MODELOS DE REGRESIÓN

Prof. Susana Martín Fernández

REGRESIÓN SIMPLE

Objetivo

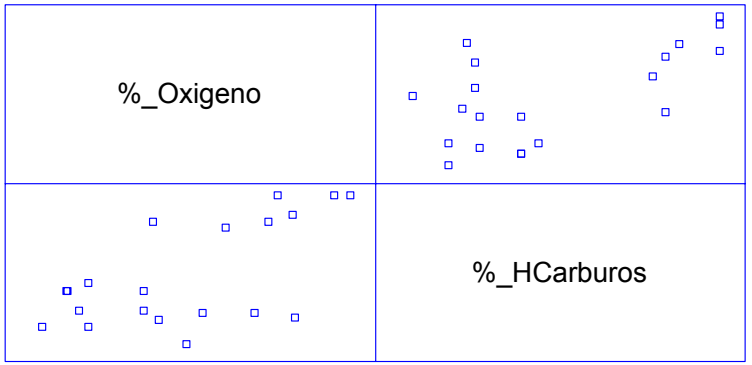
- Sean x_1 x_2 ...x_n, *n* valores de la variable numérica *X.* Sea Y=(y_1, y_2,..., y_n) un vector aleatorio de *n* variables aleatorias independientes, **el modelo de regresión estudia la dependencia lineal del vector *Y*, respecto a la variable *X*.**
- Cuando el conocimiento de una variable determina totalmente el valor de otra habrá una relación *funcional* entre ambas variables.
- Si el conocimiento de una variable no aporta información sobre el valor de otra, ambas variables son *independientes.*
- En general el conocimiento de una variable predice en mayor o menor grado el conocimiento de otra. Se dice que entre ellas hay una relación *estocástica.*

Metodología

1. Representación gráfica de los datos.
2. Planteamiento del modelo.
3. Estimación de los parámetros.
4. Contraste de simplificación del modelo.
5. Comprobación de las hipótesis básicas por análisis de residuos.
6. Análisis del Coeficiente de Correlación
7. Validación del modelo
8. Aplicación del modelo

Representación gráfica de los datos.

Ej. En una planta de producción de oxigeno, se cree que la pureza del oxígeno producido con un proceso de fraccionamiento está relacionada con el porcentaje de hidrocarburos en el condensador principal de la unidad de procesamiento.



Planteamiento del modelo.

$$Y_i = \alpha_0 + \alpha_1 x_i + \varepsilon_i \quad \forall i \in [1,n]$$

Las variables del vector $\varepsilon=(\varepsilon_1,\varepsilon_2,...\varepsilon_n)$ representan la perturbación aleatoria, y se asume que cumplen que:

- Son independientes
- Siguen un distribución normal
- Todas tienen la misma varianza homocedasticidad, σ^2.
- $E[\varepsilon_i] = 0$

Planteamiento del modelo.

De forma matricial el modelo quedaría de la siguiente forma :

$$Y = \beta X' + \varepsilon$$

Donde:

$$\beta = \begin{pmatrix} \alpha_0 & \alpha_1 \end{pmatrix}$$

$$X = \begin{pmatrix} 1 & x_1 \\ 1 & x_2 \\ \dots & \dots \\ 1 & x_n \end{pmatrix}$$

Planteamiento del modelo.

En el ejemplo, para una muestra concreta sería:

$$Y = \beta X' + \varepsilon$$

$$\begin{pmatrix} 86.91 \\ 89.85 \\ 90.28 \\ 86.34 \\ 92.58 \\ 87.33 \\ 86.29 \\ 91.86 \\ 95.61 \\ 89.86 \\ 96.73 \\ 99.42 \\ 98.66 \\ 96.07 \\ 93.65 \\ 87.31 \\ 95 \\ 96.85 \\ 85.2 \\ 90.56 \end{pmatrix} = (\alpha_0, \alpha_1) \begin{pmatrix} 1 & 1.02 \\ 1 & 1.11 \\ 1 & 1.43 \\ 1 & 1.11 \\ 1 & 1.01 \\ 1 & 0.95 \\ 1 & 1.11 \\ 1 & 0.87 \\ 1 & 1.43 \\ 1 & 1.02 \\ 1 & 1.46 \\ 1 & 1.55 \\ 1 & 1.55 \\ 1 & 1.55 \\ 1 & 1.4 \\ 1 & 1.15 \\ 1 & 1.01 \\ 1 & 0.99 \\ 1 & 0.95 \\ 1 & 0.98 \end{pmatrix}' + (\varepsilon_1, \varepsilon_2, \ldots \varepsilon_{20})$$

Estimación de los parámetros.

La función de verosimilitud para los parámetros α_0, α_1, σ^2 , es la siguiente:

$$f(Y,\alpha_0,\alpha_1,\sigma^2) = \frac{1}{(2\pi)^{n/2}\sigma^n} e^{-\frac{1}{2\sigma^2}\sum_{i=1}^{n}(y_i - \alpha_0 - \alpha_1 x_i)^2}$$

Se calculan los valores de α_0, α_1, que hacen máxima la función de verosimilitud:

$$\hat{\alpha}_0 = \sum_{i=1}^{n}\frac{y_i}{n} - \hat{\alpha}_i\bar{x} = \bar{y} - \hat{\alpha}_i\bar{x}$$

$$\hat{\alpha}_1 = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^{n}(x_i - \bar{x})^2} = \frac{Cov(Y,X)}{\sigma^2_x}$$

Estimación de los parámetros.

