

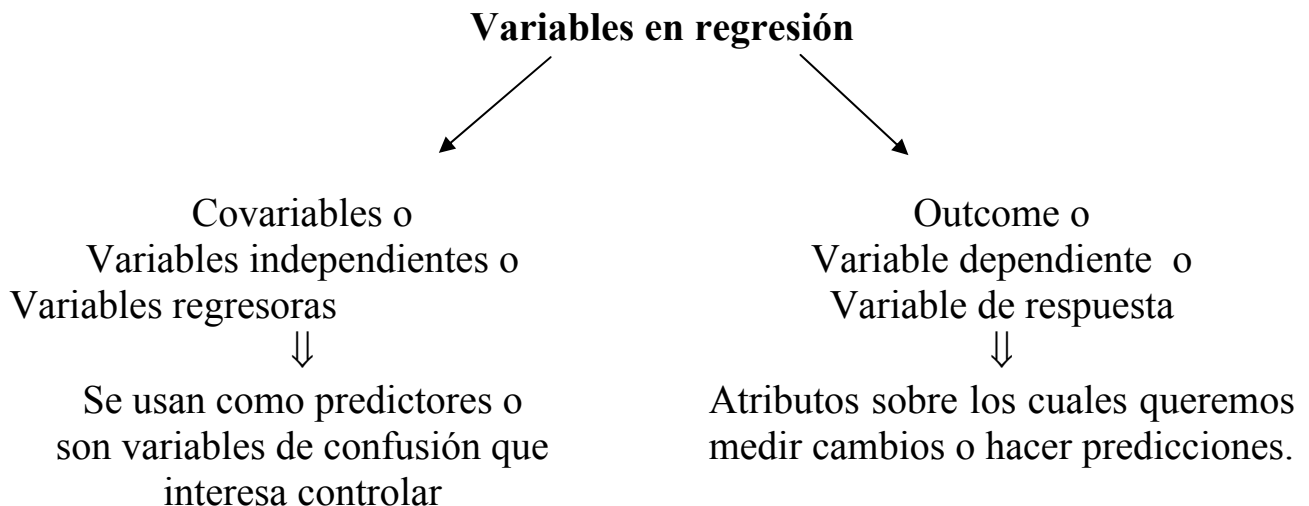
ANÁLISIS DE REGRESIÓN

El **análisis de regresión** involucra el estudio la relación entre dos variables CUANTITATIVAS. En general interesa:

- ✓ Investigar *si existe una asociación* entre las dos variables testeando la hipótesis de independencia estadística.
- ✓ Estudiar la *fuerza de la asociación*, a través de una medida de asociación denominada *coeficiente de correlación*.
- ✓ Estudiar la *forma de la relación*. Usando los datos propondremos un modelo para la relación y a partir de ella será posible predecir el valor de una variable a partir de la otra.

Para ello proponemos un **MODELO** que relaciona una variable dependiente (Y) con una variable independiente (X).

La decisión sobre qué análisis usar en una situación particular, depende de la naturaleza del OUTCOME y del tipo de función que se propone para relacionar el outcome y la variable independiente.



MODELOS

Llamaremos **MODELO MATEMÁTICO** a la función matemática que proponemos como forma de relación entre la variable dependiente (Y) y la o las variables independientes.

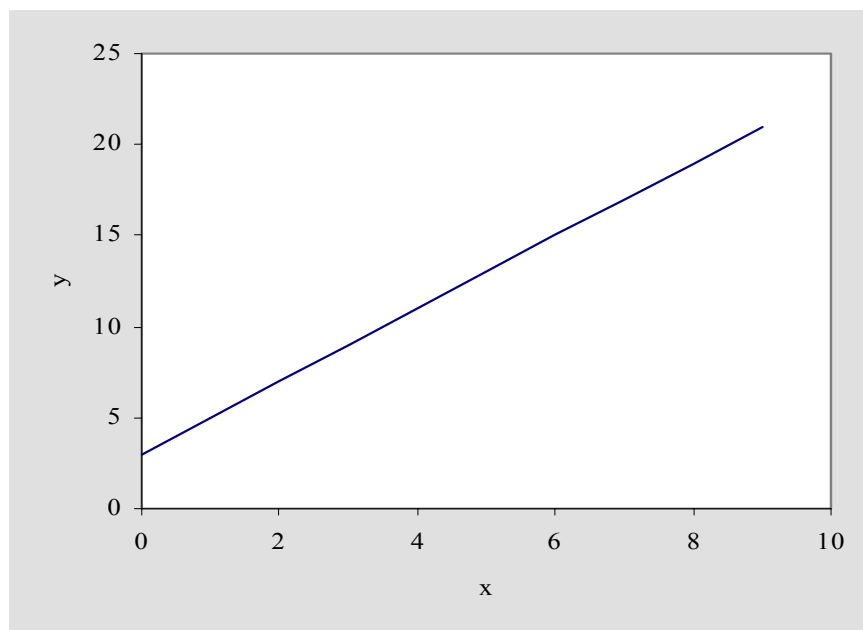
La función más simple para la relación entre dos variables es la **FUNCIÓN LINEAL**

$$Y = a + b X$$

- ♦ Esta expresión es una aproximación de la verdadera relación entre X e Y.
- ♦ Para un dado valor de X el modelo predice un cierto valor para Y.
- ♦ Mientras mejor sea la predicción, mejor es el modelo para explicar el fenómeno.

Por ejemplo,

$$Y = 2 X + 3$$



Interpretación de los coeficientes:

- ✓ el coeficiente **a** es la **PENDIENTE** de la recta, mide el cambio en Y por cada unidad de cambio en X, en el ejemplo la pendiente es 2.
- ✓ El coeficiente **b** es la **ORDENADA AL ORIGEN**, el punto donde la recta intercepta el eje Y, es decir el valor de Y cuando X = 0.

Consideremos el modelo $Y = aX + b$

- ◆ Este modelo es una aproximación de la verdadera relación entre X e Y.
- ◆ Para un dado valor de X el modelo predice un cierto valor para Y.
- ◆ Mientras mejor sea la predicción, mejor es el modelo.

Un MODELO DETERMINÍSTICO supone que bajo condiciones ideales, el comportamiento de la variable dependiente puede ser totalmente descrito por una función matemática de las variables independientes (o por un conjunto de ecuaciones que relacionen las variables). Es decir, en condiciones ideales el modelo permite predecir SIN ERROR el valor de la variable dependiente.

- ◆ Ejemplo: Ley de la Gravedad.

Podemos predecir exactamente la posición de un objeto que cae en caída libre y en el vacío para cada instante de tiempo.

Un MODELO ESTADÍSTICO permite la incorporación de un COMPONENTE ALEATORIO en la relación. En consecuencia, las predicciones obtenidas a través de modelos estadísticos tendrán asociado un error de predicción.

- ◆ Ejemplo: Relación de la altura con la edad en niños.

Niños de la misma edad seguramente no tendrán la misma altura. Sin embargo, a través de un modelo estadístico es posible concluir que la altura aumenta con la edad. Es más, podríamos predecir la altura de un niño de cierta edad y asociarle un ERROR DE PREDICCIÓN que tiene en cuenta: ERRORES DE MEDICIÓN y VARIABILIDAD ENTRE INDIVIDUOS.

En problemas biológicos, trabajando en “condiciones ideales” es posible evitar los errores de medición, pero no la variabilidad individual, por eso es indispensable incluir el componente aleatorio en los modelos estadísticos.

En este curso trataremos sobre Regresión Lineal. Haremos énfasis en este tipo de modelos porque

- ◆ son de amplia aplicación,
- ◆ son más simples de implementar,

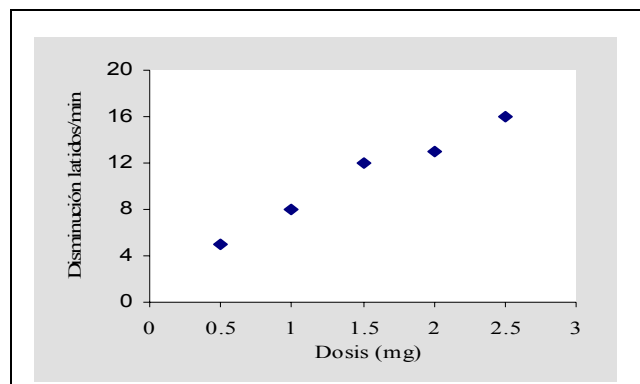
- ♦ otros procedimientos estadísticos más complejos pueden ser mejor comprendidos luego de estudiar regresión lineal.

REGRESIÓN LINEAL SIMPLE

Consideremos el siguiente experimento controlado y aleatorizado para estudiar el efecto de una nueva droga sobre la frecuencia cardiaca de ratas sanas.

Cinco ratas fueron asignadas aleatoriamente a una de cinco dosis y se registró la máxima disminución observada en la frecuencia cardiaca en una hora. Los datos obtenidos son:

Dosis (mg)	Máxima disminución de la FC (DFC)
0.5	5
1.0	8
1.5	12
2.0	13
2.5	16



La relación respuesta-dosis es aparentemente lineal. Parece razonable proponer

$$DFC = \beta_0 + \beta_1 * DOSIS + \text{error}$$

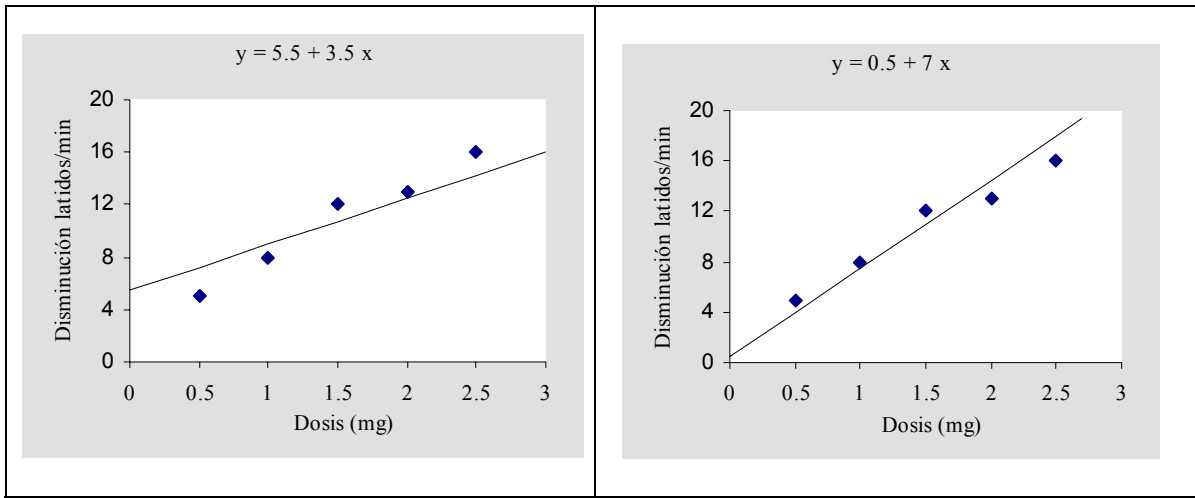
$$Y_i = \beta_0 + \beta_1 * X_i + \epsilon_i$$

Podríamos intentar ajustar una recta “a ojo”. Propuestas:

$$y_i = 5.5 + 3.5 * x_i$$

$$y_i = 0.5 + 7.0 * x_i$$

¿Cuál recta es “mejor”? ¿Cómo decidir? Veamos los gráficos.



Para decidir cuál de las dos rectas ajusta mejor estos datos consideraremos una medida de cuán lejos está cada dato de la recta propuesta ⇒ RESIDUO.

