_1, x_2]$, la recta ajustada en ese tramo causará un gran error, para todas las extrapolaciones en el intervalo $x \in [x_2, x_3]$.

Nota 2.8. En situaciones donde la curva que explica la relación entre Y y X es difícilmente representada por una recta o alguna función suave de x, como la ilustrada en la Figura 2.4, es preferible el uso de métodos no paramétricos como la regresión local, los suavizamientos splines, entre otros.

2.9. Análisis de varianza para probar la significancia de la regresión

El análisis de varianza o ANOVA, consiste en la descomposición de la variabilidad total observada en la variable respuesta, denotada por SST, en la suma de componentes o fuentes de variabilidad, de acuerdo al modelo propuesto. Recuerde que el modelo de regresión lineal simple plantea que la respuesta es igual a la suma de una componente real no aleatoria (la función de regresión dada por $\beta_0 + \beta_1 X$) y un error aleatorio E. Se espera que la recta ajustada explique en forma significativa la variabilidad observada en Y. Bajo los supuestos del modelo, es posible demostrar que la la variabilidad total muestral de la respuesta satisface la siguiente descomposición:

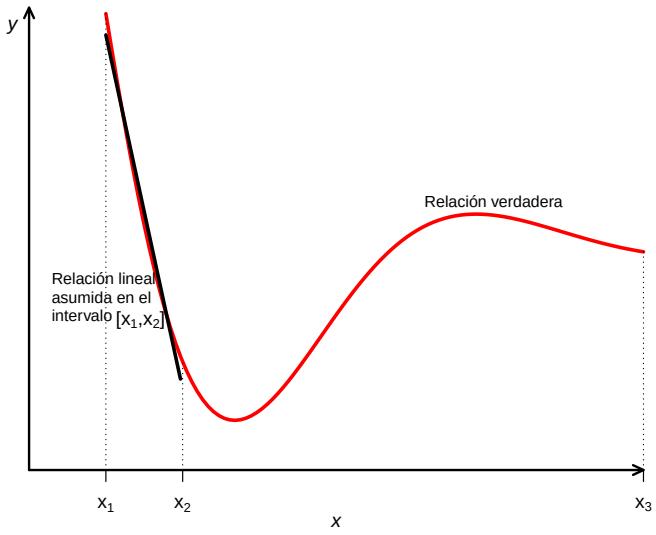


Figura 2.4: Extrapolación dañina. Al considerar solamente datos (x, y) en el intervalo $x_1 \leq x \leq x_2$, el modelo de regresión lineal lograría una buena aproximación de la verdadera relación, pero al extrapolar con la recta ajustada por fuera del rango $x_1 \leq x \leq x_2$, ya no tendríamos un buen desempeño del modelo, pues claramente, la extrapolación causaría errores significativos con respecto a la estimación de la relación verdadera.

$$\underbrace{\sum_{i=1}^{n}(Y_i - \bar{Y})^2}_{\textbf{Variabilidad total}} = \underbrace{\sum_{i=1}^{n}(\widehat{Y}_i - \bar{Y})^2}_{\textbf{Variabilidad explicada}} + \underbrace{\sum_{i=1}^{n}(Y_i - \widehat{Y}_i)^2}_{\textbf{Variabilidad no explicada}} \tag{2.33}$$

De donde (repase resultados en Tabla 2.2):

$$\text{SST} = \text{SSR} + \text{SSE} = \widehat{\beta}_1 S_{xy} + \text{SSE} = \widehat{\beta}_1^2 S_{xx} + \text{SSE} \tag{2.34}$$

En virtud de la anterior igualdad, podemos también establecer la siguiente identidad para los grados de libertad (g.l) de las sumas de cuadrados:

$$\underbrace{\text{g.l}\,(\text{SST})}_{\textbf{n−1}} = \underbrace{\text{g.l}\,(\text{SSR})}_{\textbf{1}} + \underbrace{\text{g.l}\,(\text{SSE})}_{\textbf{n−2}} \tag{2.35}$$

Nota 2.9. Bajo el supuesto de que los errores del modelo son variables aleatorias independientes e idénticamente distribuidos como una normal de media cero, es decir, satisfacen que $E_i \overset{\text{iid}}{\sim} N(0, \sigma^2)$, pueden demostrarse los siguientes resultados;

1. Cuando $\beta_1 = 0$, $\text{SSR}/\sigma^2 \sim \chi^2_1$

2. $\text{SSE}/\sigma^2 \sim \chi^2_{n-2}$

3. SSR/σ^2 y SSE/σ^2 son estadísticamente independientes.

4. De las anteriores propiedades podemos decir que bajo $H_0 : \beta_1 = 0$, el estadístico[1],

$$F_0 = \frac{\text{SSR}/\text{g.l}\,(\text{SSR})}{\text{SSE}/\text{g.l}\,(\text{MSE})} = \frac{\text{SSR}/1}{\text{SSE}/\,(n-2)} = \frac{\text{SSR}}{\text{MSE}} \sim f_{1,n-2} \tag{2.36}$$

5. F_0 de la ecuación anterior es igual al cuadrado del estadístico T_0 del test de significancia de la pendiente β_1 dado en la Tabla 2.5, es decir[2],

$$F_0 = \frac{\text{SSR}}{\text{MSE}} = \left[\frac{\widehat{\beta}_1}{\sqrt{\dfrac{\text{MSE}}{S_{xx}}}}\right]^2 \sim f_{1,n-2} \tag{2.37}$$

De lo anterior, se concluye que en el caso de la regresión lineal simple, la prueba sobre la significancia de la regresión es equivalente a probar si la pendiente de la recta es significativamente diferente de cero, es decir, el test puede realizarse de dos maneras: test t para β_1 o mediante el análisis de varianza, pero en este último, el criterio de decisión a un nivel de significancia α es usando un valor crítico $f_{\alpha,1,n-2}$, con el cual rechazamos la hipótesis nula de que la variabilidad en la variable respuesta es debida sólo al error aleatorio (para aceptar la hipótesis de que la regresión en x es significativa), si $F_0 > f_{\alpha,1,n-2}$, donde $f_{\alpha,1,n-2}$ es tal que $P\,(f_{1,n-2} > f_{\alpha,1,n-2}) = \alpha$.