El valor de σ^2 que hace máxima la función de verosimilitud es:

$$\hat{\sigma}^2 = \frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{\alpha}_0 - \hat{\alpha}_1 x_i)^2 = \frac{1}{n}\sum e_i^2$$

El estimador resultante de la varianza o *varianza residual* es:

$$\hat{S}_R^{\,2} = \frac{1}{n-2}\sum_{i=1}^{n} e_i^{\,2}$$

Los residuos tienen que cumplir dos restricciones que proceden del cálculo de los estimadores de máxima verosimilitud:

$$\sum e_i = 0 \qquad \sum e_i\, x_i = 0$$

Estimación de los parámetros

Análisis de Regresión - Modelo Lineal Y = a + b*X

Variable dependiente: %_Oxigeno
Variable independiente: %_HCarburos

Parámetro	Estimación	Error estándar	Estadístico T	P-Valor
Ordenada	77.8633	4.19889	18.5438	0.0000
Pendiente	11.801	3.48512	3.38612	0.0033

Estimación α_0

Estimación α_1

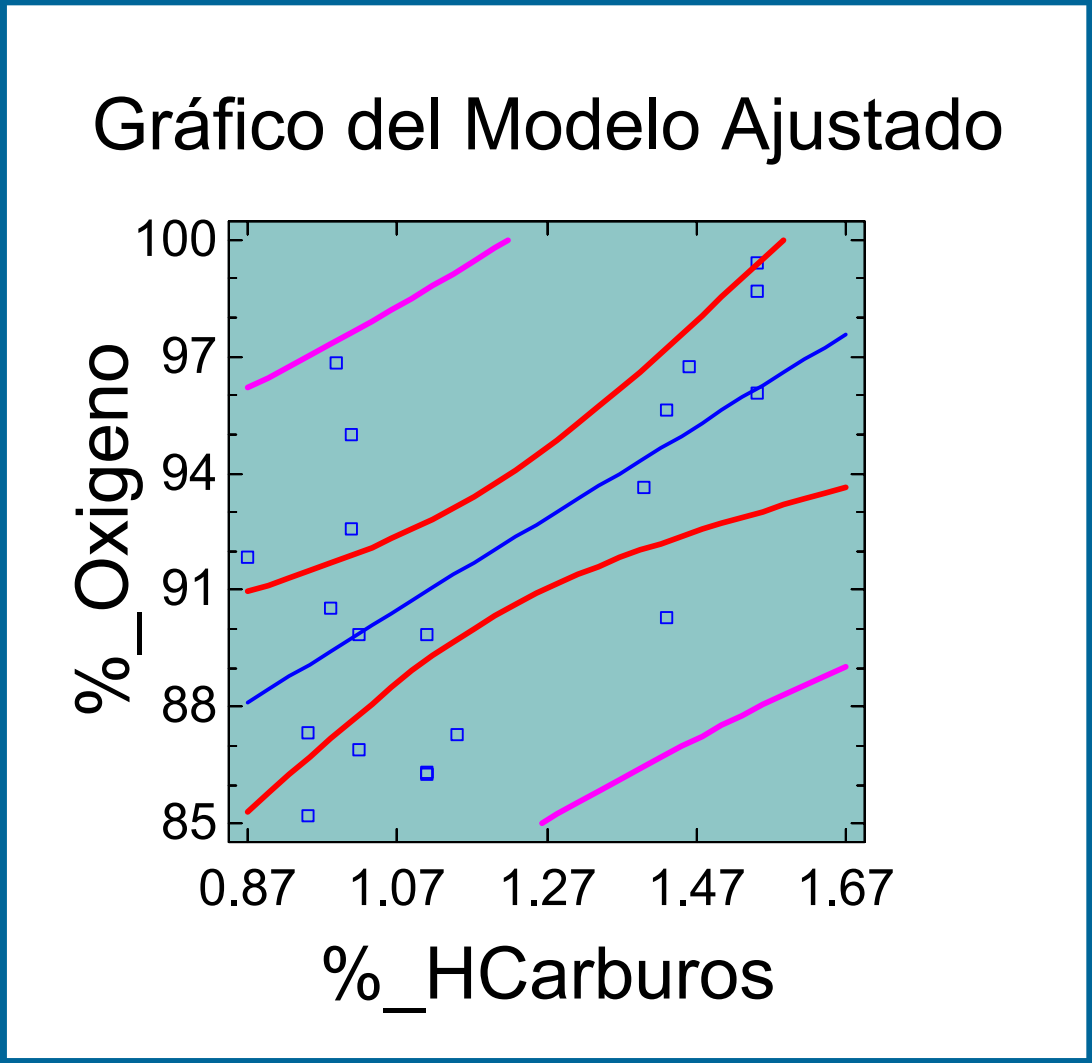
Análisis de la Varianza

Fuente	Suma de cuadrados	GL	Cuadrado medio	Cociente-F	P-Valor
Modelo	148.313	1	148.313	11.47	0.0033
Residuo	232.834	18	12.9352		
Total (Corr.)	381.147	19			

Coeficiente de Correlación = 0.623797
R-cuadrado = 38.9122 porcentaje
R-cuadrado (ajustado para g.l.) = 35.5185 porcentaje
Error estándar de est. = 3.59656
Error absoluto medio = 2.84593
Estadístico de Durbin-Watson = 1.91084 (P=0.3683)
Autocorrelación residual en Lag 1 = 0.0226275

Estimación σ

Estimación de los parámetros




```
%_Oxigeno = 77.8633 + 11.801*%_HCarburos
```

Simplificación del Modelo

Los contrastes de simplificación del modelo son los siguientes:

1. El modelo no es lineal: H_0: $\alpha_1=0$

2. El término independiente es cero: H_0: $\alpha_0=0$

Simplificación del Modelo

El modelo no es lineal: $H_0: \alpha_1=0$

Bajo la hipótesis nula, los estimadores de los parámetros son:

$$\hat{\alpha}_{00} = \bar{y} \qquad \hat{\alpha}_{10} = 0 \qquad \hat{\sigma}_0^2 = \frac{1}{n}\sum_{i=1}^{n}(y_i - \bar{y})^2$$

$$F = \frac{(Y - \breve{\beta}_0 X')(Y - \breve{\beta}_0 X')' - (Y - \hat{\beta} X')(Y - \hat{\beta} X')'}{(Y - \hat{\beta} X')(Y - \hat{\beta} X')'}$$

$$F = \frac{\sum_{i=1}^{n}(y_i - \bar{y})^2 - \sum_{i=1}^{n}(y_i - \bar{y} + \hat{\alpha}_1\bar{x} - \hat{\alpha}_1 x_i)^2}{\sum_{i=1}^{n}(y_i - \bar{y} + \hat{\alpha}_1\bar{x} - \hat{\alpha}_1 x_i)^2} = \frac{\hat{\alpha}_1^2\sum_{i=1}^{n}(x_i - \bar{x})^2}{\sum_{i=1}^{n}(y_i - \bar{y} + \hat{\alpha}_1\bar{x} - \hat{\alpha}_1 x_i)^2}$$

El estadístico *(n-2)/1 F* sigue una distribución F-snedecor con *(1, n-2)* grados de libertad

