CONTRASTE DE LINEALIDAD

Este contraste sirve para probar si el modelo lineal se ajusta adecuadamente a los datos, y sólo puede realizarse cuando, para cada uno de los k valores x_i de X, se realizan n_i observaciones, $y_{i1}, y_{i2}, \ldots, y_{in_i}$, que se representan en una tabla como la siguiente:

X	x_1	x_1	$\ldots$	x_1	x_2	x_2	$\ldots$	x_2	$\ldots$	$\ldots$	x_k	x_k	$\ldots$	x_k
Y	y_{11}	y_{12}	$\cdots$	y_{1n_1}	y_{21}	y_{22}	$\cdots$	y_{2n_2}	$\cdots$	$\cdots$	y_{k1}	y_{k2}	$\cdots$	y_{kn_k}

La hipótesis nula es: H_0' : la regresión es lineal (El modelo lineal SÍ SE AJUSTA adecuadamente a los datos, y cada $\hat{y}_i$ es un estimador centrado de la media $\mu_{Y|x_i}$);

La hipótesis alternativa es: H_1' : la regresión no es lineal. (El modelo lineal NO SE AJUSTA adecuadamente a los datos)

El contraste de la linealidad de la regresión consiste en descomponer la variabilidad residual del ADEVA de regresión en dos efectos diferentes:

- Un efecto aleatorio, debido a la distribución de Y para cada valor de X.
- Un efecto sistemático, si la verdadera función de regresión no es lineal.

Estos dos efectos sólo pueden estimarse por separado si se realizan n_i observaciones, $y_{i1}, y_{i2}, \ldots, y_{in_i}$, para cada uno de los k valores x_i de X, con un total de $n = \sum_{i=1}^{k} n_i$ pares de valores.

Estadísticos auxiliares para calcular las estimaciones $\hat{\alpha}, \hat{\beta}, \hat{\sigma}_e^2$ y $\tilde{\mu}_{y|x_i}$ a partir de estos k conjuntos de datos, por el método de mínimos cuadrados:

$$T_x = \sum_{i=1}^{k} n_i x_i, \quad T_{i\cdot} = \sum_{j=1}^{n_i} y_{ij}, \quad T_y = \sum_{i=1}^{k} T_{i\cdot}, \quad \bar{x} = \frac{T_x}{n}, \quad \bar{y}_{i\cdot} = \frac{T_{i\cdot}}{n_i}, \quad \bar{y} = \frac{T_y}{n}$$

$$T_{xx} = \sum_{i=1}^{k} n_i x_i^2, \quad T_{yy} = \sum_{i=1}^{k} \sum_{j=1}^{n_i} y_{ij}^2, \quad T_{xy} = \sum_{i=1}^{k} x_i T_{i\cdot}$$

$$S_{xx} = \sum_{i=1}^{k} n_i (x_i - \bar{x})^2 = T_{xx} - \frac{1}{n} T_x^2, \quad S_{yy} = \sum_{i=1}^{k} \sum_{j=1}^{n_i} (y_{ij} - \bar{y})^2 = T_{yy} - \frac{1}{n} T_y^2$$

$$S_{xy} = \sum_{i=1}^{k} (x_i - \bar{x}) \sum_{j=1}^{n_i} (y_{ij} - \bar{y}) = T_{xy} - \frac{1}{n} T_x T_y$$

Las estimaciones de los parámetros de la regresión que se deducen de estos estadísticos son:

$$\hat{\beta} = \frac{S_{xy}}{S_{xx}} \qquad \hat{\alpha} = \bar{y} - \frac{S_{xy}}{S_{xx}} \bar{x} \qquad \hat{\sigma}_e^2 = \frac{S_{yy} - \hat{\beta} S_{xy}}{n - 2}$$

ADEVA DE LINEALIDAD

El contraste de linealidad se realiza mediante un ADEVA DE LINEALIDAD, que está basado en la igualdad

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

$$(\quad SCT \quad = \quad SCX \quad + \quad SCF \quad + \quad SCD \quad)$$

En este ADEVA, la suma de cuadrados residual,

$$SCR=\sum_{i=1}^{k}\sum_{j=1}^{n_i}(y_{ij}-\hat{y}_i)^2$$

se ha dividido en dos sumas de cuadrados independientes:

$$SCF=\sum_{i=1}^{k}\sum_{j=1}^{n_i}(\bar{y}_i-\hat{y}_i)^2 \quad \text{y} \quad SCD=\sum_{i=1}^{k}\sum_{j=1}^{n_i}(y_{ij}-\bar{y}_i)^2$$

con $k-2$ y $n-k$ grados de libertad, respectivamente

SCF indica la falta de ajuste del modelo lineal a la población, midiendo la variabilidad debida al sesgo de las estimaciones de $\mu_{Y|x_i}$ por $\hat{y}_i$