El análisis de varianza suele presentarse en forma de Tabla, conocida como la tabla ANOVA, donde los cuadrados medios corresponden a las sumas de cuadrados que allí se discriminan, divididas por sus respectivos grados de libertad:

Tabla 2.7: Tabla ANOVA del modelo de regresión lineal simple

Fuente de variación	Suma de cuadrados	Grados de libertad	Cuadrados medios	Cuadrados medios esperados	Estadístico F_0	Valor P
Regresión	SSR	1	$\text{MSR} = \text{SSR}/1$	$\text{E}\,[\text{MSR}] = \sigma^2 + \beta_1^2 S_{xx}$	MSR/MSE	$P\,(f_{1,n-2} > F_0)$
Error	SSE	$n-2$	$\text{MSE} = \text{SSE}/\,(n-2)$	$\text{E}\,[\text{MSE}] = \sigma^2$		
Total	SST	$n-1$	$\text{MST} = \text{SST}/\,(n-1)$			

También podemos evaluar el valor p de la prueba (significancia más pequeña, según los datos, que conduce al rechazo de H_0) el cual es igual a $P(f_{1,n-2} > F_0)$, y si es "pequeño", entonces se rechaza la hipótesis:

[1] Aquí debe recordar la definición de una variable aleatoria f_{ν_1,ν_2} como el cociente de dos variables aleatorias chi cuadrados independientes: $X_1 \sim \chi^2_{\nu_1}$, y $X_2 \sim \chi^2_{\nu_2}$, divididas por sus respectivos grados de libertad.

[2] Un resultado asociado es el siguiente: Si $T \sim t_\nu$, entonces $T^2 \sim f_{1,\nu}$.

H_0 : "El modelo lineal de Y en X no es significativo para explicar la variabilidad de Y"

$\iff$

$H_0 : \beta_1 = 0.$

vs.

H_1 : "El modelo lineal de Y en X es significativo para explicar la variabilidad de Y"

$\iff$

$H_1 : \beta_1 \neq 0.$ (2.38)

De lo anterior, note que con el test ANOVA se está rechazando la hipótesis nula cuando se tiene un estadístico F_0 "grande" bajo la distribución $f_{1,n-2}$. De otro lado, dada la equivalencia entre el test t sobre β_1 y el test ANOVA para el modelo, la conclusión obtenida por el análisis de varianza debe ser la misma que la obtenida cuando se prueba la significancia individual de β_1.

2.10. R^2 de una regresión

Esta cantidad que aparece en los resultados de la regresión lineal, proviene de la razón SSR/SST y por tanto, podemos interpretarla como *la proporción de la variabilidad total observada en la variable respuesta, que es explicada por la relación lineal con la variable predictora considerada*. Esta cantidad también es conocida como el coeficiente de determinación y ha sido utilizada erróneamente como medida para evaluar la bondad del ajuste lineal, pues si bien valores cercanos a 1 indican una mayor asociación lineal, no necesariamente garantiza que los supuestos básicos del modelo lineal se estén cumpliendo y menos que no haya carencia de ajuste lineal (Montgomery et. al., 2012; Kutner et. al., 2005). Kutner et. al., 2005 lista como los tres errores de interpretación más comunes del R^2, los siguientes:

1. *Creer que un R^2 alto indica que el modelo puede hacer predicciones útiles.* Hay casos donde se tiene un R^2 alto y sin embargo, los intervalos de predicción son muy amplios indicando poca precisión del pronóstico.

2. *Creer que un R^2 alto indica que la recta de regresión ajustada tiene buen ajuste.* Hay casos en los cuales se ajusta una recta obteniendo un R^2 cuando la verdadera relación no es lineal.

3. *Creer que un R^2 cercano a cero indica que X y Y no están relacionados.* Cuando existe una relación no lineal entre X y Y, puede ocurrir que al ajustar considerando linealidad, el R^2 dé cercano a cero.

2.11. Pasos en el Análisis de regresión

1. Comprensión del problema

2. Realizar análisis exploratorio de los datos mediante diagramas de dispersión para establecer el tipo de función de regresión apropiada.

3. Aplicar transformaciones para estabilizar varianza o para simplificar la modelación.

4. Desarrollar uno o más modelos de regresión tentativos.

5. Ajustar los modelos tentativos.

6. Evaluar los modelos ajustados

   a) Analizar de residuales para:

   - Verificar si el modelo es adecuado: Gráfico de residuos vs. x para chequear ausencia de patrones sistemáticos, test de carencia de ajuste.
   - Verificar si los supuestos sobre el término de error se cumplen: Gráficos y test de probabilidad normal; gráficos de residuos vs. valores ajustados de la respuesta para chequear varianza constante y ausencia de patrones sistemáticos.

   b) Diagnóstico de observaciones atípicas e influenciales.

7. Para los modelos que pasen las pruebas en 6:

   - Evaluar calidad del ajuste.
   - Hacer predicciones: Sólo dentro del rango de valores considerados para la variable predictora o valores cercanos a dicho rango y evaluar calidad de pronósticos.