Simplificación del Modelo

El término independiente es 0: $H_0: \alpha_0 = 0$

$$F = \frac{\hat{\alpha}_0^{\,2}\, n \sum_{i=1}^{n} (x_i - \bar{x})^2 \Big/ \sum_{i=1}^{n} x_i^{\,2}}{\sum_{i=1}^{n} (y_i - \bar{y} + \hat{\alpha}_1 \bar{x} - \hat{\alpha}_1 x_i)^2}$$

El estadístico *(n-2)/1 F* sigue una distribución F-snedecor con *(1, n-2)* grados de libertad

Simplificación del Modelo

Análisis de Regresión - Modelo Lineal Y = a + b*X

Variable dependiente: %_Oxigeno
Variable independiente: %_HCarburos

Parámetro	Estimación	Error estándar	Estadístico T	P-Valor
Ordenada	77.8633	4.19889	18.5438	0.0000
Pendiente	11.801	3.48512	3.38612	0.0033

Test $\alpha_0=0$

Test $\alpha_1=0$

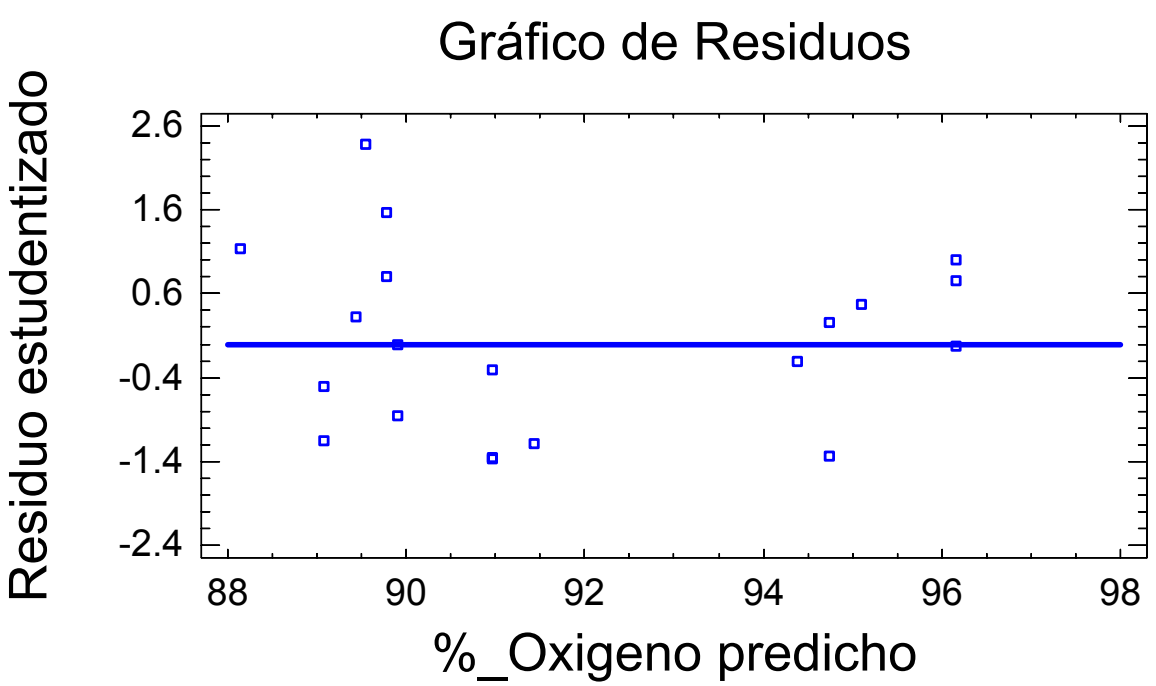
Análisis de la Varianza

Fuente	Suma de cuadrados	GL	Cuadrado medio	Cociente-F	P-Valor
Modelo	148.313	1	148.313	11.47	0.0033
Residuo	232.834	18	12.9352		
Total (Corr.)	381.147	19			

Coeficiente de Correlación = 0.623797
R-cuadrado = 38.9122 porcentaje
R-cuadrado (ajustado para g.l.) = 35.5185 porcentaje
Error estándar de est. = 3.59656
Error absoluto medio = 2.84593
Estadístico de Durbin-Watson = 1.91084 (P=0.3683)
Autocorrelación residual en Lag 1 = 0.0226275

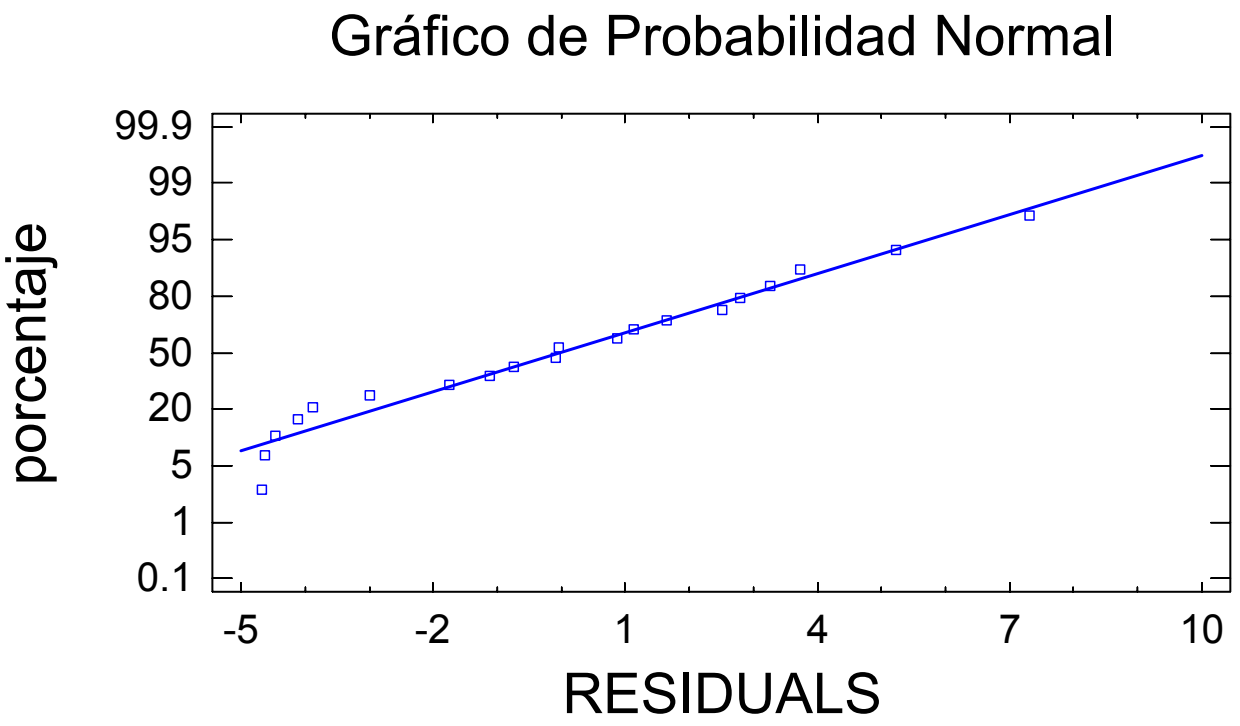
Comprobación Hipótesis Básicas de los Residuos

```
Estadístico de Durbin-Watson = 1.91084 (P=0.3683)
Autocorrelación residual en Lag 1 = 0.0226275
```