SCD indica la variabilidad aleatoria de las respuestas y_{ij} a un mismo valor x_i respecto de su media muestral $\bar{y}_i$

TABLA DE ADEVA PARA LOS CONTRASTES DE LINEALIDAD Y REGRESIÓN

Fuente de variación	Sumas de cuadrados	Grados de libertad	Medias de cuadrados	valor de F_l
Regresión	$SCX = \hat{\beta} S_{xy}$	1	$MCX = SCX$	$\frac{MCX}{MCR}$
Residual	$SCR = S_{yy} - \hat{\beta} S_{xy}$	$n-2$	$MCR = \frac{SCR}{n-2}$	$---$
Falta de ajuste	$SCF = SCR - SCD$	$k-2$	$MCF = \frac{SCF}{k-2}$	$\frac{MCF}{MCD}$
Error	$SCD = T_{yy} - \sum_i \frac{T_{i\cdot}^2}{n_i}$	$n-k$	$MCD = \frac{SCD}{n-k}$	$---$
Total	$SCT = S_{yy}$	$n-1$	$---$	$---$

Si se acepta la hipótesis nula de linealidad y se rechaza la hipótesis nula de falta de regresión, pueden efectuarse contrastes sobre los parámetros de la recta de regresión y pueden construirse intervalos de confianza para ellos y para las predicciones

El valor $\hat{y}_0 = \hat{\alpha} + \hat{\beta} x_0$ se utiliza de dos formas diferentes:

Para estimar la media $\mu_{Y|x_0} = \alpha + \beta x_0$

Para predecir el valor de Y que se obtendría para un valor x_0 dado.

Ejemplo 2: **Se han preparado 30 cultivos iguales y en cada uno se han introducido 20 colonias de estafilococos del mismo peso; estos cultivos se han distribuído al azar en 6 cubetas, cada una de las cuales se ha mantenido a una temperatura constante durante 48 horas. Terminado ese tiempo, se ha contado el número de colonias en cada cultivo, encontrando los siguientes datos:**

Temperaturas (^oC)	Número de colonias para cada temperatura					
10	37	35	32	41	37	36
15	41	48	41	45	50	
20	55	53	60	51	55	
25	62	58	57	60		
30	62	68	67	70	65	
35	72	68	70	71	69	

Contraste con un ADEVA de linealidad, a un nivel del 10%, si existe relación lineal entre la temperatura y el crecimiento de estas bacterias. En caso afirmativo, estime la recta de regresión; en caso negativo, indique cuál sería la función de regresión más ajustada a estos datos.

En la resolución de este ejemplo utilizaremos las siguientes notaciones:

x_i : cada una de las temperaturas

n_i: número de respuestas al valor x_i

y_{ij} : j - ésima respuesta al valor x_i

$k = 6$: valores distintos de ***X***

$n = 30$: número de valores de ***Y***

$\bar{y}_i$:valor medio de ***Y*** para x_i

ESTADÍSTICOS PREVIOS AL ADEVA DE REGRESIÓN Y LINEALIDAD

x_i	n_i	$x_i n_i$	y_{ij}						$T_{i.}$	$x_i T_{i.}$	$\bar{y}_i$
10	6	60	37	35	32	41	37	36	218	2180	36.33
15	5	75	41	48	41	45	50		225	3375	45.00
20	5	100	55	53	60	51	55		274	5480	54.80
25	4	100	62	58	57	60			237	5925	59.25
30	5	150	62	68	67	70	65		332	9960	66.40
35	5	175	72	68	70	71	69		350	12250	70.00

$\bar{x} = 22 \quad T_x = 660; \qquad T_y = 1636; \quad \bar{y} = 54.533$

$T_x^2/n = 14520 \quad T_y^2/n = 89216.533 \quad T_x T_y/n = 35992 \quad T_{xy} = 39170$

x_i^2	$x_i^2 n_i$	y_{ij}^2						$T_{i.}^2/n_i$
100	600	1369	1225	1024	1681	1369	1296	7920.67
225	1125	1681	2304	1681	2025	2500		10125.00
400	2000	3025	2809	3600	2601	3025		15015.20
625	2500	3844	3364	3249	3600			14042.25
900	4500	3844	4624	4489	4900	4225		22044.80
1225	6125	5184	4624	4900	5041	4761		24500.00