8. Escoger el mejor modelo.

9. Interpretar los parámetros del modelo seleccionado a la luz de los datos.

10. Hacer las inferencias de interés.

11. Reportar resultados

2.12. Regresión lineal en R

En R la función disponible para regresión lineal clásica es la función `lm( )`, en la cual se formula el modelo usando la sintaxis de fórmulas R admisibles, como se ejemplifica en la Tabla 2.8. Todos estos modelos ajustados con la función `lm( )` son estimados por mínimos cuadrados ordinarios.

Tabla 2.8: Fórmulas R en modelos de regresión lineal

Fórmula R	Modelo a ajustar	Corrida con `lm( )`
Y ∼ X, o bien, Y ∼ 1 + X	$Y_i = \beta_0 + \beta_1 X_i + E_i$ $E_i \overset{\text{iid}}{\sim} N(0, \sigma^2)$	lm(Y ∼ X), o bien, lm(Y ∼ 1 + X)
Y ∼ −1 + X, o bien, Y ∼ 0 + X	$Y_i = \beta_1 X_i + E_i$ $E_i \overset{\text{iid}}{\sim} N(0, \sigma^2)$	lm(Y ∼ −1 + X), o bien, lm(Y ∼ 0 + X)
log (Y) ∼ X	$\log(Y_i) = \beta_0 + \beta_1 X_i + E_i$ $E_i \overset{\text{iid}}{\sim} N(0, \sigma^2)$	lm(log (Y) ∼ X)
Y ∼ X1 + X2	$Y_i = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + E_i$ $E_i \overset{\text{iid}}{\sim} N(0, \sigma^2)$	lm(Y ∼ X1 + X2)
Y ∼ X + I(X^2)	$Y_i = \beta_0 + \beta_1 X_i + \beta_2 X_i^2 + E_i$ $E_i \overset{\text{iid}}{\sim} N(0, \sigma^2)$	lm(Y ∼ X + I(X^2))
Y ∼ X1 ∗ X2	$Y_i = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + \beta_3 X_{1i} * X_{2i} + E_i$ $E_i \overset{\text{iid}}{\sim} N(0, \sigma^2)$	lm(Y ∼ X1 ∗ X2)

Otras funciones importantes son

- `summary( )`: Para la obtención de la tabla de parámetros estimados cuando se aplica a un objeto `lm`.
- `anova( )`: Para Tabla ANOVA.
- `residuals( )`: para la obtención de los valores de los residuos de ajuste.
- `fitted( )`: Para la obtención de los valores ajustados en la variable respuesta.
- `predict( )`: Para la obtención de predicciones puntuales y por intervalos de predicción. También permite calcular valores ajustados y los intervalos de confianza para $\mu_{Y|x}$.
- `qqnorm( )`, `qqline( )`: Para gráfico de probabilidad normal.
- `shapiro.test( )`: Para tes de normalidad Shapiro-Wilk.

Código R 2.1. *A continuación se muestra el código básico en regresión lineal simple. Se supone que previamente se han creados los objetos* **`x`**, **`y`**, **`xnuevo`**.

```
#Ajuste del modelo
modelo=lm(y~x)
summary(modelo) #Tabla de parámetros estimados
anova(modelo)   #Tabla ANOVA

#Gráficos con residuales
plot(x,residuals(modelo))              #Residuos vs. x
plot(fitted(modelo),residuals(modelo)) #Residuos vs. Y ajustado
qqnorm(residuals(modelo))              #Gráfico de probabilidad normal
qqline(residuals(modelo))              #Adición de la recta de probabilidad normal
```

```
#Gráfico de dispersión con recta ajustada superpuesta
plot(x,y)                              #Gráfico de dispersión de Y vs. x
lines(x,fitted(modelo),lty=1,col=2)    #Adición de la recta de regresión ajustada

#Predicciones para un vector de valores x guardado como xnuevo
#con intervalos de predicción del 95% de confianza
predict(modelo,newdata=data.frame(x=xnuevo),interval="prediction",level=0.95)

#con intervalos de confianza de la respuesta media del 95% de confianza
predict(modelo,newdata=data.frame(x=xnuevo),interval="confidence",level=0.95)
```

Otras funciones que pueden ser útiles son las siguientes:

- `coef( )`: Extrae en un vector los valores estimados de los parámetros de la regresión.
- `confint( )`: Retorna una matriz en la cual en cada fila se dan los límites de los intervalos de confianza para los parámetros del modelo.
- `vcov( )`: Produce la matriz de varianzas y covarianzas estimadas, para el vector de parámetros estimado.

A continuación se ilustra el uso de estas funciones sobre el modelo ajustado en el problema presentado en la Sección 2.13.2.

Código R 2.2. *Extracción de coeficientes estimados, intervalos de confianza y matriz de covarianzas del vector de parámetros estimados,*

```
modelo=lm(Rapidez.Grabado~Flujo.Cloro)
coef(modelo)
confint(modelo,level=0.95)
vcov(modelo)
```

Salida R 2.1. *Salida en consola R de la ejecución de las funciones invocadas en el Código R 2.2.*

```
> coef(modelo)
(Intercept) Flujo.Cloro
   6.448718   10.602564
> confint(modelo,level=0.95)
                2.5 %   97.5 %
(Intercept) -0.1593835 13.05682
Flujo.Cloro  8.2415238 12.96360
> vcov(modelo)
            (Intercept) Flujo.Cloro
(Intercept)    7.809606   -2.658589
Flujo.Cloro   -2.658589    0.996971
```