Este gráfico es muestra la heteroscedasticidad de los residuos. Su variabilidad cambia
al aumentar los valores de la variable dependiente.

Comprobación Hipótesis Básicas de los Residuos



Comprobación Hipótesis Básicas de los Residuos

Tests de Bondad de Ajuste para RESIDUALS

Contraste Chi-cuadrado

Límite Inferior	Límite Superior	Frecuencia Observada	Frecuencia Esperada	Ch
menor o igual	-3.3866	5	3.33	
-3.3866	-1.50783	2	3.33	
-1.50783	-4.4E-7	4	3.33	
-4.4E-7	1.50783	2	3.33	
1.50783	3.3866	4	3.33	
mayor 3.3866		3	3.33	

Chi-cuadrado = 2.20003 con 3 g.l. P-Valor = 0.531938

Estadístico DMAS de Kolmogorov = 0.115795
Estadístico DMENOS de Kolmogorov = 0.0909808
Estadístico DN global de Kolmogorov = 0.115795
P-Valor aproximado = 0.951365

Se acepta normalidad

Análisis del Coeficiente de Correlación

El coeficiente de correlación mide la relación lineal existente entre dos variables.

$$\rho = \frac{cov(Y,X)}{S_y S_X}$$

Su valor varía entre -1 y 1.

Si ρ=0, no existe relación lineal. Si las variables son normales, además son independientes.

La dependencia entre las variables es completa cuando ρ=±1

Análisis del Coeficiente de Correlación

Contrastes de hipótesis sobre el coeficiente de correlación:

1. H_0: $\rho=0$ frente a H_1: $\rho\neq 0$

Estadístico: $t = \sqrt{n-2}\,\dfrac{r}{\sqrt{1-r^2}} \longrightarrow t_{n-2}$

2. H_0: $\rho=\rho_0 \neq 0$ frente a H_1: $\rho\neq\rho_0$

Estadístico: $z = \dfrac{1}{2}\ln\left(\dfrac{1+r}{1-r}\right) \longrightarrow$ Normal

$$E(z) = \frac{1}{2}\ln\left(\frac{1+\rho_0}{1-\rho_0}\right) + \frac{\rho_0}{2(n-1)}$$

$$\operatorname{var}(z) = \frac{1}{n-3}$$

Análisis del Coeficiente de Correlación

Ej.

Correlaciones

	%_HCarburos	%_Oxigeno
%_HCarburos		0.6238 (20) 0.0033
%_Oxigeno	0.6238 (20) 0.0033	

Validación del modelo

- Análisis de la Varianza
- Test de Falta de Ajuste
- Detección de Residuos Atípicos
- Determinación de Puntos Influyentes

Validación del modelo- Análisis de la Varianza

Análisis de la Varianza

Fuente	Suma de cuadrados	GL	Cuadrado medio	Cociente-F	P-Valor
Modelo	148.313	1	148.313	11.47	0.0033
Residuo	232.834	18	12.9352		
Total (Corr.)	381.147	19			

La hipótesis nula es que el modelo no es válido.

La descomposición de la variabilidad es la siguiente:

$$\sum_{i=1}^{n}(y_i-\bar{y})^2=\sum_{i=1}^{n}(y_i-\hat{y}_i)^2+\sum_{i=1}^{n}(\hat{y}_i-\bar{y})^2$$

Validación del modelo- Test de Falta de Ajuste

Este test comprueba el ajuste de los datos al modelo de regresión lineal.

H_0: La regresión es lineal

Requisitos:

-Normalidad

-Independencia

-Varianza constante.

Observaciones reales duplicadas.

Ej. %_HCarb =1’02 aparece 2 veces.

Validación del modelo- Test de Falta de Ajuste

Análisis de Varianza con Falta de ajuste

Fuente	Suma de cuadrados	GL	Cuadrado medio	Cociente-F	P-Valor
Modelo	148.313	1	148.313	11.47	0.0033
Residuo	232.834	18	12.9352		
Falta de ajuste	194.581	10	19.4581	4.07	0.0292
Error puro	38.2538	8	4.78173		
Total (Corr.)	381.147	19			

El error residual los separa en 2 grupos:

$$\sum_{i=1}^{m}\sum_{j=1}^{n_i}\left(y_{ij}-\hat{y}_i\right)^2=\sum_{i=1}^{m}\sum_{j=1}^{n_i}\left(y_{ij}-\bar{y}_i\right)^2+\sum_{i=1}^{m}\sum_{j=1}^{n_i}\left(\bar{y}_i-\hat{y}_i\right)^2$$

$\bar{y}_i$ Valor medio de las n_i observaciones en x_i

Validación del modelo- Detección de Residuos Atípicos

- Valor atípico es una observación extrema.
- No son representativos del resto de datos.
- Método de Stefansky (1971) para detectarlos:

$$\frac{|e_i|}{\sqrt{\sum e_i^2}} \geq 2$$

Validación del modelo- Detección de Residuos Atípicos

Los residuos atípicos se pueden deber a:

1. Medición incorrecta
2. Análisis incorrecto
3. Registro incorrecto de datos

} **Se eliminan**

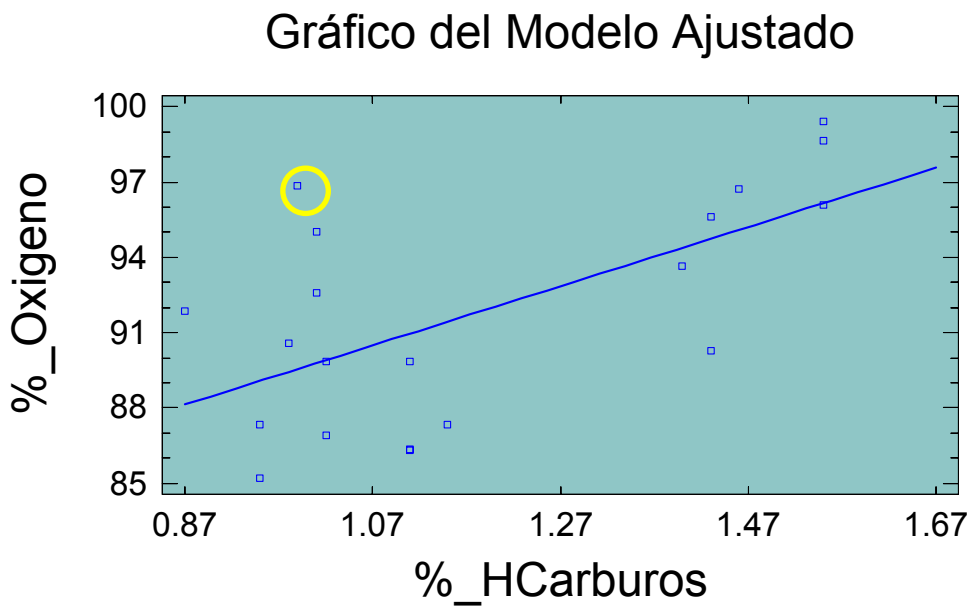
4. Observación extraordinaria factible

} **Permanecen.** Puede controlar propiedades clave del modelo

Validación del modelo- Detección de Residuos Atípicos

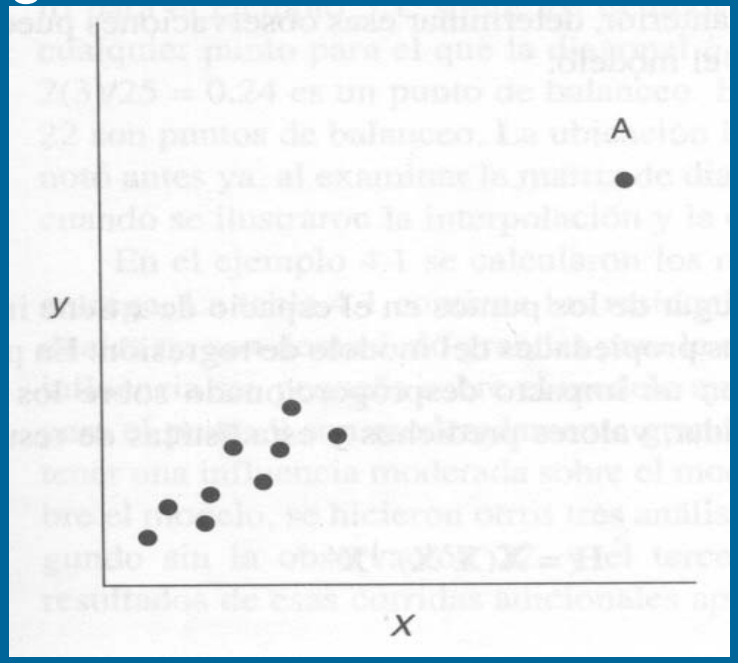
Residuos Atípicos

Fila	X	Y	Y Predicha	Residuo	Residuo Estudentizado
18	0.99	96.85	89.5463	7.3037	2.38

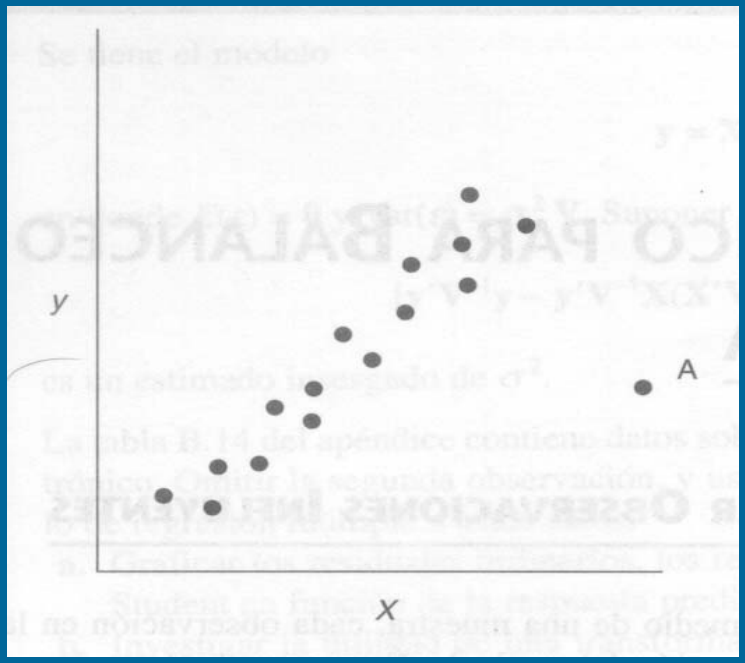


Validación del modelo - Puntos Influyentes

El punto influyente (Outlier) es aquél que tiene influencia sobre los coeficientes de regresión y/o las propiedades del modelo como R^2, y los errores estándar de los coeficientes de regresión...



Pto. de balanceo



Pto. influyente

Validación del modelo - Puntos Influyentes

Balanceo o Leverage (Apalancamiento)

Mide la **influencia** de cada observación en la determinación de los **coeficientes de regresión**.