RESIDUOS



RESIDUOS



x	Y _{obs}	Y _{ajus}	(Y _{obs} -Y _{ajus})	(Y _{obs} -Y _{ajus}) ²	x	Y _{obs}	Y _{ajus}	(Y _{obs} - Y _{ajus})	(Y _{obs} -Y _{ajus}) ²
0.5	5.0	7.3	-2.3	5.1	0.5	5.0	4.0	1.0	1.00
1.0	8.0	9.0	-1.0	1.0	1.0	8.0	7.5	0.5	0.25
1.5	12.0	10.8	1.3	1.6	1.5	12.0	11.0	1.0	1.00
2.0	13.0	12.5	0.5	0.3	2.0	13.0	14.5	-1.5	2.25
2.5	16.0	14.3	1.8	3.1	2.5	16.0	18.0	-2.0	4.00
Total=			0.3	10.9	Total=			-1	8.50



$$\sum (y_i - 5.5 - 3.5x_i)^2$$



$$\sum (y_i - 0.5 - 7x_i)^2$$

La mejor recta sería aquella que minimice la suma de las distancias al cuadrado de los puntos a la recta, es decir deberíamos encontrar $\hat{\beta}_0$ y $\hat{\beta}_1$ tales que

$$\sum (y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i)^2 \leq \sum (y_i - b_0 - b_1 x_i)^2$$

para cualquier elección de b_0 y b_1 que hagamos.

Este método para encontrar la recta que mejor ajuste a los datos se conoce como **MÉTODO DE MÍNIMOS CUADRADOS**.

Afortunadamente no es necesario probar con diferentes rectas cuál de ellas es la que produce la menor suma de cuadrados, ya que es posible encontrar analíticamente las expresiones para $\hat{\beta}_0$ y $\hat{\beta}_1$. En el caso general en que tenemos n pares de observaciones (X_i, Y_i) , $\hat{\beta}_0$ y $\hat{\beta}_1$ son las soluciones del sistema de ecuaciones normales:

$$\frac{\partial}{\partial \beta_0} \sum (y_i - \beta_0 - \beta_1 x_i)^2 = 0$$

$$\frac{\partial}{\partial \beta_1} \sum (y_i - \beta_0 - \beta_1 x_i)^2 = 0$$

y se obtiene:

$$\hat{\beta}_1 = \frac{\sum_{i=1}^n (X_i - \bar{X})(Y_i - \bar{Y})}{\sum_{i=1}^n (X_i - \bar{X})^2}$$

$$\hat{\beta}_0 = \bar{Y} - \hat{\beta}_1 \bar{X} \quad \Leftarrow \quad \text{la recta pasa por el punto } (\bar{X}, \bar{Y})$$

En el ejemplo de dosis-respuesta los estimadores de mínimos cuadrados para β_0 y β_1 resultan ser:

$$\hat{\beta}_1 = \frac{(0.5 - 1.5)(5.0 - 10.8) + \dots + (2.5 - 1.5)(16 - 10.8)}{(0.5 - 1.5)^2 + \dots + (2.5 - 1.5)^2} = \frac{13.5}{2.5} = 5.4$$

$$\hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x} = 10.8 - 5.4 \cdot 1.5 = 2.7$$

La RECTA AJUSTADA para nuestros datos es

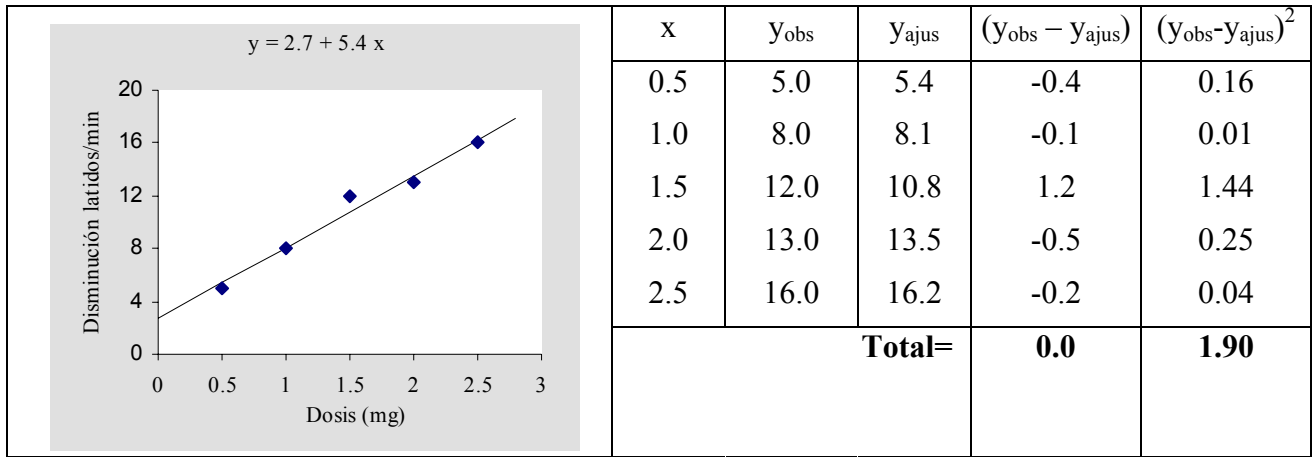
$$\hat{Y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_i = 2.7 + 5.4 x_i$$

¿Qué nos indican los valores de los coeficientes?

- ♦ $\hat{\beta}_0$ = ORDENADA AL ORIGEN (intercept) = 2.7 \Rightarrow es el punto donde la recta corta el eje vertical, es decir, la disminución esperada en el número de pulsaciones cuando la dosis es cero.

No interpretable si el 0 no está contenido en el rango de valores de X.

- ♦ $\hat{\beta}_1$ = PENDIENTE = 5.4 \Rightarrow nos dice que por cada mg de aumento en la dosis se espera un cambio de 5.4 pulsaciones/min en la FC.
- ♦ Si $\hat{\beta}_1 = 0$ entonces $\hat{\beta}_0 = \bar{Y}$. La media de los datos es el estimador de mínimos cuadrados cuando no hay variables regresoras.



Notación

VALORES ESTIMADOS DE LOS PARÁMETROS $\hat{\beta}_0, \hat{\beta}_1$

VALOR PREDICHO $\hat{Y}_i = (\hat{\beta}_0 + \hat{\beta}_1 x_i)$

RESIDUO o RESIDUAL = outcome observado – valor predicho
 $= Y_i - (\hat{\beta}_0 + \hat{\beta}_1 x_i)$

PENDIENTE ESTANDARIZADA

La pendiente $\hat{\beta}_1$ nos indica si hay relación entre las dos variables, su signo nos indica si la relación es positiva o negativa, pero no mide la FUERZA de la asociación.

La razón es que su valor numérico depende de las unidades de medida de las dos variables. Un cambio de unidades en una de ellas puede producir un cambio drástico en el valor de la pendiente.

Ejemplo

x	y	recta ajustada	x'	y	recta ajustada
2	10	5.7 + 2.3 x	2	1.0	0.57 + 0.23 x
3	13		3	1.3	
4	15		4	1.5	
5	17		5	1.7	

Por esa razón, puede resultar interesante considerar una versión estandarizada de la pendiente

$$\hat{\beta}_1^* = \hat{\beta}_1 \frac{s_x}{s_y}$$

donde s_x y s_y son las desviaciones estándares de las X's y de las Y's respectivamente.

Esta es la pendiente que se obtendría al hacer la regresión de los scores Z de la variable dependiente respecto de los scores Z de la variable regresora.

INTERESANTE!!!

$$\hat{\beta}_1^* = \hat{\beta}_1 \frac{s_x}{s_y} = r$$

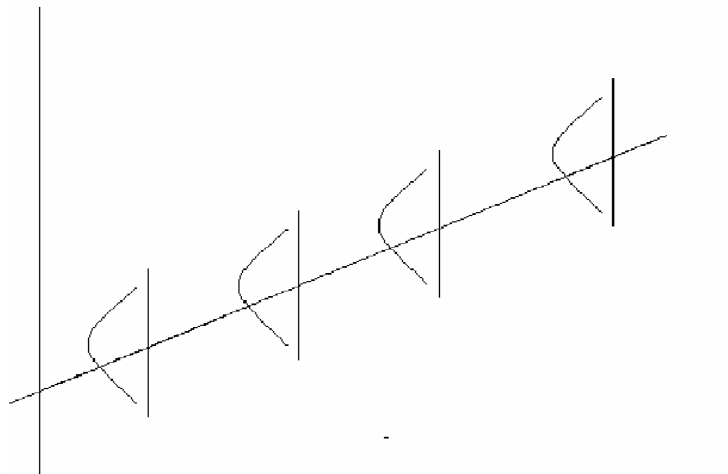
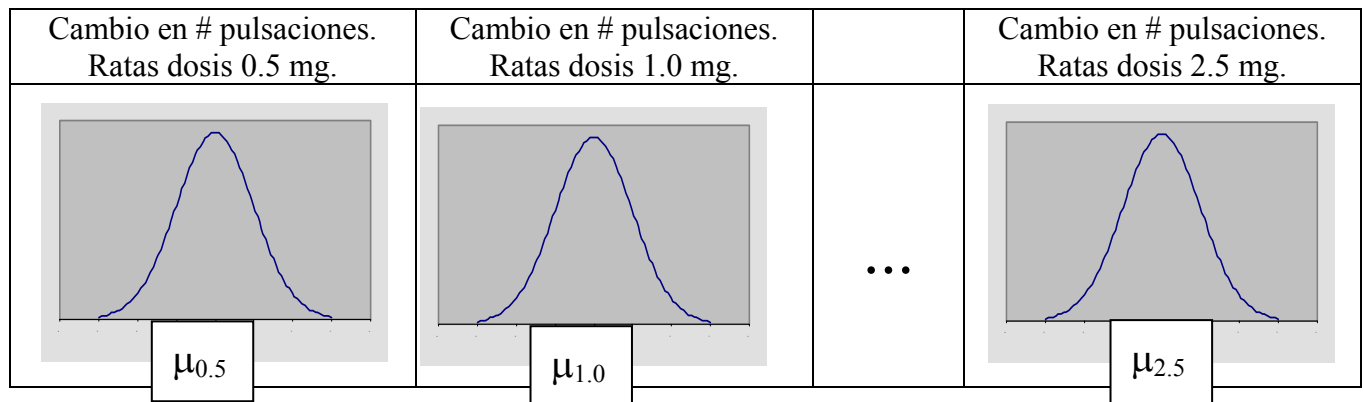
donde r es el **coeficiente de correlación de Pearson**. Notar que si $s_x = s_y$ tenemos $\hat{\beta}_1^* = \hat{\beta}_1 = r$.

Esta relación directa entre el coeficiente de correlación de Pearson y la pendiente de la recta de regresión sólo es válida en el contexto de regresión simple (una variable regresora) no vale para el caso de regresión múltiple (más de una variable regresora).

Propiedades del coeficiente de correlación (de la pendiente estandarizada)

- r mide la fuerza de la asociación LINEAL entre X e Y .
- $-1 \leq r \leq 1$
- $r = 0$ implica que no hay relación lineal
- $r = \pm 1$ cuando todos los puntos caen sobre la recta
- r tiene el mismo signo que la pendiente
- mientras mayor el valor absoluto de r mayor la fuerza de la asociación
- el valor de r no depende de las unidades de medición
- el coeficiente de correlación trata a X e Y simétricamente. Si ajustamos $Y = \alpha + \beta X$ o $X = \alpha^* + \beta^* Y$, en ambos casos obtendremos el mismo coeficiente de correlación, pero no la misma pendiente!!