$T_{xx} = 16850 \qquad T_{yy} = 93864 \qquad \sum_i T_{i.}^2/n_i = 93647.92$

$$S_{xx} = T_{xx} - T_x^2/n = 2330;$$

$$S_{yy} = T_{yy} - T_y^2/n = 4647.47$$

$$S_{xy} = T_{xy} - T_x T_y/n = 3178$$

TABLA DE ADEVA PARA LOS CONTRASTES DE LINEALIDAD Y REGRESIÓN

Fuente de variación	Sumas de cuadrados	Gr. de lib.	Medias de cuadrados	valor de F_l
Regresión	$SCX = \hat{\beta} S_{xy} = 4334.63$	1	$MCX = 4334.63$	$\frac{MCX}{MCR} = 388$
Residual	$SCR = S_{yy} - \hat{\beta} S_{xy} = 312.84$	28	$MCR = 11.173$	$---$
Falta de ajuste	$SCF = SCR - SCD = 96.76$	4	$MCF = 24.19$	$\frac{MCF}{MCD} = 2.687$
Error	$SCD = T_{yy} - \sum_i \frac{T_{i.}^2}{n_i} = 216.08$	24	$MCD = 9.003$	$---$
Total	$SCT = S_{yy} = 4647.47$	29	$---$	$---$

Estimaciones de β y α:

$$\hat{\beta} = b_0 = \frac{S_{xy}}{S_{xx}} = \frac{3178}{2330} = 1.364 \qquad \hat{\alpha} = a_0 = \bar{y} - b_0 \bar{x} = 54.533 - 1.364 \times 22 = 24.525$$

Estimación de σ_e^2: $\hat{\sigma}_e^2 = \dfrac{S_{yy} - \hat{\beta} S_{xy}}{n-2} = \dfrac{4647.47 - 4334.63}{28} = \dfrac{312.84}{28} = 11.173 = MCR$

Recta de regresión: $\hat{y} = 24.525 + 1.364\, x$. **Coeficiente de determinación:** $R^2 = 0.933$

En resumen, los contrastes para el modelo lineal son:

Contraste de regresión

H_0 : $\beta = 0$ (No existe regresión lineal)

H_1 : $\beta \neq 0$. (Sí existe regresión lineal)

ADEVA de regresión

En el contexto del ADEVA, en contraste anterior puede enunciarse de la siguiente forma:

H_0 : La variabilidad de Y no está explicada por el modelo lineal

H_1 : Una parte significativa de la variabilidad de Y está explicada con el modelo lineal

Contraste de linealidad (o contraste de Falta de ajuste)

H_0 : El modelo lineal ajusta adecuadamente los datos

H_1 : El modelo lineal NO ajusta adecuadamente los datos

OTRAS FUNCIONES DE AJUSTE

Otros modelos que podemos considerar son aquellos que se reducen a modelos lineales, mediante transformaciones adecuadas de una de las variables o de las dos a la vez. Por ejemplo:

- *Función exponencial,* $y = ab^x$

  Si se toman logaritmos en ambos lados de la igualdad, se transforma en la ecuación

  $$\log(y) = \log(a) + x\log(b)$$

  Si definimos la variable aleatoria $Z = \log(Y)$, los coeficientes de la función exponencial se calculan ajustando una recta de regresión de Z sobre X,en la que hallaríamos los logaritmos de los coeficientes a y b.

- *Función potencial o multiplicativa,* $y = cx^d$,

  Si se toman logaritmos en ambos lados de la igualdad, con lo que se transforma en la ecuación

  $$\log(y) = \log(c) + d\log(x)$$

  Si definimos la variable aleatoria $Z = \log(Y)$ y la variable $W = \log(X)$, los coeficientes de la función potencial se calculan ajustando una recta de regresión de Z sobre W, en la que hallaríamos el coeficiente d y el logaritmo del coeficiente c.

Otros modelos de ajuste son:

Inversa de Y: $Y = \dfrac{1}{a + b\,X} \longrightarrow 1/Y = a + bX$

Inversa de X: $Y = a + \dfrac{b}{X}$

Doble Inversa: $Y = \dfrac{1}{a + b/X}$

Logarítmo de X: $Y = a + b \cdot lnX$

Multiplicativo: $Y = a\,X^b$

Exponencial: $Y = \exp\,(a + bX)$

Raíz Cuadrada de X: $Y = a + b\sqrt{X}$

Raíz Cuadrada de Y: $Y = (a + b\,X)^2$

Curva S : $Y = \exp\,(a + b/X)$