Se detectan a partir de: $$\hat{y} = X\hat{\beta} = X(X'X)^{-1}X'y = Hy$$

La diagonal de H es una medida de la distancia de la i-ésima observación al centro del espacio X. Hay apalancamiento si este valor es mayor que $$2\sum_{i=1}^{n} h_{ii} / n$$

Validación del modelo - Puntos Influyentes

DFFITS

Este método estudia la influencia de la eliminación de la i-ésima observación sobre la predicción.

$$\text{DFFITS}_i = \frac{\hat{y}_i - \hat{y}_{(i)}}{\sqrt{S^2_{(i)} h_{ii}}} \quad i = 1....n$$

$\hat{y}_{(i)}$ es el valor ajustado de y_i sin utilizar la i-ésima observación

Un punto se analiza si: $|\text{DFFITS}_i| > 2\sqrt{\sum h_{ii}/n}$

Validación del modelo - Puntos Influyentes

DISTANCIA DE MAHALANOBIS

Distancia no euclídea que considera la correlación entre variables.

$$D^2(\mathbf{y}) = (\mathbf{y} - \mathbf{X})\mathbf{S}^{-1}(\mathbf{y} - \mathbf{X})`$$

D es la distancia al cuadrado desde cada punto y al conjunto de variables X
S representa la matriz de covarianzas de X.
X es el vector que contiene los valores medios de las variables independientes.

Transformación de una variable aleatoria

Los modelos de regresión suponen:

1. Los errores tienen media 0, varianza constante y no están correlacionados.
2. Los errores tienen distribución normal.
3. La forma del modelo es correcta.

Si no se cumple alguna de estas suposiciones se pueden TRANSFORMAR LOS DATOS. La transformación se realiza de forma empírica.

Transformación de una variable aleatoria. Estabilización de la varianza.

Relación entre σ^2 y E[y]	Transformación
$\sigma^2 \sim$ Constante	y'=y
$\sigma^2 \sim$ E[y]	$y' = y^{1/2}$ (Raíz cuadrada, datos de Poisson)
$\sigma^2 \sim$ E[y] [1-E[y]]	$y' = 1/\text{sen}(y^{1/2})$,(proporciones binomiales)
$\sigma^2 \sim E[y]^2$	Y'=ln(y) (logarítmica)
$\sigma^2 \sim E[y]^3$	$Y' = y^{-1/2}$ (raíz cuadrdada recíproca)
$\sigma^2 \sim E[y]^4$	$Y' = y^{-1}$ (recíproca)

Transformación de una variable aleatoria. Linealización del Modelo

La no linealidad del modelo se detecta:

1. Con el test de falta de ajuste
2. Con el gráfico de dispersión
3. De forma empírica

Transformación de una variable aleatoria. Método Box-Cox.

- ✓ Se transforma la variable y para corregir la no normalidad y/o la varianza no constante.
- ✓ Es una transformación de potencia y^{λ}
- ✓ Se determinan los parámetros de la recta α_0, α_1 y λ por el método de máxima verosimilitud.

Transformación de una variable aleatoria. Método Box-Cox.

Problema cuando λ=0.

Solución realizando el siguiente cambio:

$$y^{(\lambda)} = \begin{cases} \dfrac{y^{\lambda} - 1}{\lambda\left(\ln^{-1}\left[(1/n)\sum_{i=1}^{n} \ln y_i\right]\right)^{\lambda-1}}, & \lambda \neq 0 \\ \left(\ln^{-1}\left[(1/n)\sum_{i=1}^{n} \ln y_i\right]\right)\ln y, & \lambda = 0 \end{cases}$$

Se ajusta el modelo $y^{(\lambda)}=\beta X'+\varepsilon$

Transformaciones Box-Cox - Poder = 1.5123 Cambio = 0.0

Variable dependiente: %_Oxigeno
Variable independiente: %_HCarburos

Parámetro	Estimación	Error Estándar	Estadístico T	P-Valor
Ordenada	47.6995	4.19789	11.3627	0.0000
Pendiente	11.8709	3.48429	3.40699	0.0031

Análisis de la Varianza

Fuente	Suma de Cuadrados	Gl	Cuadrado Medio	F-Ratio	P-Valor
Modelo	150.075	1	150.075	11.61	0.0031
Residuo	232.723	18	12.9291		
Total (Corr.)	382.799	19			

Coeficiente de Correlación = 0.626137
R-cuadrado = 39.2048 porcentaje
Error Estándar de la Est. = 3.5957

El StatAdvisor

Este procedimiento está diseñado para permitirle comparar el efecto de varias transformaciones de poder de %_Oxigeno en la regresión lineal entre él y %_HCarburos. La ecuación del modelo ajustado, mostrado como una línea continua, es

BoxCox(%_Oxigeno) = 47.6995 + 11.8709*%_HCarburos

donde

BoxCox(%_Oxigeno) = 1 + (%_Oxigeno^1.5123-1)/(1.5123*91.7145^0.512304)

Aplicación del Modelo

- Predicción de nuevas observaciones

$$\hat{y}_0 = \hat{\alpha}_0 + \hat{\alpha}_1 x_0$$

- Intervalos de confianza:
  - De la respuesta media E(y)
  - De nuevas predicciones

Aplicación del Modelo

Intervalos de confianza de la respuesta media, E(y).

Se fija un valor de interés x_0, y se trata de encontrar int. de confianza de $E(y/x_0)$.

Estimador de $E(y/x_0)$: $\hat{E}(y/x_0) = \hat{\alpha}_0 + \hat{\alpha}_1 x_0$

Su varianza es:

$$\text{var}(\hat{E}(y/x_0)) = \text{var}(\hat{\alpha}_0 + \hat{\alpha}_1 x_0) = \text{var}(\bar{y} + \hat{\alpha}_1 (x_0 - \bar{x})) =$$