EL MODELO DE REGRESIÓN LINEAL (ORDINARIO)



OBSERVACIONES $(X_1, Y_1), \dots, (X_N, Y_N)$

$$Y_i = \beta_0 + \beta_1 x_i + \varepsilon_i \quad i = 1, 2, \dots, n$$

con ε_i independientes y $\varepsilon_i \sim N(0, \sigma^2)$

Y_i = disminución en la FC de la rata i

x_i = dosis de droga recibida por la rata i

ε_i = término error para la rata i

Supuestos

1. NORMALIDAD. Para cada valor de X, Y es una variable aleatoria con distribución Normal con media μ_x . [La distribución de la DFC para cada dosis de la droga es Normal con media μ_x].
2. HOMOSCEDASTICIDAD. Todas las distribuciones poblacionales tienen la misma varianza. [La varianza de DFC es la misma para todas las dosis].
3. LINEALIDAD. Las medias μ_x de las distintas poblaciones están relacionadas linealmente con X. [La media poblacional en la DFC cambia linealmente con la dosis].

$$\mu_x = E(Y/ X= \mathbf{x}) = \beta_0 + \beta_1 \mathbf{x}$$

- $\Rightarrow \beta_0$ = ordenada al origen = MEDIA POBLACIONAL de la variable resultante (DFC) cuando la variable regresora (dosis) toma valor 0.
 - $\Rightarrow \beta_1$ = pendiente = cambio en la MEDIA POBLACIONAL de la variable resultante (DFC) cuando la variable regresora (dosis) aumenta en 1 unidad.
4. INDEPENDENCIA. Los valores de Y son estadísticamente independientes. [Este supuesto no se cumple, por ejemplo, si para algunos de los individuos tenemos observaciones repetidas].

Comentarios.

- Generalmente no sabemos si los supuestos son verdaderos, ni conocemos los valores de los parámetros β_0 y β_1 . El proceso de estimar los parámetros de un modelo lineal y valorar si el modelo es adecuado para nuestros datos se denomina Análisis de Regresión.
- En este modelo suponemos que la variable X NO TIENE ERROR!!! El término de error (ε) mide la variabilidad de la variable aleatoria Y para cada nivel FIJO de la variable X.
- En nuestro ejemplo dosis-frecuencia cardíaca los valores de la variable explicativa fueron FIJADOS por el investigador. En el caso general, en que ambas variables se miden simultáneamente (edad materna y peso del niño al

nacer, por ejemplo) suponemos que los valores de la variable regresora no tienen error. Esto difícilmente sea cierto!!!!

ESTIMACIÓN DE β_0 Y β_1

Los parámetros del modelo lineal se estiman a través del método de mínimos cuadrados. Llamamos $\hat{\beta}_0$ y $\hat{\beta}_1$ a los estimadores de mínimos cuadrados de β_0 y β_1 , para obtenerlos no es necesario hacer los supuestos 1,2 y 4, sólo el de LINEALIDAD.

- ♦ $\hat{\beta}_0$ es un *estimador insesgado* de β_0
- ♦ $\hat{\beta}_1$ es un *estimador insesgado* de β_1

Esto significa que:

- ♦ $\hat{\beta}_0$ tiene una distribución de muestreo con media β_0 y
- ♦ $\hat{\beta}_1$ tiene una distribución de muestreo con media β_1

Recordar. La distribución de muestreo de $\hat{\beta}_0$ y $\hat{\beta}_1$ se obtendría empíricamente repitiendo infinitas veces el experimento (5 ratas con las mismas dosis y en las mismas condiciones), y calculando para cada repetición las estimaciones $\hat{\beta}_0$ y $\hat{\beta}_1$ de los parámetros. Con las infinitas estimaciones de $\hat{\beta}_0$ construimos un histograma, que corresponde a la distribución de muestreo del estimador $\hat{\beta}_0$. Del mismo modo para $\hat{\beta}_1$.

$\hat{\beta}_0$ y $\hat{\beta}_1$ son INSESGADOS aún cuando los supuestos de homoscedasticidad y normalidad sean falsos!!!

IMPORTANTE!!!

Hemos escrito TRES ecuaciones de REGRESIÓN para el mismo problema:

Ecuación	Notación	
Recta de Regresión Poblacional	$\mu_x = \beta_0 + \beta_1 X$	DESCONOCIDO
Modelo de Regresión Lineal	$Y = \beta_0 + \beta_1 X + \varepsilon$	DESCONOCIDO
Recta de Regresión Estimada	$\hat{Y} = \hat{\beta}_0 + \hat{\beta}_1 X$	CONOCIDO

Nuestra ecuación $\hat{Y} = \hat{\beta}_0 + \hat{\beta}_1 x$ es una estimación de la verdadera recta poblacional.

RECAPITULEMOS

⇒ Hasta aquí:

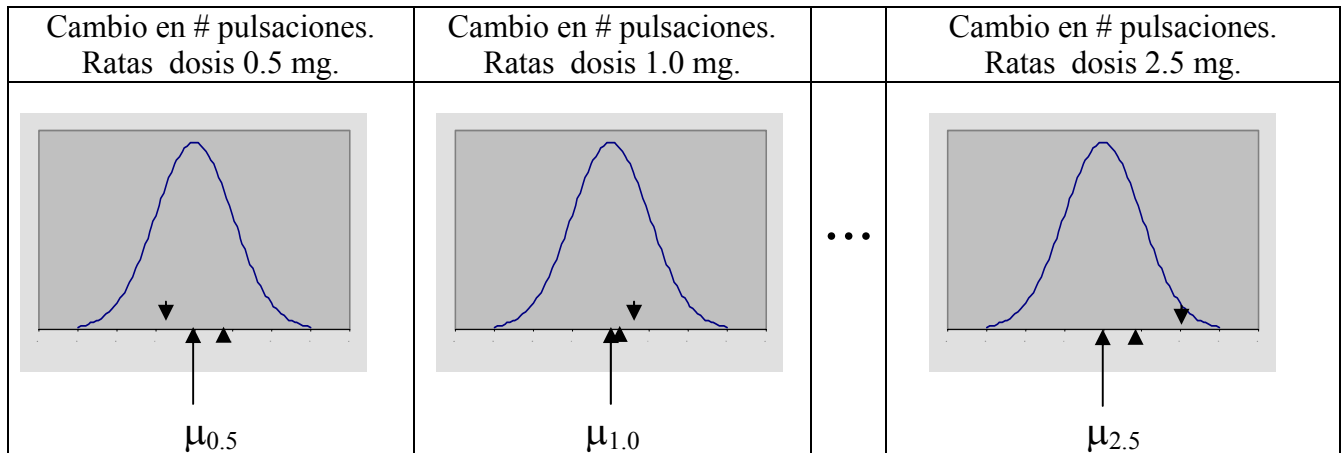
1. Planteamos el Modelo de regresión lineal homoscedástico
2. Presentamos estimadores de mínimos cuadrados para la pendiente y la ordenada al origen.
 - . ¿Cómo se obtienen los estimadores? → Métodos de Mínimos Cuadrados ordinarios.

⇒ En lo que sigue:

3. Estimaremos σ^2 , la varianza común a las distintas poblaciones.
4. Obtendremos el error estándar de $\hat{\beta}_1$ y el de $\hat{\beta}_0$ para construir:
 - . Tests de hipótesis para β_0 y β_1 .
 - . Intervalos de confianza para β_0 y β_1 .
5. Construiremos la Tabla de Análisis de Varianza.

¿CÓMO ESTIMAMOS LA VARIANZA σ^2 COMÚN A TODAS LAS POBLACIONES?

Recordemos nuestro modelo



▼ Valor observado (Y_i) ▲ Valor predicho (\hat{Y}_i , que estima a μ_i)

Distancia entre ▼ y ▲ = $(Y_i - \hat{Y}_i) =$ residuo del dato $i \quad i = 1, 2, \dots, 5$

Cada residuo $(Y_i - \hat{Y}_i)$ provee oportunidad de medir la variabilidad en cada población individual. Entonces, un candidato natural para estimar la varianza poblacional sería

$$\frac{\sum_{i=1}^5 (Y_i - \hat{Y}_i)^2}{5}$$

Sin embargo, no tenemos 5 residuos independientes porque existen dos vínculos entre ellos. En consecuencia, tenemos sólo 3 GRADOS DE LIBERTAD en la suma de los residuos. Entonces estimamos σ^2 con

$$\frac{\sum_{i=1}^5 (Y_i - \hat{Y}_i)^2}{(5 - 2)}$$

Si el tamaño de muestra fuera n usaríamos

$$\frac{\sum_{i=1}^n (Y_i - \hat{Y}_i)^2}{(n - 2)}$$

Notación

Varianza de cada población (Varianza error)

$$\sigma^2$$

Suma de Cuadrados Residual

$$RSS = \sum_{i=1}^n (Y_i - \hat{Y}_i)^2$$

Varianza de los Residuos (Estima a σ^2)

$$s_e^2 = \frac{\sum_{i=1}^n (Y_i - \hat{Y}_i)^2}{(n - 2)} = \frac{RSS}{(n - 2)}$$

Desviación estándar de los Residuos

$$s_e = \sqrt{s_e^2}$$

PROPIEDAD: Bajo los supuestos del modelo la varianza residual s_e^2 es un estimador insesgado de σ^2 .

Ejemplo (continuación)

x_i	y_i	$\hat{y}_i=2.7+5.4x_i$	$(y_i - \hat{y}_i)^2$
0.5	5.0	5.4	0.16
1.0	8.0	8.1	0.01
1.5	12.0	10.8	1.44
2.0	13.0	13.5	0.25
2.5	16.0	16.2	0.04
Total=			1.90

Suma de Cuadrados Residual

$$RSS = \sum_{i=1}^n (Y_i - \hat{Y}_i)^2 = 1.90$$

Varianza de los Residuos

$$s_e^2 = \frac{RSS}{(n - 2)} = \frac{1.90}{3} = 0.633$$

Desviación estándar de los Residuos

$$s_e = \sqrt{0.633} = 0.796$$

INFERENCIA PARA LA PENDIENTE

Intentaremos construir un intervalo de confianza y un test para β_1 .

Bajo los supuestos del modelo lineal (normalidad, homogeneidad de varianzas, linealidad e independencia) el estimador $\hat{\beta}_1$ de la pendiente tiene distribución normal con media β_1 y varianza $Var(\hat{\beta}_1)$, y por lo tanto

$$T = \frac{\hat{\beta}_1 - \beta_1}{SE(\hat{\beta}_1)} \sim t_{n-2}$$

donde $SE(\hat{\beta}_1)$, el error estándar del estimador de la pendiente, se define como:

$$SE^2(\hat{\beta}_1) = \frac{s_e^2}{\sum_{i=1}^n (x_i - \bar{x})^2} = \frac{s_e^2}{(n-1)s_x}$$

Recordemos que $SE(\hat{\beta}_1)$ es un estimador de la desviación estándar de la distribución de muestreo de $\hat{\beta}_1$.

Notemos que la varianza de $\hat{\beta}_1$ disminuye (la estimación es más precisa) cuando:

- (1) La varianza σ^2 disminuye.
- (2) La varianza de la variable regresora aumenta \Rightarrow Mientras más amplio el rango de valores de la covariable, mayor la precisión en la estimación.
- (3) El tamaño de muestra aumenta.

Ejemplo (continuación)

x_i	y_i	$\hat{y}_i=2.7+5.4x_i$	$(y_i - \hat{y}_i)^2$	$(x_i - \bar{x})^2$
0.5	5.0	5.4	0.16	1.50
1.0	8.0	8.1	0.01	0.25
1.5	12.0	10.8	1.44	0.00
2.0	13.0	13.5	0.25	0.25
2.5	16.0	16.2	0.04	1.00
Total=			1.90	2.50

$$SE^2(\hat{\beta}_1) = \frac{s_e^2}{\sum_{i=1}^n (x_i - \bar{x})^2} = \frac{0.633}{2.50} = 0.2533$$

$$SE(\hat{\beta}_1) = \sqrt{0.2533} = 0.503$$

Nota. Un valor pequeño de $SE(\hat{\beta}_1)$ nos indica que la estimación de la pendiente variará poco de muestra en muestra (para este conjunto dado de valores de X).