La matriz resultante con `vcov(modelo)` corresponde a la siguiente, en donde los cálculos de varianzas y covarianzas estimadas se basan en la fórmulas vistas en la Sección 2.6, usando al MSE en lugar de σ^2:

$$\begin{bmatrix} \widehat{\text{Var}}\left[\widehat{\beta}_0\right] & \widehat{\text{Cov}}\left(\widehat{\beta}_0, \widehat{\beta}_1\right) \\ \widehat{\text{Cov}}\left(\widehat{\beta}_0, \widehat{\beta}_1\right) & \widehat{\text{Var}}\left[\widehat{\beta}_1\right] \end{bmatrix} = \begin{bmatrix} \frac{1}{n}\sum_{i=1}^{n} x_i^2 & -\bar{x} \\ -\bar{x} & 1 \end{bmatrix} \times \frac{\text{MSE}}{S_{xx}} = \begin{bmatrix} 7.809606 & -2.658589 \\ -2.658589 & 0.996971 \end{bmatrix}.$$

2.13. Problemas

2.13.1. Relación lineal en un proceso de destilación

En un proceso de destilación químico se desea establecer la relación entre la pureza del oxígeno producido (Y, en %) y el porcentaje de hidrocarburos presentes en el condensador principal de la unidad de destilación (X). Los datos son los siguientes:

Tabla 2.9: Datos del proceso de destilación

X: % hidrocarburos	Y: % pureza O_2
0.99	90.01
1.02	89.05
1.15	91.43
1.29	93.74
1.46	96.73
1.36	94.45
0.87	87.59
1.23	91.77
1.55	99.42
1.40	93.65
1.19	93.54
1.15	92.52
0.98	90.56
1.01	89.54
1.11	89.85
1.20	90.39
1.26	93.25
1.32	93.41
1.43	94.98
0.95	87.33

Los datos se presentan en la Figura 2.5 donde se ha superpuesto una curva de ajuste no paramétrico conocida como curva LOESS, que puede ayudar a apreciar la tendencia.Se ajusta el modelo de regresión lineal simple y los resultados R del ANOVA y parámetros estimados se muestran en las Tablas 2.10 y 2.11. Interprete a la luz del problema y realice las pruebas de interés sobre el modelo. Las Figuras 2.6(a) y 2.6(b) muestran los gráficos de residuales del modelo lineal ajustado. Evalúe en cada uno el supuesto de varianza constante. En la Figura 2.6(c) se presenta el gráfico de probabilidad normal sobre los residuos y en la Figura 2.7 el gráfico de la recta ajustada, con límites de confianza y de predicción.

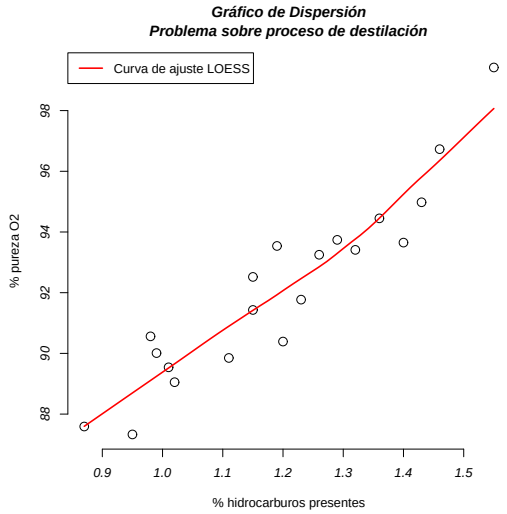


Figura 2.5: Dispersión de los pares de datos y ajuste LOESS datos proceso de destilación

Tabla 2.10: ANOVA en datos proceso de destilación. Compare lo resaltado en color con lo resaltado en Salida R 2.2

Fuente	Df	Sum Sq	Mean Sq	F_0	$P\,(f_{1,18} > F_0)$
% Hidrocarb	1	152.13	152.13	128.86	1.227×10^{-9}
Error	18	21.25	1.18		

Salida R 2.2. *Salida R originada por* `anova(modelo)`

```
> anova(modelo) #Obteniendo ANOVA
Analysis of Variance Table

Response: Porc.PurezaO2
              Df Sum Sq Mean Sq F value    Pr(>F)
Porc.Hidrocarb  1 152.13 152.127  128.86 1.227e-09 ***
Residuals      18  21.25   1.181
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Tabla 2.11: Tabla de Parámetros Estimados. Compare lo resaltado en color con lo resaltado en Salida R 2.3

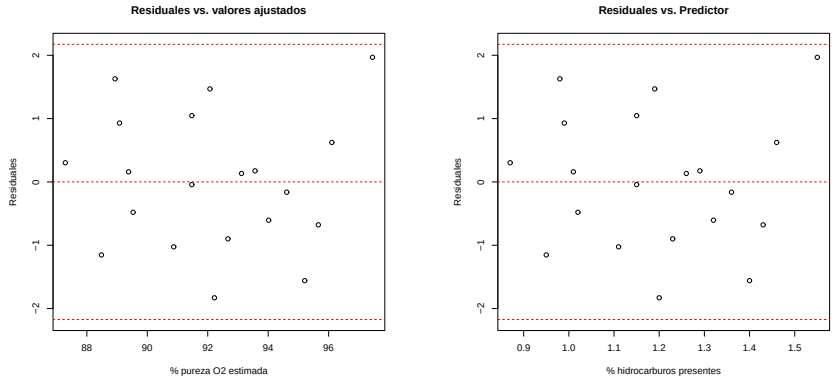
Parámetro	Estimación	Std. Error	T_0	$P\left(\lvert t_{18}\rvert > \lvert T_0\rvert\right)$
β_0	74.2833	1.593	46.62	$< 2 \times 10^{-16}$
β_1	14.9475	1.317	11.35	1.23×10^{-9}
$\sqrt{\text{MSE}} = 1.087$, $R^2 = 0.8774$, $R^2_{adj} = 0.8706$, $F_0 = 128.9$, VP $= P(f_{1,18} > F_0) = 1.227 \times 10^{-9}$.				
La ecuación ajustada es $\widehat{y} = 74.283 + 14.947x$.				

Salida R 2.3. *Salida R originada por* `summary(modelo)`