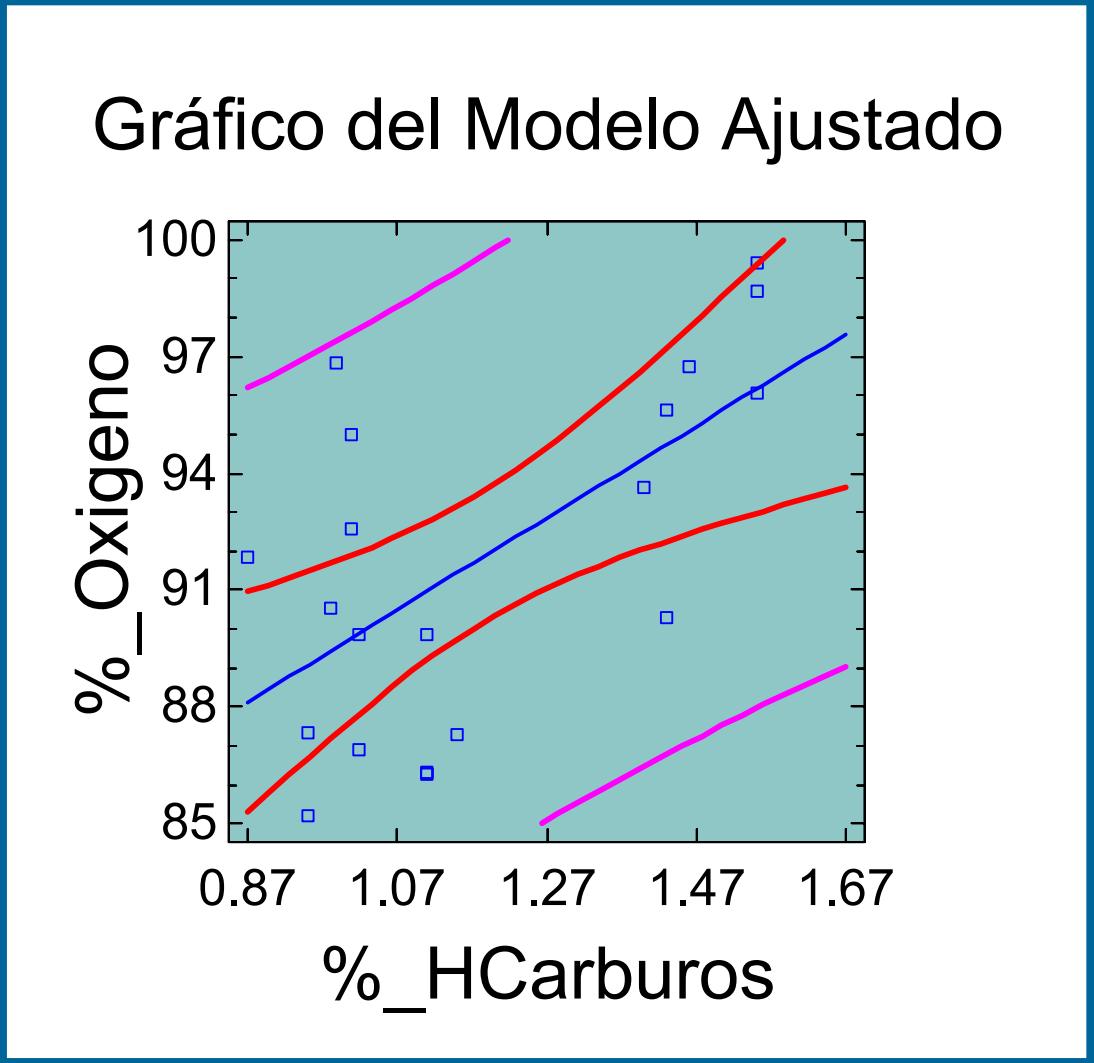
$$= \frac{\sigma^2}{n} + \sigma^2 \frac{(x_0 - \bar{x})^2}{S_{xx}}$$

Aplicación del Modelo

Intervalos de confianza de la respuesta media, E(y), para un nivel de confianza 1-α es:

$$E(y/x_0) \in \left(\hat{\mu}_{y|x_0} \pm t_{\alpha/2,n-2} \sqrt{\frac{S_{res}^2}{n} + S_{res}^2 \frac{(x_0 - \overline{x})^2}{S_{xx}}} \right)$$

Intervalos de Confianza



%_Oxigeno = 77.8633 + 11.801*%_HCarburos

Aplicación del Modelo

Intervalos de confianza de nuevas predicciones

Si el valor de interés de la variable independiente es x_0 entonces , $\hat{y}_0 = \hat{\alpha}_0 + \hat{\alpha}_1 x_0$ es el valor estimado de y_0.

$$\text{var}(y_0 - \hat{y}_0) = \sigma^2 + \frac{\sigma^2}{n} + \sigma^2 \frac{(x_0 - \bar{x})^2}{S_{xx}}$$

Aplicación del Modelo

Intervalos de confianza de nuevas predicciones

Y por tanto el intervalo de confianza, para un nivel de confianza 1-α, es

$$y_0 \in \left(\hat{y}_0 \pm t_{\alpha/2,n-2} \sqrt{S_{res}^2 + \frac{S_{res}^2}{n} + S_{res}^2 \frac{(x_0 - \bar{x})^2}{S_{xx}}} \right)$$

Regresión Múltiple

- El objetivo de la regresión múltiple es construir un modelo probabilístico que relacione un variable dependiente *Y* con dos o más variables matemáticas independientes x_1, x_2, ... x_k. La expresión de dicho modelo es la siguiente:

$$Y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_k x_k + \varepsilon$$

Donde:

- β_i, es el coeficiente que representa el efecto sobre la variable dependiente al aumentar en una unidad el valor de la variable independiente x_i.
- ε, representa la perturbación aleatoria.

- ε, verifica las siguientes hipótesis:
- Su media es 0.
- Su varianza es constante, $\sigma 2$.
- Las perturbaciones son independientes entre sí.
- Siguen una distribución Normal.

- **Estimación de los parámetros.**
- Aplicando el método de mínimos cuadrados, (número de observaciones es *n,)* la función a minimizar es:

$$M = \sum \left(y_i - (\beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + \cdots + \beta_k x_{ki}) \right)^2$$

- Derivando respecto a β_0, se verifica, si se llama $e_i = y_i - \hat{y}_i$, la siguiente ecuación:

$$\sum e_i = 0$$

- Derivando respecto a βi, se verifica:

$$\sum e_i x_{ji} = 0 \qquad j = 1, \cdots k$$

- El sistema de ecuaciones definido por las expresiones anteriores se puede escribir de la siguiente manera:

$$\sum y_i = n\hat{\beta}_0 + \hat{\beta}_1 \sum x_{1i} + \hat{\beta}_2 \sum x_{2i} + \cdots + \hat{\beta}_k \sum x_{ki}$$

$$\sum y_i\, x_{1i} = \hat{\beta}_0 \sum x_{1i} + \hat{\beta}_1 \sum x_{1i}^2 + \hat{\beta}_2 \sum x_{2i}\, x_{1i} + \cdots + \hat{\beta}_k \sum x_{ki}$$

$$\vdots$$

$$\sum y_i\, x_{ki} = \hat{\beta}_0 \sum x_{ki} + \hat{\beta}_1 \sum x_{1i}\, x_{ki} + \hat{\beta}_2 \sum x_{2i}\, x_{ki} + \cdots + \hat{\beta}_k \sum x_{ki}^2$$