INTERVALO DE CONFIANZA PARA β_1

Un intervalo de confianza de nivel $(1 - \alpha)$ para el parámetro β_1 (pendiente de la recta de regresión poblacional) está dado por:

$$\hat{\beta}_1 \pm t_{n-2, \alpha/2} SE(\hat{\beta}_1),$$

donde $t_{n-2, \alpha/2}$ es el percentil de la distribución t de Student con $n - 2$ grados de libertad que deja a su derecha un área $\alpha/2$.

Ejemplo (continuación)

Intervalo de confianza de nivel 95% para la pendiente

$$\hat{\beta}_1 \pm t_{n-2, \alpha/2} SE(\hat{\beta}_1) = 5.4 \pm 3.182 \cdot 0.503 = 5.4 \pm 1.60 = (3.8, 7.00)$$

percentil de la distribución t con 3 grados de libertad que deja a su derecha 2.5% del área

¿Cómo interpretamos este intervalo?

El intervalo puede no tener sentido si un incremento en X de 1 unidad es relativamente grande o relativamente pequeño en términos prácticos. Si en nuestro ejemplo nos interesara un IC para un cambio en la dosis de 0.2 mg simplemente lo obtenemos multiplicando los extremos del IC anterior por la constante. Obtendríamos $(3.8 \cdot 0.2, 7.0 \cdot 0.2) = (0.78, 1.4)$.

TEST DE HIPÓTESIS PARA LA PENDIENTE

Queremos un test de nivel α para las hipótesis

$$H_0: \beta_1 = \beta_1^* \quad \text{versus} \quad H_1: \beta_1 \neq \beta_1^*$$

β_1^* es algún valor propuesto por el investigador.

El test para $H_0: \beta_1 = 0$ se conoce como test de *independencia* o de *no asociación* ya que nos dice si las variables están asociadas o no. (El test t que hace STATA considera $\beta_1^* = 0$).

El test se basa en el estadístico

$$T = \frac{\hat{\beta}_1 - \beta_1^*}{SE(\hat{\beta}_1)} \sim t_{n-2}$$

y rechaza H_0 cuando el valor del estadístico observado en la muestra da grande y positivo o grande y negativo, es decir, el p-valor da menor que el nivel α .

Ejemplo (continuación)

Hacemos un test de nivel $\alpha = 0.05$ para las hipótesis

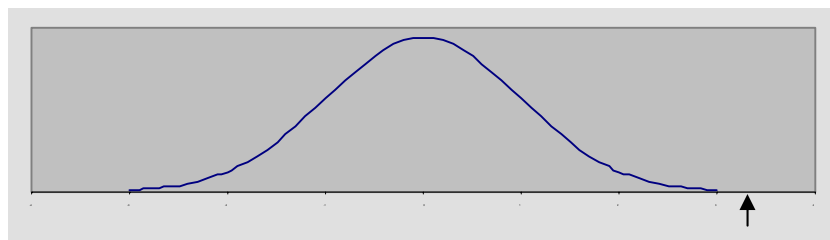
$$H_0: \beta_1 = 0 \quad \text{versus} \quad H_1: \beta_1 \neq 0$$

El valor del estadístico obtenido a partir de nuestros datos es

$$T = \frac{\hat{\beta}_1 - 0}{SE(\hat{\beta}_1)} = \frac{5.4 - 0}{0.503} = 10.736$$

Para calcular el p-valor, utilizamos la distribución t_3 . El área a la derecha de 10.736 es 0.00085, entonces $p = 0.0017$.

Distribución t_3



10.736

Concluimos que la pendiente es significativamente diferente de 0. Es decir, existe una relación positiva entre dosis y respuesta, que no puede ser atribuida al azar. Nuestros datos no son consistentes con la hipótesis nula de no relación entre FDC y dosis.

ESTIMACIÓN DE LA ORDENADA AL ORIGEN (β_0)

Tal como ocurre para $\hat{\beta}_1$, bajo el modelo propuesto $\hat{\beta}_0$ tiene distribución normal con media β_0 y varianza $\sigma^2(\hat{\beta}_0)$ que se estima como

$$SE^2(\hat{\beta}_0) = s_e^2 \left(\frac{1}{n} + \frac{\bar{x}^2}{\sum_{i=1}^n (x_i - \bar{x})^2} \right)$$

Construimos intervalos de confianza y test de hipótesis para β_0 de forma análoga a como lo hicimos para β_1 .

- La mayoría de los paquetes estadísticos devuelve el test para $H_0: \beta_0 = 0$, pero este test en general no tiene interés.
- El parámetro β_0 en general carece de interpretación, salvo que el rango de variación de los datos contenga a $X=0$.

FUENTES DE VARIABILIDAD EN NUESTROS DATOS

Habíamos observado cinco valores de disminución de la FC en nuestras ratas

5, 8, 12, 13 y 16

¿Por qué varían estas cinco respuestas?

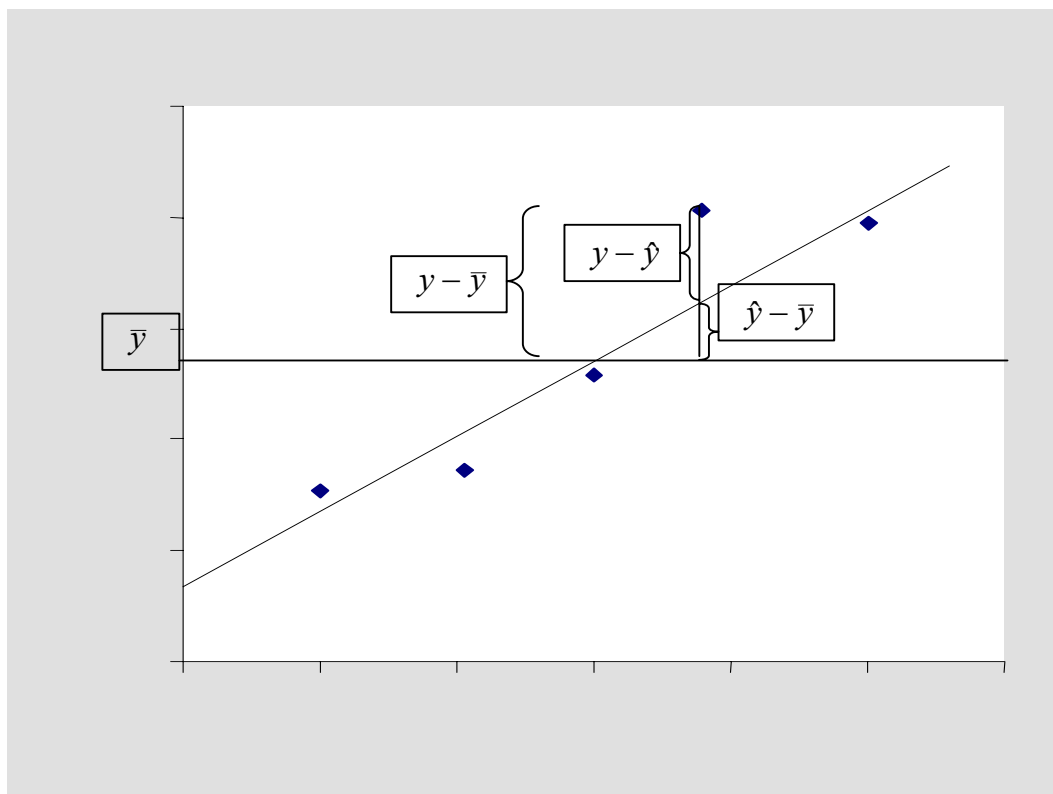
- (1) Porque las diferentes ratas recibieron diferentes dosis de la droga.

VARIABILIDAD EXPLICADA POR LA VARIABLE REGRESORA

(2) Porque aunque hubieran recibido la misma dosis la respuesta no hubiera sido *exactamente igual* en las 5 ratas debido a diferentes causas. Por ejemplo,

- las ratas no responden exactamente igual a la misma dosis,
- diferente manejo del investigador al manipular las ratas que afecta la FC,
- condiciones basales de las ratas ligeramente diferentes,
- errores en los instrumentos de medición, etc. etc.

VARIABILIDAD RESIDUAL (NO EXPLICADA POR LA DOSIS).



Tenemos entonces,

1. Una medida de la VARIABILIDAD TOTAL de la variable Y (cuando no tenemos en cuenta la variable regresora) es la suma de las desviaciones a la media al cuadrado.

$$\text{Total Sum of Squares} = \text{TSS} = \sum_{i=1} (Y_i - \bar{Y})^2$$

2. Una medida de la VARIABILIDAD NO EXPLICADA por la variable regresora es la suma de los residuos al cuadrado.

$$\text{Residual Sum of Squares} = \text{RSS} = \sum_{i=1} (Y_i - \hat{Y}_i)^2$$

3. Finalmente, una medida de cuánto contribuye la variable X a explicar la variabilidad de Y (VARIABILIDAD EXPLICADA POR EL MODELO DE REGRESIÓN) está dada por

$$\text{Regression Sum of Squares} = \text{RegSS} = \sum_{i=1} (\hat{Y}_i - \bar{Y})^2$$

$$\text{Resultado interesante} \Rightarrow \text{TSS} = \text{RegSS} + \text{RSS}$$

Trataremos de construir una medida de la fuerza de la relación entre la variable dependiente e independiente, que nos indique cuán buen predictor de Y es X. Se trata de decidir si el hecho de conocer el valor de X (dosis)

Si uno puede predecir Y mucho mejor usando la recta de regresión $\hat{Y} = \hat{\beta}_0 + \hat{\beta}_1 x$ que sin conocer el valor de X, entonces las variables están asociadas.

La medida de asociación que propondremos se construye con 4 elementos:

- Una regla para predecir Y cuando no conocemos X $\Rightarrow \bar{Y}$
- Una regla para predecir Y cuando conocemos X $\Rightarrow \hat{Y} = \hat{\beta}_0 + \hat{\beta}_1 x$
- Una medida resumen de los errores que se cometen con cada regla
 - $\Rightarrow \text{TSS}$ para la Regla 1
 - $\Rightarrow \text{RSS}$ para la Regla 2
- Una medida de cuánto se reduce el error al usar la regla más sofisticada

$$R^2 = \frac{\text{TSS} - \text{RSS}}{\text{TSS}} = \frac{\text{RegSS}}{\text{TSS}}$$

llamada **COEFICIENTE DE DETERMINACIÓN**

R^2 nos dice qué porcentaje de la variabilidad total en la variable Y puede ser explicada por la variable regresora, en consecuencia es una medida de la capacidad de PREDICCIÓN del modelo.

R^2 también puede verse como es una medida de la fuerza de la ASOCIACIÓN LINEAL entre X e Y. (Hacemos énfasis en la palabra lineal porque fue obtenido bajo un modelo lineal)

Propiedades de R^2

- $0 \leq R^2 \leq 1$
- No depende de las unidades de medición.
- Es el cuadrado del coeficiente de correlación de Pearson (deberíamos usar la notación r^2 , pero ...)
- Mientras mayor es R^2 mayor es la fuerza de la variable regresora para predecir el outcome.
- Mientras mayor sea R^2 menor es la RSS y por lo tanto, más cercanos están los puntos a la recta.
- Toma el mismo valor cuando usamos a X para predecir a Y o cuando usamos a Y para predecir a X.