```
> summary(modelo) #Obteniendo tabla de parámetros ajustados
Call:
lm(formula = Porc.PurezaO2 ~ Porc.Hidrocarb)

Residuals:
     Min       1Q   Median       3Q      Max
-1.83029 -0.73334  0.04497  0.69969  1.96809

Coefficients:
               Estimate Std. Error t value Pr(>|t|)
(Intercept)      74.283      1.593   46.62  < 2e-16 ***
Porc.Hidrocarb   14.947      1.317   11.35 1.23e-09 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 1.087 on 18 degrees of freedom
Multiple R-squared:  0.8774,    Adjusted R-squared:  0.8706
F-statistic: 128.9 on 1 and 18 DF,  p-value: 1.227e-09
```

Con las gráficas de residuales aquí presentadas se evaluará si no hay evidencia en contra de:

- La media de los errores E_i es cero para todo i;
- La varianza de los E_i es igual para todo i;
- Adicionalmente, se determinará si no hay evidencia de carencia de ajuste del modelo, es decir si la función de regresión representa apropiadamente a la media de la variable respuesta.



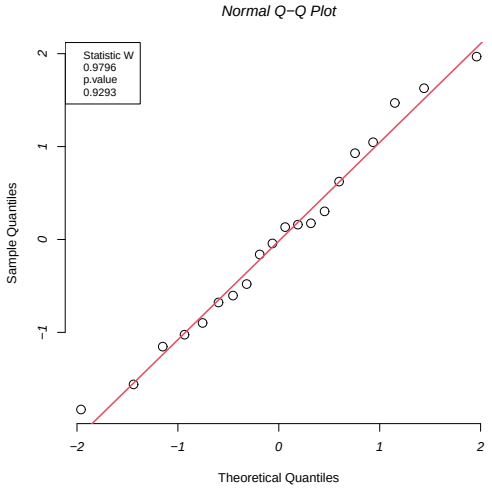


Figura 2.6: Residuos en el problema de destilación. (a) Gráfico de $\widehat{E}_i$ vs. $\widehat{y}_i$; (b) gráfico de $\widehat{E}_i$ vs. x_i, (c) gráfico de probabilidad construido con los $\widehat{E}_i$ resultados para Test Shapiro Wilk aparecen en la leyenda de esta gráfica. Estos gráficos serán usados para evaluar la validez de supuestos sobre los errores E_i. También para identificar observaciones conocidas como atípicas u outliers.

Con la gráfica de probabilidad normal y el test Shapiro Wilk se probará si no hay evidencia en contra del supuesto $E_i \sim N(0, \sigma^2)$, asumiendo que es válido el supuesto de independencia[3].

[3]Todos los tests de bondad de ajuste distribucional han sido desarrollados bajo el supuesto de que las observaciones con las cuales opera el test, provienen de una muestra aleatoria de la distribución supuesta. Recuerde la definición de muestra aleatoria: el conjunto de variables $X_1, X_2, \ldots, X_n$ es una muestra aleatoria de una distribución $F(x)$ si y sólo si son independientes e idénticamente distribuidas, con distribución $F(x)$.

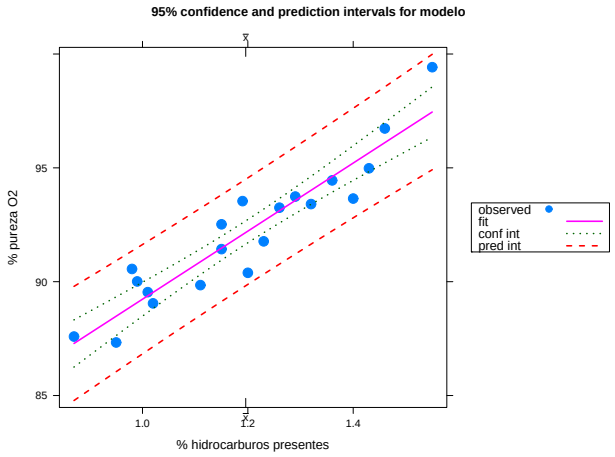


Figura 2.7: Gráfico de dispersión con recta ajustada, intervalos de confianza para la media del % Pureza O_2 e intervalos de predicción para % Pureza O_2.

Con la ecuación de regresión ajustada se estima la media del % de pureza del O_2 y sus límites de confianza del 95 %, para un nivel de % de hidrocarburos presentes en el condensador de 0.80 % y de 0.92 %,

Tabla 2.12: Estimación media e I.C del 95 %. Resultados R son como en Salida 2.4

% hidrocarb (x_0)	$\widehat{\mu}_{Y\|x_0}$	lim. Inf	Lim. Sup
0.80	86.24	85.03	87.45
0.92	88.03	87.12	88.95

Salida R 2.4. *Salida R que origina* `predict( )` *para predicción puntual e intervalos de confianza para la respuesta media en* $x_0 = 0.8$ *y* $x_0 = 0.92$.