- Las ecuaciones anteriores se pueden expresar de forma matricial:

$$\begin{bmatrix} 1 & 1 & \cdots & 1 \\ x_{11} & x_{12} & \cdots & x_{1n} \\ \vdots & \vdots & \vdots & \vdots \\ x_{k1} & x_{k2} & \cdots & x_{kn} \end{bmatrix} \begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{bmatrix} = \begin{bmatrix} 1 & 1 & \cdots & 1 \\ x_{11} & x_{12} & \cdots & x_{1n} \\ \vdots & \vdots & \vdots & \vdots \\ x_{k1} & x_{k2} & \cdots & x_{kn} \end{bmatrix} \begin{bmatrix} 1 & x_{11} & \cdots & x_{k1} \\ 1 & x_{12} & \cdots & x_{k2} \\ \vdots & \vdots & \vdots & \vdots \\ 1 & x_{1n} & \cdots & x_{kn} \end{bmatrix} \begin{bmatrix} \hat{\beta}_0 \\ \hat{\beta}_1 \\ \vdots \\ \hat{\beta}_k \end{bmatrix}$$

De la expresión anterior se puede despejar el valor de los parámetros buscados:

$$\underset{\_}{\hat{\beta}} = \left(X' X \right)^{-1} \quad X' Y$$

- Si la matriz de covarianzas es la siguiente:

$$S = \begin{bmatrix} s_{yy} & s_{yx_1} & \cdots & s_{yx_k} \\ s_{x_1 y} & s_{x_1 x_1} & \cdots & s_{x_1 x_k} \\ \vdots & \vdots & \vdots & \vdots \\ s_{x_k y} & s_{x_k x_1} & \cdots & s_{x_k x_k} \end{bmatrix}$$

- La expresión de cada parámetro βi con *i* de *1 a k,* es la siguiente:

$$\hat{\beta}_i = \frac{-\left|S_{yx_i}\right|}{\left|S_{yy}\right|}$$

- Donde:
- $|S_{yxi}|$ es el determinante del mínimo complementario correspondiente a los órdenes de las variables *y* y x_i. En este caso, estos órdenes serán 1 para la variable *y* e i+1 para la variable x_i.
- El término independiente será:

$$\hat{\beta}_0 = \bar{y} - \sum \hat{\beta}_i \, \bar{x}_i$$

- La varianza de la perturbación aleatoria, σ^2. se estima a partir de la varianza residual, estimador máximo-verosímil en la hipótesis de normalidad. El número de grados de libertad de los residuos es *n-k-1*, por haber *k+1* restricciones:

$$s_R^2 = \frac{\sum e_i^2}{n-k-1}$$

- **Descomposición de la variabilidad**
- La variabilidad de la respuesta puede descomponerse de la siguiente manera:

$$\sum \left(y_i - \bar{y} \right)^2 = \sum \left(\hat{y}_i - \bar{y} \right)^2 + \sum \left(y_i - \hat{y}_i \right)^2$$

que expresa la variación total *VT* como suma de la variación explicada por el modelo *VE* y la residual o no explicada *VNE.*

- El contraste de regresión comprobará que el modelo es válido. La hipótesis nula será la más sencilla y es que el vector de parámetros de regresión sea nulo.

- La tabla ADEVA es la siguiente:

Fuente de variación	**Suma de cuadrados**	**Grados de libertad**	**Varianza**	**Contraste**
VE	$\sum\left(\hat{y}_i - \bar{y}\right)^2$	k	$\hat{S}_e^{\,2}$	$F = \hat{S}_e^2 / \hat{S}_R^2$
VNE	$\sum\left(y_i - \hat{y}_i\right)^2$	n-k-1	$\hat{S}_R^{\,2}$	
VT	$\sum\left(y_i - \bar{y}\right)^2$	n-1	$\hat{S}_y^{\,2}$	

Correlación en Regresión Múltiple

Coeficiente de determinación o de correlación múltiple

Este coeficiente mide la correlación entre la variable dependiente y el conjunto de las variables independientes:

$$R^2 = \frac{VE}{VT}$$

- Inconvenientes:
- Al aumentar el número de variables que intervienen en el modelo, su valor aumenta, aunque el efecto de esta nueva variable no sea significativo.
- Es muy sensible a la elección de variable dependiente. Dos modelos formalmente iguales, pueden tener diferente valor del coeficiente de determinación.
- Se utiliza el coeficiente de determinación corregido,

$$\bar{R}^2 = 1 - \frac{Varianza\,residual}{Varianza\,de\,y}$$

$$\bar{R}^2 = \sqrt{1 - \frac{|S|}{\sigma_y{}^2 |S_{yy}|}}$$

Coeficiente de correlación parcial

Dado un conjunto de variables, x_1, x_2, ... x_k, el coeficiente de correlación parcial entre dos cualesquiera de ellas, es una medida de su relación lineal, cuando se elimina de ambas el efecto debido al resto de las variables.

Por ejemplo si se quiere calcular el coeficiente de correlación parcial entre x_1 y x_2, se calculará primero los hiperplanos de regresión de x_1 respecto a x_3 x_4, ... x_k y de x_2, respecto a x_3 x_4, ... x_k , si llamamos $e_{1.345\ldots k}$ y $e_{2.345\ldots k}$ los residuos de los dos ajustes anteriores, el coeficiente de regresión parcial será:

$$r_{12\,.\,3\cdots k} = \frac{E\left[e_{1.34\ldots k}\ e_{2.34..k}\right]}{\sqrt{v(e_{1.34..k})\,v(e_{2.34..k})}} = \frac{-\left|S_{x_1 x_2}\right|}{\sqrt{\left|S_{x_1 x_1}\right|\left|S_{x_2 x_2}\right|}}$$

- Supongamos que se están estudiando solamente 3 variables x_1, x_2 y x_3., se pueden relacionar los coeficientes de correlación simple y parcial a través de la siguiente expresión:

$$r_{12.3} = \frac{r_{33} r_{12} - r_{13}\, r_{23}}{\sqrt{\left(1 - r_{13}^2\right)\left(1 - r_{23}^2\right)}}$$

- Donde r_{ij} es el coeficiente de correlación simple entre las variables x_i, x_j.