Ejemplo (continuación)

En nuestro ejemplo
$$R^2 = \frac{ESS}{TSS} = \frac{72.9}{74.80} = 0.975.$$

Entonces, el 97% de la variación observada en los datos de DFC es explicada por la dosis de droga. La dosis es un excelente predictor de la DFC.

Pero CUIDADO !!! Cuando tenemos sólo dos observaciones ($n = 2$), se obtiene $R^2 = 1$ independientemente de los datos ... porque dos puntos determinan una línea recta, así que mínimos cuadrados daría un ajuste perfecto!!

$$RSS = 0 \quad \Rightarrow \quad RegSS = TSS \quad \Rightarrow \quad R^2 = 1$$

TABLA DE ANALISIS DE LA VARIANZA (ANOVA)

Las sumas de cuadrados correspondientes a las tres fuentes de variación que hemos descripto arriba se presentan habitualmente en una Tabla, denominada Tabla de ANOVA.

Fuente de Variación (Source)	SS	d.f.	MS
Model	RegSS	1	RegSS/1
Residual	RSS	n – 2	RSS/(n–2) ← s_e^2
Total	TSS	n – 1	

La columna d.f. indica los grados de libertad de cada SS.

- El modelo tiene d.f. = # de parámetros en el modelo – 1= # de covariables en el modelo.
- La suma de cuadrados residual tiene n – 2 grados de libertad (estamos estimando dos parámetros en el modelo)
- La suma de cuadrados total tiene n – 1 grados de libertad (hay un vínculo que liga las desviaciones respecto de la media).

La columna MS (Mean Square) se obtiene como el cociente entre la SS y sus correspondientes grados de libertad.

OTRO TEST PARA LA PENDIENTE ...

A partir de la Tabla es posible derivar un test para $H_0 : \beta_1 = 0$.

[En el contexto de regresión lineal simple ya hemos obtenido el test t que resuelve este punto, este test será más importante en Regresión Múltiple].

El razonamiento es el siguiente: Bajo los supuestos del modelo de regresión,

- (1) La distribución de muestreo de RMS = RSS/(n – 2) tiene esperanza σ^2
- (2) La distribución de muestreo de EMS=RegSS/1 tiene esperanza

$$\sigma^2 + \beta_1^2 \sum_{i=1}^n (x_i - \bar{x})^2$$

Entonces, cuando H_0 es verdadera el MS del modelo y el MS residual deberían parecerse mucho, o su cociente debería parecerse a 1. Por lo tanto, es razonable considerar el estadístico

$$F = \frac{\text{RegMS}}{\text{RMS}} = \frac{\text{RegMS}}{\text{RSS}/(n-2)}$$

como candidato para testear la hipótesis $H_0 : \beta_1 = 0$. Esperamos que F esté cerca de 1 si H_0 es verdadera y que F sea grande y positiva cuando H_0 es falsa.

Bajo los supuestos del modelo lineal y cuando H_0 es verdadera

$F \sim$ Distribución de Fisher con $(1, n-2)$ grados de libertad

Por lo tanto un test de nivel α para las hipótesis

$$H_0 : \beta_1 = 0 \quad \text{versus} \quad H_1 : \beta_1 \neq 0$$

rechazará H_0 si el valor del estadístico para los datos de la muestra produce un p-valor (calculado sobre la distribución F de Fisher) menor que el nivel α .

Nota 1. El test es de dos colas, porque detecta apartamientos de H_0 tanto en la dirección positiva como en la dirección negativa. Sin embargo, la hipótesis nula **NO SE RECHAZA** si el cociente F es pequeño!!!

Nota 2. En el caso de Regresión Lineal Simple, el test que hemos construido basado en la distribución F de Fisher y el test basado en la distribución t (a dos colas) son equivalentes, ya que ambas distribuciones están relacionadas. Además ambos testean la misma hipótesis. Esto no es así para el caso de Regresión Lineal Múltiple.

STATISTIX

Datos: Y X

Statistics / Linear Model / Linear Regression

Dependent Variable => Y

Independent Variable => X

UNWEIGHTED LEAST SQUARES LINEAR REGRESSION OF Y

PREDICTOR VARIABLES	COEFFICIENT	STD ERROR	STUDENT'S T	P
CONSTANT	2.70000	0.83467	3.23	0.0480
X	5.40000	0.50332	10.73	0.0017
R-SQUARED	0.9746	RESID. MEAN SQUARE (MSE)		0.63333
ADJUSTED R-SQUARED	0.9661	STANDARD DEVIATION		0.79582

SOURCE	DF	SS	MS	F	P
REGRESSION	1	72.9000	72.9000	115.11	0.0017
RESIDUAL	3	1.90000	0.63333		
TOTAL	4	74.8000			

CASES INCLUDED 5 MISSING CASES 0

ESTIMACIÓN DE LA MEDIA DE Y PARA UN DADO VALOR DE X

Supongamos que nos interesa estimar la MEDIA de la disminución de FC en la población de ratas tratadas con cierta dosis x_0 de la droga, por ejemplo 1.5 mg. [Esta población es hipotética, y se la obtendría usando la misma dosis en ratas en las mismas condiciones que las de nuestro experimento].

Nuestro estimador de μ cuando $x = x_0$ es

$$\hat{\mu}_{x_0} = \hat{Y}_{x_0} = \hat{\beta}_0 + \hat{\beta}_1 x_0$$

Por ejemplo, para $x = 1.5$ mg. estimamos una reducción media de la FC

$$\hat{Y}_{1.5} = 2.7 + 5.4 \cdot 1.5 = 10.8 \text{ latidos/min.}$$

Para decir cuán precisa es nuestra estimación, necesitamos un ESTIMADOR del ERROR ESTÁNDAR de $\hat{\mu}_x$ es

$$SE(\hat{\mu}_{x_0}) = s_e \sqrt{\frac{1}{n} + \frac{(x_0 - \bar{x})^2}{\sum_{i=1}^n (x_i - \bar{x})^2}}$$

Notar que el error estándar depende de la distancia de x_0 a \bar{x} . Cuanto más cercano a la media el valor de x , menor el error de estimación.

Bajo los supuestos del modelo, un INTERVALO DE CONFIANZA DE NIVEL $(1-\alpha)$ para μ_x (el valor esperado de FDC para la población de ratas sometidas a dosis x) es

$$\hat{\mu}_x \pm t_{n-2, \alpha/2} SE(\hat{\mu}_x)$$

donde $t_{n-2, \alpha/2}$ es el percentil de la distribución t con $n - 2$ grados de libertad que deja a su derecha un área $\alpha/2$.

Ejemplo (continuación)

Para nuestro ejemplo, obtenemos los siguiente intervalos de confianza para la media de la disminución de pulsaciones de poblaciones de ratas con distintos valores de dosis.

Dosis (x_0)	$SE(\hat{\mu}_{x_0})$	Intervalo de Confianza 95%	Longitud
0.5	0.6164	(3.44, 7.36)	3.92
1.0	0.4359	(6.71, 9.49)	2.78
1.5	0.3559	(9.67, 11.93)	2.26
2.0	0.4359	(12.11, 14.89)	2.78
2.5	0.6164	(14.24, 18.16)	3.92

- Notar el cambio de amplitud.
- ¿Es correcto hacer estimaciones para valores de X distintos de aquellos observados en nuestros datos? Por ejemplo para $X = 0.85$?
- ¿Es correcto hacer estimaciones para valores de X fuera del rango de los valores de X de nuestros datos? Por ejemplo para $X = 3.7$?

PREDICCIÓN DE UN NUEVO VALOR DE Y PARA UN DADO X.

¿Si repitiéramos el experimento en iguales condiciones, cuál sería la DFC de una rata tomada al azar que recibiera una dosis x_0 ?

Antes queríamos estimar la media de la hipotética población de ratas que recibieron dosis x_0 . Ahora queremos predecir el valor en UNA RATA tomada al azar de esa población.

Predicción: Una regla para calcular a partir de los datos un valor que nos permita “adivinar” el valor que puede tomar una VARIABLE ALEATORIA.

Estimación: Una regla para calcular a partir de los datos un valor que nos permita “adivinar” el valor que puede tomar un PARÁMETRO POBLACIONAL.

Nuestra mejor predicción es nuevamente $\hat{\mu}_{x_0} = \hat{Y}_{x_0} = \hat{\beta}_0 + \hat{\beta}_1 x_0$, pero ahora el error asociado será mayor. Estimamos el ERROR ESTÁNDAR de la PREDICCIÓN con

$$s_e \sqrt{1 + \frac{1}{n} + \frac{(x_0 - \bar{x})^2}{\sum_{i=1}^n (x_i - \bar{x})^2}}$$

A partir de este error estándar podemos construir un INTERVALO de CONFIANZA de nivel $(1 - \alpha)$ para el valor predicho de Y cuando $X = x_0$.

Ejemplo (continuación)

El intervalo de confianza del 95% para predecir la DFC de una rata que recibió una dosis de 2.0 mg es

$$\hat{Y}_{2.0} \pm 3.182 \cdot 0.79582 \sqrt{1 + \frac{1}{5} + \frac{(2 - 1.5)^2}{2.50}} = 13.5 \pm 2.887 = (10.61, 16.39)$$

STATISTIX

Datos: Y X

Statistics / Linear Model / Linear Regression

Dependent Variable => Y

Independent Variable => X

En la pantalla de resultados de la regresión:

Results / Prediction

Prediction Value => 2.0 (elegimos 1 de los valores)

PREDICTED/FITTED VALUES OF Y

LOWER PREDICTED BOUND	10.612	LOWER FITTED BOUND	12.113
PREDICTED VALUE	13.500	FITTED VALUE	13.500
UPPER PREDICTED BOUND	16.388	UPPER FITTED BOUND	14.887
SE (PREDICTED VALUE)	0.9074	SE (FITTED VALUE)	0.4359

UNUSUALNESS (LEVERAGE)	0.3000
PERCENT COVERAGE	95.0
CORRESPONDING T	3.18

PREDICTOR VALUES: X = 2.0000

De donde:

- IC 95% para la media de Y cuando X = 2.0 es (12.1, 14.9) [fitted value]
- IC 95% para un nuevo valor de Y cuando X = 2.0 es (10.6, 16.4) [predicted value]

REGRESIÓN EN ESTUDIOS OBSERVACIONALES

A menudo el investigador no selecciona los valores de la variable independiente, sino que toma una muestra de alguna población y observa simultáneamente X e Y para cada miembro de la muestra.

¿Son válidos los resultados que hemos derivado suponiendo que los valores de X eran fijos?

- ♦ Estimadores de β_0 y β_1

- ◆ Estimadores de σ^2 , $\sigma^2(\hat{\beta}_0)$, $\sigma^2(\hat{\beta}_1)$
- ◆ Intervalos de confianza y test para β_0 y β_1
- ◆ Estimadores e intervalos de confianza para μ_x
- ◆ Predicción para Y cuando $X = x$ y su intervalo de confianza.

SI, SON VÁLIDOS los resultados si se cumple que para cada valor de X:

- ❖ Y tiene distribución normal con media μ_x y varianza σ^2 (NORMALIDAD)
- ❖ La media μ_x es una función lineal de X (LINEALIDAD)
- ❖ La varianza σ^2 es la misma para todo nivel de X.

ESTUDIOS OBSERVACIONALES vs EXPERIMENTALES

Consideremos los datos de dosis-respuesta en ratas del estudio experimental de nuestro ejemplo y los datos de un estudio observacional en el que se registra la tasa de nacimiento y la actividad económica femenina en 26 países, y se pretende predecir la tasa de nacimientos en función de la actividad económica femenina.