```
> predict(modelo,newdata=data.frame(Porc.Hidrocarb=c(0.8,0.92)),level=0.95,
          interval="confidence")
      fit      lwr      upr
1 86.2413 85.03272 87.44987
2 88.0350 87.11657 88.95343
```

Se calcula la predicción del % de pureza del O_2y el intervalo de predicción del 95 %, para un nivel de % de hidrocarburos presentes en el condensador de 0.80 % y de 0.92 %.

Tabla 2.13: Predicción puntual e I.P del 95 %. Resultados R son como en Salida 2.5

% hidrocarb (x_0)	$\widehat{Y}_0$	lim. Inf	Lim. Sup
0.80	86.24	83.66	88.82
0.92	88.03	85.57	90.50

Salida R 2.5. *Salida R que origina* `predict( )` *para predicción puntual e intervalos de predicción de la respuesta futura en* $x_0 = 0.8$ *y* $x_0 = 0.92$.

```
> predict(modelo,newdata=data.frame(Porc.Hidrocarb=c(0.8,0.92)),
          interval="prediction",level=0.95)
      fit      lwr      upr
1 86.2413 83.65839 88.82421
2 88.0350 85.57445 90.49554
```

Código R 2.3. *A continuación, se presenta el código R usado para la obtención de resultados de este problema.*

```
rm(list=ls(all=TRUE))
#Lectura de datos ingresando por teclado. #columna1 Porcentaje de hidrocarburos,
#columna2 Pureza O2
datos=data.frame(scan(what=list(Porc.Hidrocarb=0,Porc.PurezaO2=0)))
0.99 90.01
1.02 89.05
1.15 91.43
1.29 93.74
1.46 96.73
1.36 94.45
0.87 87.59
1.23 91.77
1.55 99.42
1.40 93.65
1.19 93.54
1.15 92.52
0.98 90.56
1.01 89.54
1.11 89.85
1.20 90.39
1.26 93.25
1.32 93.41
1.43 94.98
0.95 87.33

datos          #dé un vistazo para verificar que se leyó correctamente los datos
attach(datos)  #Disponibilizando las variables guardadas en data.frame "datos"

#Gráfico de dispersión con curva LOESS
plot(Porc.Hidrocarb,Porc.PurezaO2,
      main="Gráfico de Dispersión problema sobre proceso de destilación",
      xlab="% hidrocarburos presentes",ylab="% pureza O2",cex=1.5,bty="n",
      font=3,font.main=4)
```

```
lines(loess.smooth(Porc.Hidrocarb,Porc.PurezaO2,family="gaussian",span=0.8),
      lty=1,lwd=2,col="red")
legend("topleft",legend="Curva de ajuste LOESS",col=2,lwd=2)

#Ajustando MRLS
modelo=lm(Porc.PurezaO2~Porc.Hidrocarb)
anova(modelo)   #Obteniendo ANOVA
summary(modelo) #Obteniendo tabla de parámetros ajustados

#Graficas de residuales
win.graph()
plot(fitted(modelo),residuals(modelo),
     ylim=c(min(residuals(modelo),-2*summary(modelo)$sigma),
            max(residuals(modelo),2*summary(modelo)$sigma)),
     xlab="% pureza O2 estimada",ylab="Residuales",
     main="Residuales vs. valores ajustados")

abline(h=c(-2*summary(modelo)$sigma,0,2*summary(modelo)$sigma),lty=2,col=2)

win.graph()
plot(Porc.Hidrocarb,residuals(modelo),
      ylim=c(min(residuals(modelo),-2*summary(modelo)$sigma),
             max(residuals(modelo),2*summary(modelo)$sigma)),
      xlab="% hidrocarburos presentes",ylab="Residuales",
      main="Residuales vs. Predictor")

abline(h=c(-2*summary(modelo)$sigma,0,2*summary(modelo)$sigma),lty=2,col=2)

test=shapiro.test(residuals(modelo));test #Test de normalidad sobre residuales
#Gráfico de normalidad con información del test Shapiro
win.graph()
qqnorm(residuals(modelo),cex=1.5,bty="n",font=3,font.main=3)
qqline(residuals(modelo),lty=1,lwd=2,col=2)
legend("topleft",legend=rbind(c("Statistic W","p.value"),
                              round(c(test$statistic,test$p.value),digits=4)),cex=0.8)

#Gráfico de dispersión con recta ajustada
#intervalos de confianza para la respuesta media
#y de predicción para la respuesta
library(HH)
win.graph(width=8,height=6)
ci.plot(modelo,xlab="% hidrocarburos presentes",ylab="% pureza O2",lty=c(2,1,3),
        cex=1.5,conf.level=.95)

#Para obtener predicciones en x=0.80, 0.92, I.C y I.P del 95%
#Respuesta media estimada  e I.C del 95%
```

```
predict(modelo,newdata=data.frame(Porc.Hidrocarb=c(0.8,0.92)),level=0.95,
        interval="confidence")

#Predicciones e I.P del 95%
predict(modelo,newdata=data.frame(Porc.Hidrocarb=c(0.8,0.92)),
        interval="prediction",level=0.95)

detach(datos)
```

2.13.2. Relación lineal en un proceso de grabado de semiconductores

El grabado con plasma es esencial para la transferencia de figuras de líneas finas en los procesos de fabricación de semiconductores. En un experimento aleatorio se obtuvieron los siguientes datos sobre el flujo de cloro (variable X, en SCMM) en el mecanismo grabador y la rapidez de grabado (Variable Y, en 100A/min). Los datos obtenidos se muestran a continuación.

Tabla 2.14: Datos del proceso de grabado de semiconductores

x (SCMM)	y (100A/min)	x (SCMM)	y (100A/min)
1.5	23.0	1.5	24.5
2.0	25.0	2.5	30.0
2.5	33.5	3.0	40.0
3.5	40.5	3.5	47.0
4.0	49.0		

Vea el gráfico de dispersión en la Figura 2.8. Vea ANOVA en Tabla 2.15 y parámetros estimados del modelo lineal en 2.16. Los residuos se muestran en las Figuras 2.9(a) y 2.9(b); la Figura 2.9(c) presenta el gráfico de probabilidad normal sobre los residuos y en la Figura 2.10 el gráfico de la recta ajustada, con límites de confianza y de predicción. Analice los resultados de ajuste.

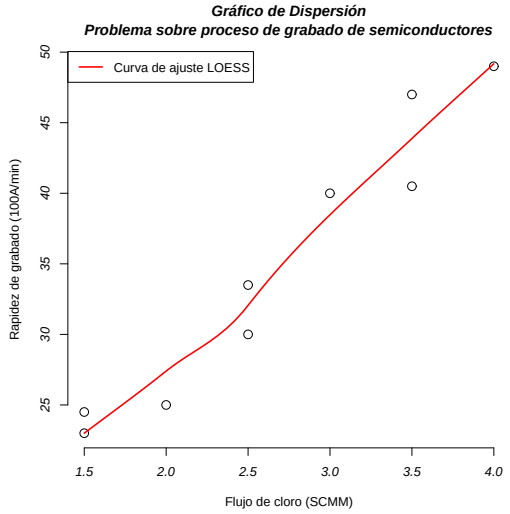


Figura 2.8: Dispersión de los pares de datos y ajuste LOESS datos proceso de grabado de semiconductores

Tabla 2.15: ANOVA en datos proceso de grabado de semiconductores. Compare lo resaltado en color con lo resaltado en Salida R 2.6

Fuente	Df	Sum Sq	Mean Sq	F_0	$P\left(f_{1,7} > F_0\right)$
Flujo de cloro	1	730.69	730.69	112.76	1.438×10^{-5}
Error	7	45.36	6.48		

Salida R 2.6. *Salida R originada por* `anova(modelo)`

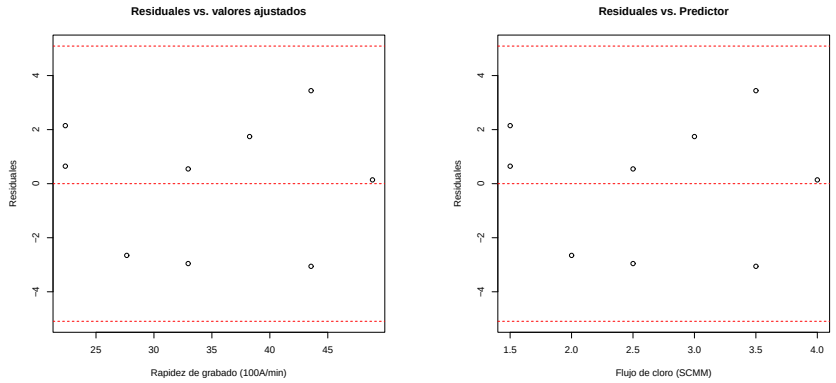
```
> anova(modelo) #Obteniendo ANOVA
Analysis of Variance Table

Response: Rapidez.Grabado
            Df Sum Sq Mean Sq F value    Pr(>F)
Flujo.Cloro  1 730.69  730.69  112.76 1.438e-05 ***
Residuals    7  45.36    6.48
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Tabla 2.16: Tabla de Parámetros Estimados. Compare lo resaltado en color con lo resaltado en Salida R 2.7

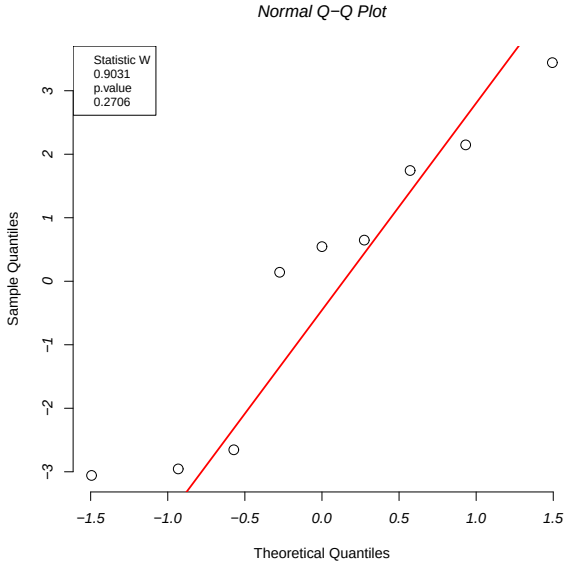
Parámetro	Estimación	Std. Error	T_0	$P\left(\lvert t_7\rvert > \lvert T_0\rvert\right)$
β_0	6.4487	2.7946	2.308	0.0544
β_1	10.6026	0.9985	10.619	1.44×10^{-5}
$\sqrt{\text{MSE}} = 2.546$, $R^2 = 0.9415$, $R^2_{adj} = 0.9332$, $F_0 = 112.8$, VP $= P(f_{1,7} > F_0) = 1.438 \times 10^{-5}$.				
La ecuación ajustada es $\widehat{y} = 6.4487 + 10.6026x$.				