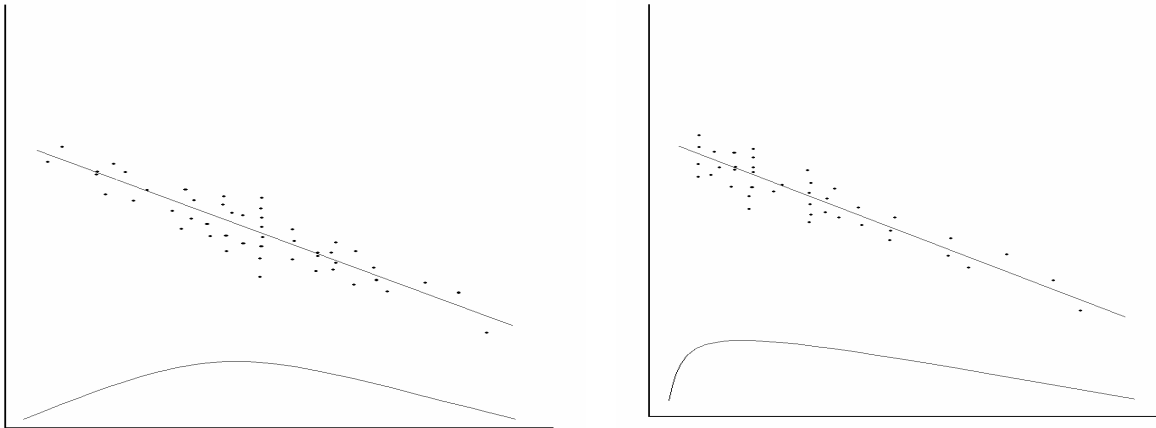
Experimento Controlado	Estudio observacional
El investigador asignó ALEATORIAMENTE las ratas a las dosis, y se preocupó de manejar de manera similar las distintas ratas.	El investigador NO tiene control del nivel de actividad económica de las mujeres en los distintos países. Países con alto índice de actividad económica femenina pueden diferir de los otros países en factores medidos y no medidos asociados con la tasa de natalidad.
Cualquier cambio estimado en la FC puede ser razonablemente atribuido al incremento en la dosis de la droga.	No es posible hacer una interpretación causal.
β_1 = cambio en la FC debido a un incremento de una unidad en la dosis.	β_1 = cambio promedio en la tasa de natalidad (B) que acompaña un incremento de una unidad en el índice de actividad económica femenina (W)

¡ β_1 no tiene interpretación CAUSAL si los datos provienen de un estudio observacional!

Supongamos que en el estudio observacional los países se hubieran seleccionado de modo tal que *la probabilidad de que un país sea seleccionado depende de la tasa de actividad femenina (W)*.

¿Sería válida nuestra inferencia acerca de β_1 en la regresión de la tasa de nacimiento respecto de la tasa de actividad femenina? **Si!!!**

Consideremos los gráficos siguientes.



¿Qué ocurre si la probabilidad de ser seleccionado depende del valor de la variable dependiente? Entonces la inferencia acerca de β_1 en el modelo $Y = \beta_0 + \beta_1 X + \varepsilon$ podría NO SER VÁLIDA!!!! En particular, $\hat{\beta}_1$ podría ser un estimador sesgado de β_1 .

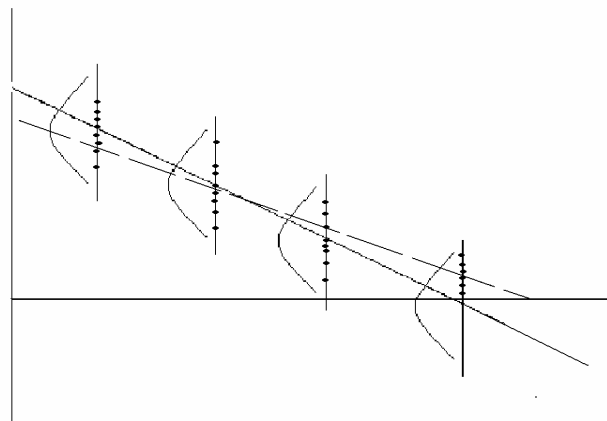
Para entender ésto consideremos los siguientes ejemplos:

Ejemplo 1. Sólo se seleccionan países con tasa de natalidad igual a un valor dado. Entonces $\hat{\beta}_1 = 0!!!!$

Ejemplo 2. Se toma una muestra aleatoria de países, pero se descartan aquellos en los que la tasa de natalidad sea menor que un cierto valor.

En resumen, para un modelo $Y = \beta_0 + \beta_1 X + \varepsilon$

Muestra con probabilidad que depende de	Inferencia acerca de β_1
X	VÁLIDA
Y	NO VÁLIDA



REGRESIÓN A TRAVÉS DEL ORIGEN

Consideremos los siguientes datos sobre el número de derrames accidentales de petróleo en el mar (X) y la cantidad de petróleo derramado (en millones de metros cúbicos, Y).

Año	# de derrames (X)	Cantidad de petróleo derramada (Y)
1973	36	84.5
1974	48	67.1
1975	45	188.0
1976	29	204.2
1977	49	213.1
1978	35	260.5
1979	65	723.5
1980	32	135.6
1981	33	45.3
1982	9	1.7
1983	17	387.8
1984	15	24.2
1985	8	15.0

La recta de regresión ajustada es

$$\hat{Y} = -44.487 + 6.957x$$

Nuestra estimación nos dice que si no ocurrieran derrames habría “44 millones recuperados”. Sabemos, que si no hay derrames ($X = 0$), la cantidad derramada debe ser cero ($Y = 0$). Parece razonable ajustar el modelo

$$Y_i = \beta_1 X_i + \varepsilon$$

donde imponemos $\beta_0 = 0$.

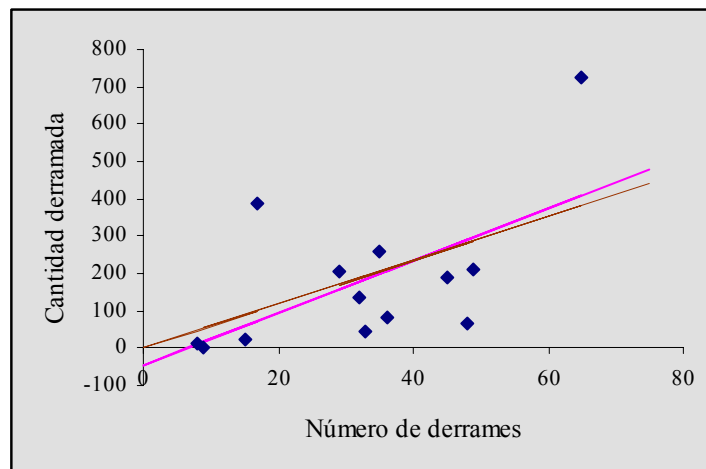
El procedimiento de mínimos cuadrados para el problema $\beta_0 = 0$ busca entre todas las rectas que pasan por el origen, aquella que minimice la suma de los residuos al cuadrado. Puede mostrarse que esta recta tiene pendiente

$$\hat{\beta}_1 = \frac{\sum_{i=1}^n X_i Y_i}{\sum_{i=1}^n X_i^2}.$$

Notar que las observaciones no se ajustan con la media en este caso.

Para los datos de petróleo se obtiene la recta

$$\hat{Y}_i = 5.861 x_i$$



Nota. Para hacer regresión por el origen en Statistix:

Statistics / Linear Model / Linear Regression

Dependent Variable => Y

Independent Variable => X

Desmarcar FIT CONSTANT

UNWEIGHTED LEAST SQUARES LINEAR REGRESSION OF CANTIDAD

NOTE: MODEL FORCED THROUGH ORIGIN

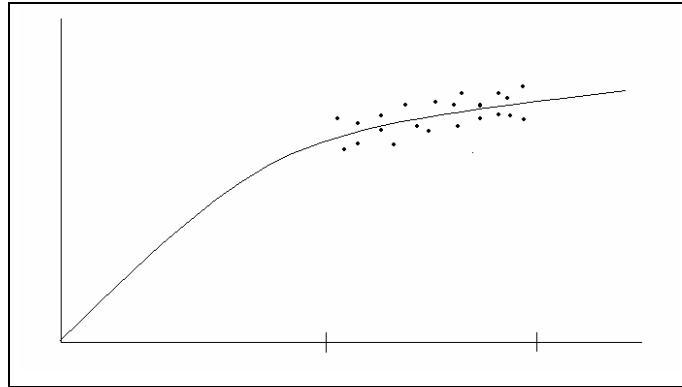
PREDICTOR VARIABLES	COEFFICIENT	STD ERROR	STUDENT'S T	P
-----	-----	-----	-----	-----
NUMERO	5.86088	1.22960	4.77	0.0005
R-SQUARED	0.6544	RESID. MEAN SQUARE (MSE)		25837.1
ADJUSTED R-SQUARED	0.6256	STANDARD DEVIATION		160.739

SOURCE	DF	SS	MS	F	P
-----	---	-----	-----	-----	-----
REGRESSION	1	587005	587005	22.72	0.0005
RESIDUAL	12	310045	25837.1		
TOTAL	13	897050			

CASES INCLUDED 13 MISSING CASES 0

¿Qué estimación de β_1 es preferible?

- ♦ Si el modelo lineal es verdadero y además $\beta_0 = 0$ entonces, la regresión por el origen resulta en un ajuste con menos error estándar para β_1 .
- ♦ Si el modelo lineal no es apropiado, aún cuando sea aproximadamente válido en el rango de valores observados de X, y se ajusta una regresión por el origen se obtendrá una estimación sesgada de β_1 .



Si nosotros sabemos que cuando $X = 0$ entonces $Y = 0$ y ajustamos el modelo completo

$$Y_i = \beta_0 + \beta_1 X_i + \varepsilon.$$

Si el test para

$$H_0 : \beta_0 = 0$$

rechaza la hipótesis nula, tenemos evidencia de que el supuesto de linealidad **NO ES VÁLIDO!!!**

Por lo tanto, cuando se sabe que la curva debe pasar por el origen, el test para $\beta_0 = 0$ es de interés para chequear apartamientos de la linealidad.

Nota. Otra situación en que β_0 podría ser conocido, aunque distinto de cero, sería el caso en que se sabe que el costo fijo diario de un servicio es β_0 y que por cada paciente que ingresa el costo se incrementa en cierta cantidad β_1 . En este caso el estimador de la pendiente es ligeramente diferente.

PROBLEMAS CON REGRESIÓN

Los siguientes son algunos problemas comunes que afectan la validez de las conclusiones de un Análisis de Regresión.

- **Relación no lineal.** El método de mínimos cuadrados que hemos presentado busca el mejor ajuste a una línea recta. Si la relación entre las variables no es lineal, la distribución propuesta para los estimadores es falsa.
- **Varianza no homogénea.** Si la varianza de los errores cambia con los valores de X (heteroscedasticidad) , entonces los errores estándares, los tests y los intervalos de confianza que hemos obtenido son inapropiados.
- **Errores correlacionados.** Los errores estándares que hemos calculado, así como los tests y los intervalos de confianza suponen que los errores son independientes. Cuando este supuesto se viola las conclusiones pueden ser erróneas.
- **Errores no normales.** Los tests e intervalos de confianza que hemos presentado, basados en la distribución t y la F, suponen que para cada valor de X la distribución de la variable Y es normal. El no cumplimiento de este supuesto invalida estos procedimientos (especialmente cuando el tamaño de muestra es pequeño).
- **Casos influyentes.** Los estimadores de mínimos cuadrados son muy poco robustos. Un único datos outlier puede modificar sustancialmente la recta estimada.
- **Variables omitidas.** Si otras variables afectan a ambas X e Y simultáneamente nuestra estimación de la pendiente puede subestimar o sobrestimar la verdadera relación entre X e Y (Confounding factors).

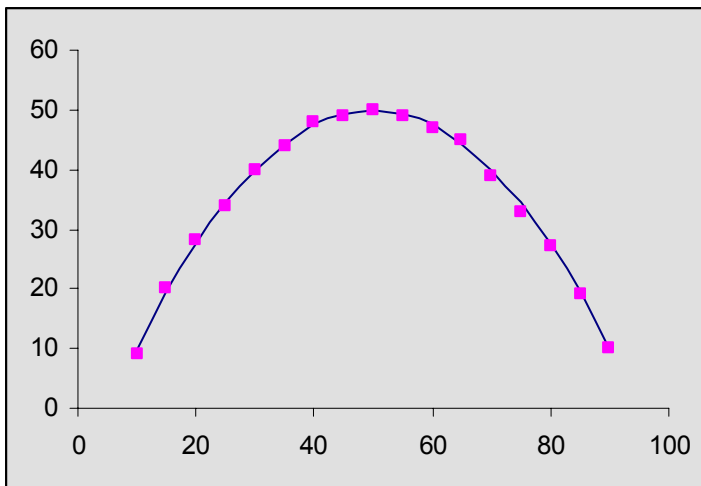
Comentaremos brevemente algunos de ellos.

Relación no lineal.

El modelo supone que la media poblacional de Y varía linealmente con X. La forma REAL de la relación es desconocida, y es muy poco probable que sea EXACTAMENTE lineal. Sin embargo, una función lineal a menudo resulta ser una buena aproximación para la verdadera relación.

La relación de talla con edad en niños es aproximadamente lineal cuando se consideran tramos cortos de edad.

Si el supuesto de linealidad claramente no se cumple, como por ejemplo en una relación en forma de U, como la del gráfico siguiente, los resultados y conclusiones del análisis de regresión pueden ser absolutamente engañosos. Por ejemplo, el test para la pendiente puede no rechazar la hipótesis de que la pendiente es cero, es decir concluiríamos que las variables NO están asociadas, las variables están fuertemente asociadas ya que la media de Y sigue una relación casi determinística con el valor de X, el problema es que esta relación no es lineal.



$$r = -0.0117$$

Recta ajustada

$$Y = 35.08 - 0.00637 X$$

Cuando no se cumple el supuesto de linealidad la hipótesis $H_0: \beta_1 = 0$ no es un test de independencia entre las variables.

Varianza no homogénea y errores no normales.

Buena noticia!!!

La recta estimada por mínimos cuadrados y los coeficientes de correlación y determinación son modos válidos para DESCRIBIR la relación entre las variables aún cuando el supuesto de normalidad no sea válido.

Lo que no vale en este caso es la inferencia.

En la práctica los supuestos de normalidad y homoscedasticidad nunca se cumplen exactamente. Sin embargo, mientras más cerca estén nuestros datos de

los supuestos del modelo lineal, más apropiados serán los tests e intervalos de confianza que construyamos.

Para muestras grandes el supuesto de distribución normal no es crucial. Una versión extendida del Teorema Central del Límite dice que el estimador de mínimos cuadrados de la pendiente tiene distribución de muestreo aproximadamente normal cuando n es grande.

Observaciones influyentes

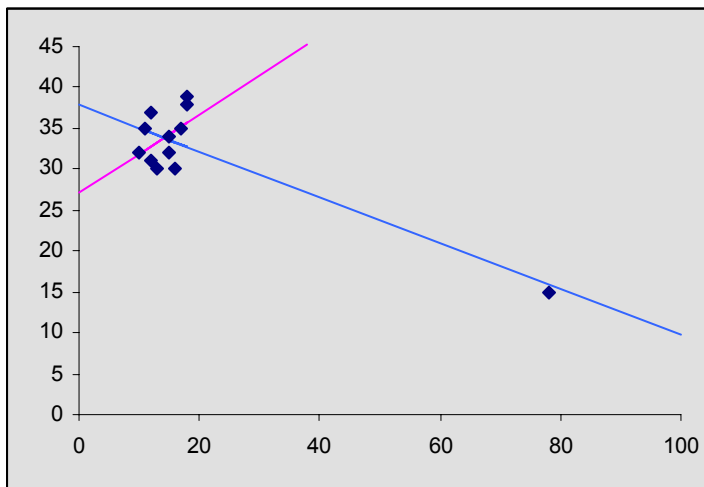
Una desventaja del método de cuadrados mínimos es que observaciones con X muy grande o muy pequeño que caigan lejos de la tendencia del resto de los datos pueden modificar sustancialmente la estimación.

Una observación se denomina INFLUYENTE si al excluirla la recta de regresión estimada cambia notablemente.

Veamos un ejemplo. Las variables de interés son (ambas tomadas en 1995) para distintos países:

Y = Tasa de nacimiento (cada 1000 habitantes) en 1995

X = Número de televisores (cada 100 habitantes)



— Sin el dato influyente
 — Con el dato influyente

Coeficiente de correlación de Pearson		Ecuación estimada
Sin dato influyente	r = 0.4255	Y = 27.0964 + 0.47732 X
Con dato influyente	r = -0.8352	Y = 37.8245 – 0.28040 X

El dato influyente es USA, el resto son países de Asia y Africa.

Un dato influyente puede ser fácilmente detectado a través de métodos gráficos, aunque también existen medidas indicadoras de cuán influyente es una observación. Las estudiaremos cuando veamos regresión múltiple.

Cuando detectamos un outlier tan severo, es importante investigarlo. Puede estar mal registrado. Si es correcto, quizás sea diferente de las otras del modo como ocurre en el ejemplo anterior y hay que preguntarse si interesa mantenerlo en el análisis.

Si el dato es correcto y no hay razones para excluirlo del análisis entonces la estimación de los parámetros debería hacerse con un método robusto.

DIAGNÓSTICO EN REGRESIÓN

¿Cómo sabemos si el modelo ajusta razonablemente bien a nuestros datos?

Comentaremos brevemente la metodología para diagnosticar:

- (1) Si hay apartamientos definidos de los supuestos del modelo
- (2) Si hay observaciones influyentes.

Recordemos nuevamente los **supuestos** de nuestro modelo lineal

- . Linealidad.
- . Distribución normal de la variable Y condicional a X.
- . Homoscedasticidad.
- . Independencia de los errores.

En la práctica es imposible verificar que los supuestos se cumplen. La idea es averiguar si existen apartamientos groseros de alguno de ellos, si no observamos apartamientos groseros entonces el modelo puede todavía ser útil.

Análisis de los residuos

La mayoría de los supuestos puede chequearse usando los residuos $(Y_i - \hat{Y}_i)$. Ellos representan la distancia de cada observación a la recta ajustada.

Normalidad. Si las observaciones provienen de distribuciones normales todas con la misma varianza σ , entonces los residuos deberían mostrar una distribución aproximadamente normal. Para chequearlo podemos usar métodos gráficos o el test de Shapiro-Wilk.

Métodos gráficos: Histograma / box-plot (box) / gráfico de probabilidad normal (normal probability plot)

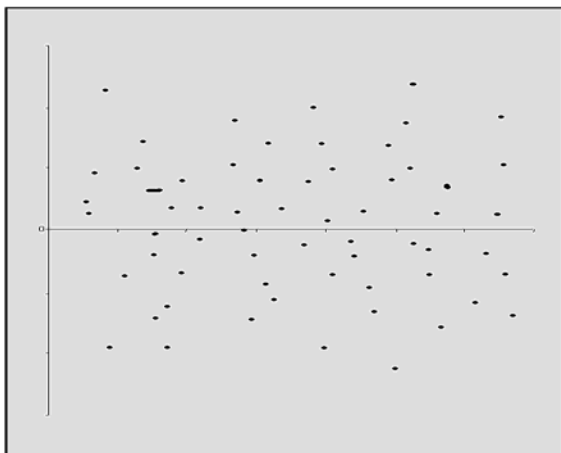
Si se detectan observaciones outliers, controlar si son correctas!!! Si lo son, investigar si son influyentes.

Linealidad y homoscedasticidad

Para chequear que el modelo lineal es una buena aproximación a la verdadera relación entre las variables y para chequear el supuesto de homogeneidad de varianzas usamos el gráfico de residuos versus valores predichos.

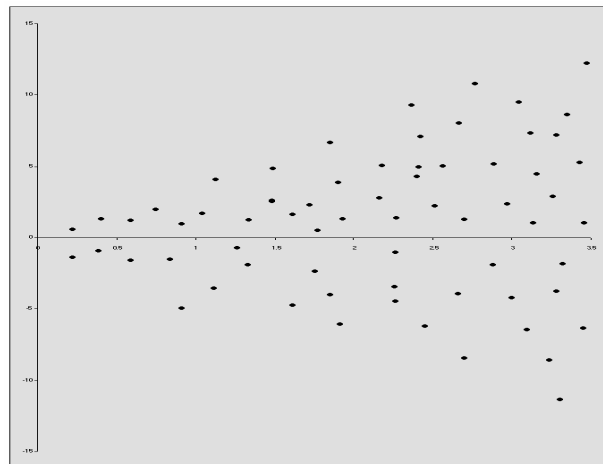
Si el gráfico muestra una nube de puntos alrededor de cero sin evidencia de estructura, tendencia o cambio de la dispersión, entonces no hay sospecha de que se violen ninguno de estos dos supuestos.

Se satisfacen los supuestos

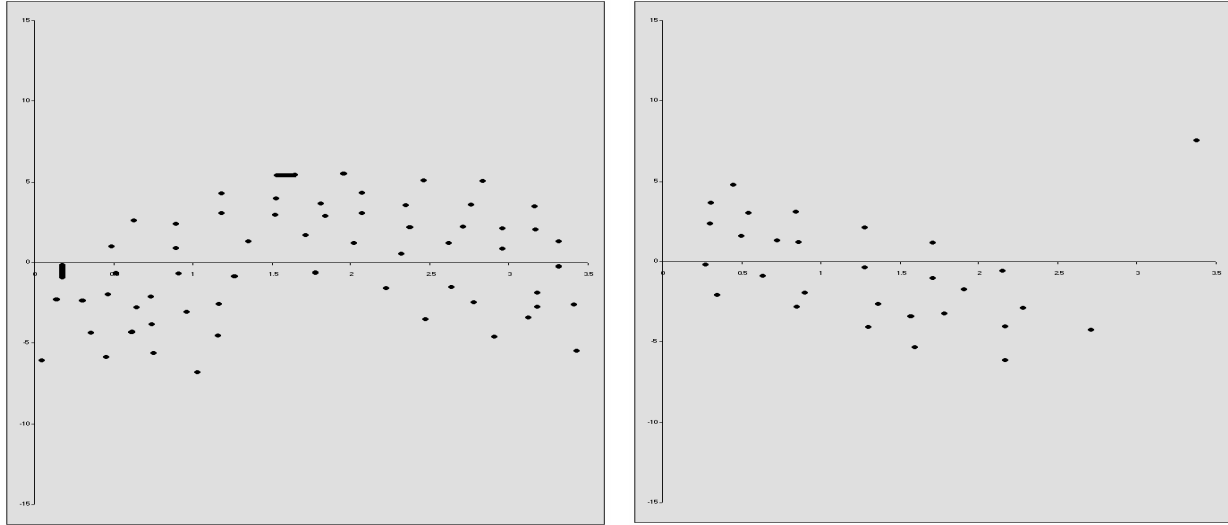


Relación no lineal

Heteroscedasticidad



Observación influyente



En general, los datos influyentes no son fácilmente detectables como outliers en gráficos de residuos vs predichos, porque ejercen “palanca” sobre la recta forzándola a pasar relativamente cerca de ellos. En consecuencia, pueden presentar residuos que no llaman la atención por su magnitud.

En la práctica los gráficos de residuos no son tan claros como estos... Recordar que aún cuando los datos satisficieran todos los supuestos, la variabilidad muestral podría hacer que el gráfico tuviera pequeños apartamientos de la imagen ideal.