Salida R 2.7. *Salida R originada por* `summary(modelo)`

```
> summary(modelo) #Obteniendo tabla de parámetros ajustados
Call:
lm(formula = Rapidez.Grabado ~ Flujo.Cloro)

Residuals:
    Min      1Q  Median      3Q     Max
-3.0577 -2.6538  0.5449  1.7436  3.4423

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)   6.4487     2.7946   2.308   0.0544 .
Flujo.Cloro  10.6026     0.9985  10.619 1.44e-05 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 2.546 on 7 degrees of freedom
Multiple R-squared:  0.9415,    Adjusted R-squared:  0.9332
F-statistic: 112.8 on 1 and 7 DF,  p-value: 1.438e-05
```



(a) (b)



(c)

Figura 2.9: Residuos en el problema proceso de grabado de semiconductores. (a) $\widehat{E}_i$ vs. $\widehat{y}_i$; (b) $\widehat{E}_i$ vs. x_i; (c) gráfico de probabilidad normal con resultados test Shapiro Wilk en la leyenda.

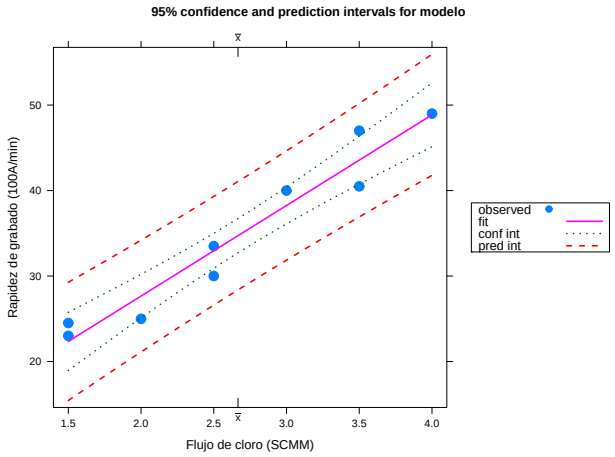


Figura 2.10: Gráfico de dispersión con recta ajustada, I.C para la media e I.P para la rapidez de grabado.

Código R 2.4. *A continuación, se presenta el código R usado para la obtención de resultados de este problema.*

```
rm(list=ls(all=TRUE))
#Lectura de datos ingresando por teclado
#columna 1: Flujo de cloro, columna 2: Rapidez de grabado
datos=data.frame(scan(what=list(Flujo.Cloro=0,Rapidez.Grabado=0)))
1.5 23.0
2.0 25.0
2.5 33.5
3.5 40.5
4.0 49.0
1.5 24.5
2.5 30.0
3.0 40.0
3.5 47.0

datos #Verifique que los datos se han ingresado correctamente
attach(datos)

#Gráfico de dispersión con curva loess adicionada
plot(Flujo.Cloro,Rapidez.Grabado,main="Gráfico de Dispersión
     Problema sobre proceso de grabado de semiconductores",
      xlab="Flujo de cloro (SCMM)",ylab="Rapidez de grabado (100A/min)",cex=1.5,
      bty="n",font=3,font.main=4)

lines(loess.smooth(Flujo.Cloro,Rapidez.Grabado,family="gaussian",span=0.8),
       lty=1,lwd=2,col="red")
legend("topleft",legend="Curva de ajuste LOESS",col=2,lwd=2) #Leyenda del gráfico
```

```
#Ajustando MRLS
modelo=lm(Rapidez.Grabado~Flujo.Cloro)
anova(modelo)    #Obteniendo ANOVA
summary(modelo) #Obteniendo tabla de parámetros ajustados

#Graficas residuales
win.graph()
plot(fitted(modelo),residuals(modelo),
      ylim=c(min(residuals(modelo),-2*summary(modelo)$sigma),
             max(residuals(modelo),2*summary(modelo)$sigma)),
      xlab="Rapidez de grabado (100A/min)",ylab="Residuales",
      main="Residuales vs. valores ajustados")
abline(h=c(-2*summary(modelo)$sigma,0,2*summary(modelo)$sigma),lty=2,col=2)

win.graph()
plot(Flujo.Cloro,residuals(modelo),
     ylim=c(min(residuals(modelo),-2*summary(modelo)$sigma),
            max(residuals(modelo),2*summary(modelo)$sigma)),
     xlab="Flujo de cloro (SCMM)",ylab="Residuales",
     main="Residuales vs. Predictor")
abline(h=c(-2*summary(modelo)$sigma,0,2*summary(modelo)$sigma),lty=2,col=2)

test=shapiro.test(residuals(modelo)) #Test de normalidad usando residuales

#Gráfico de normalidad con información del test Shapiro
win.graph()
qqnorm(residuals(modelo),cex=1.5,bty="n",font=3,font.main=3)
qqline(residuals(modelo),lty=1,lwd=2,col=2)
legend("topleft",legend=rbind(c("Statistic W","p.value"),
                              round(c(test$statistic,test$p.value),digits=4)),cex=0.8)

#Gráfico de dispersión con recta ajustada
#e intervalos de confianza para la respuesta media
#y de predicción para la respuesta
library(HH)
win.graph(width=8,height=6)
ci.plot(modelo,xlab="Flujo de cloro (SCMM)",
        ylab="Rapidez de grabado (100A/min)",lty=c(2,1,3),cex=1.5,conf.level=.95)

detach(datos)
```

Bibliografía

Fox, J. and Weisberg, S. (2019). *An R Companion to Applied Regression*, 3rd ed. Sage, Thousand Oaks CA.

Kutner, M. H., Nachtsheim, C. J., Neter, J. and Li. W. (2005). *Applied Linear Statistical Models*, 5th ed. McGraw-Hill Irwing, New York.

Montgomery, D. C., Peck, E. A. and Vining, G. G. (2012). *Introduction to Linear Regression Analysis*, 5th ed. Wiley, New Jersey.