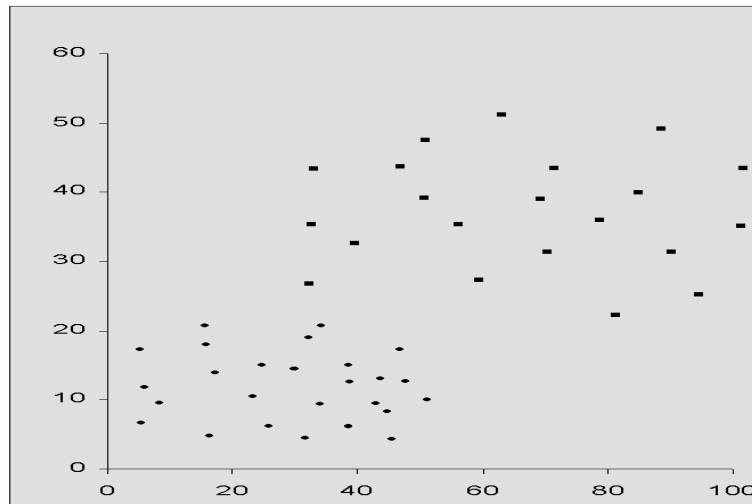
Independencia

El hecho de haber tomado una muestra aleatoria de sujetos desde alguna población asegura que, en principio, tendremos observaciones independientes. Algunas situaciones en las que este supuesto puede fallar se describen a continuación.

Estudios los datos se recolectan secuencialmente. Las observaciones consecutivas pueden no ser independientes. Determinaciones de laboratorio hechas secuencialmente en el tiempo pueden mostrar un cierto patrón, dependiendo de cómo funcionan los equipos, los observadores, etc. Modo de detección: Graficar residuos versus secuencia temporal.

Si los datos fueron obtenidos por dos observadores A y B, podríamos esperar que las observaciones de un observador tiendan a parecerse más entre ellas. Modo de detección: Gráfico de Y vs X identificando los puntos de cada grupo.

En ocasiones, la variabilidad debida a la regresión puede ser explicada por la pertenencia al grupo. Ver gráfico.



POSIBLES CONSECUENCIAS DE LA VIOLACIÓN DE LOS SUPUESTOS:

- (1) Estimación sesgada de los estimadores de los parámetros.
- (2) Estimación sesgada del error estándar de los coeficientes.
- (3) Tests e Intervalos de Confianza no válidos.
- (4) Los estimadores de mínimos cuadrados no son los eficientes (no son los estimadores de mínima varianza).

HERRAMIENTAS PARA DIAGNÓSTICO

Ejemplo.

Trabajaremos en esta sección sobre un conjunto de $n = 31$ observaciones (X, Y) donde $X =$ edad e $Y =$ años de empleo.

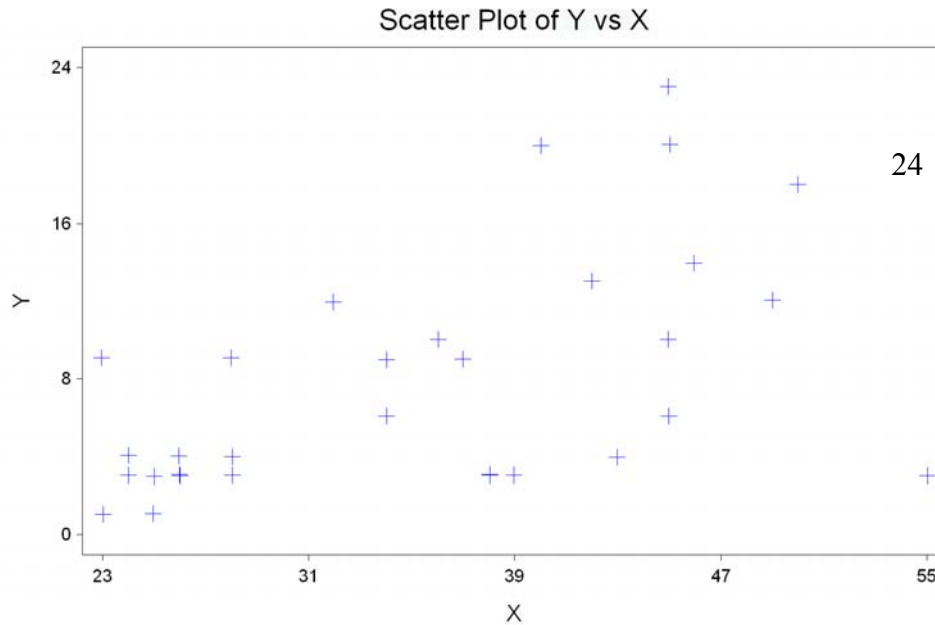
1. GRÁFICOS

En primer lugar hacemos un gráfico de las dos variables para ver si el supuesto de linealidad es adecuado.

STATISTIX

Statistics / Summary Statistics / Scatter plot

Se observa una tendencia aparentemente creciente (difícil decir si el modelo lineal es apropiado). La varianza de Y parece aumentar con X.



1.1 Análisis de la distribución de los residuos

Chequearemos los supuestos estudiando el comportamiento de los residuos. Recordemos que el residuo se define como la diferencia entre el valor observado y el valor predicho:

$$e_i = (Y_i - \hat{Y}_i)$$

Estudiamos la distribución de los residuos a través de gráficos y del test de Shapiro-Wilk

STATISTIX

Statistics / Linear Model / Linear Regression

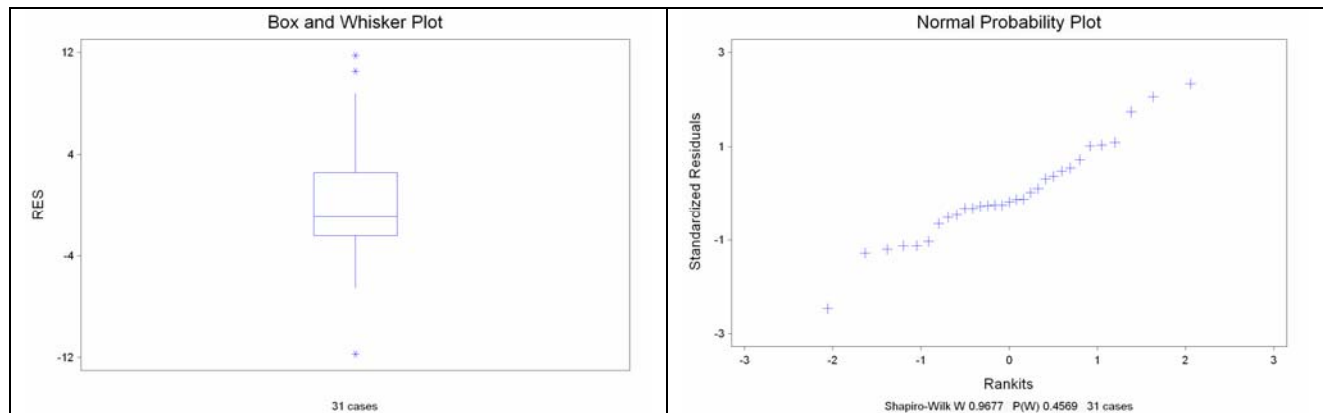
Dependent Variable => Y

Independent Variable => X

En la pantalla de resultados de la regresión:

Results / Plots / Normal Probability Plot

Save Residuals => Residual = Res Fitted = Yhat



- Box-plot. La distribución de los residuos es algo asimétrica, y muestra tres outliers.
- Normal Probability Plot. La distribución no parece muy distinta de la normal (salvo por los outliers). Test de Shapiro-Wilks $p = 0.45$ (no se rechaza la hipótesis de que los residuos provienen de una distribución normal)

1.2 Residuos versus valores predichos

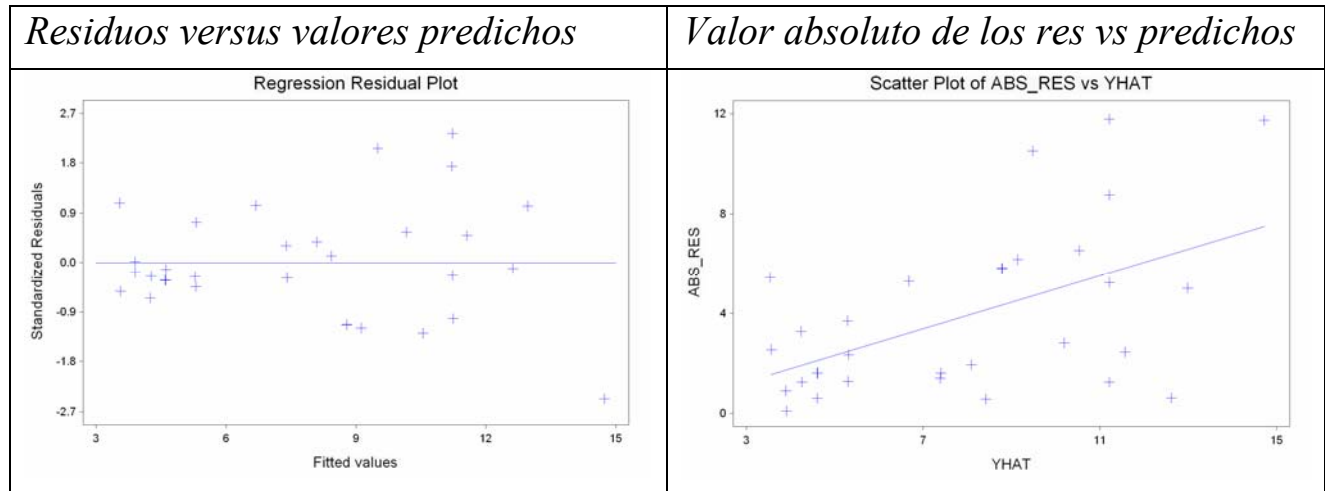
- Si los supuestos se cumplen el gráfico no debería mostrar ningún patrón, es decir deberíamos observar una nube de puntos elipsoidal.
- Hay evidencia de heterogeneidad de varianza cuando la varianza aumenta o disminuye al movernos a lo largo de los valores predichos. Imagen de “corneta” o de “rombo”.
- Los outliers aparecen como puntos muy alejados de la recta que indica el cero para los residuos.

Para graficar residuos vs predichos con Statistix:

En la pantalla de resultados de la regresión:

Results / Plots / St. Residuals vs Fitted Values

(Grafica residuos estandarizados versus los valores predichos)



¿Qué observamos?

- Heterogeneidad de varianzas.
- Hay varios datos que podrían ser outliers (8, 24, 28).
- La observación 24 es influyente.

1.3 Residuos absolutos versus predichos

Cuando se sospecha de la homogeneidad de varianzas una posibilidad es graficar los valores absolutos de los residuos (los residuos sin tener en cuenta el signo) versus los valores predichos. Además podemos graficar una curva suave que muestra los valores de las medianas de los residuos absolutos para distintos grupos de valores de los predichos. Si la curva muestra tendencia nos indica que falla el supuesto de homogeneidad de varianzas.

El gráfico indica heteroscedasticidad, ya que la magnitud del valor absoluto de los residuos aumenta cuando aumenta el valor predicho.

2. DETECCIÓN DE OBSERVACIONES INFLUYENTES

Existe una variedad de estadísticos que permiten resumir la influencia que cada observación tiene en la estimación de los parámetros. Estos estadísticos cuantifican cuánto cambian los coeficientes estimados o los valores predichos cuando la observación se elimina del conjunto de datos. La influencia de una observación depende de dos factores:

- . Cuán lejos cae el valor de Y de la tendencia general en la muestra y
- . Cuán lejos se encuentran los valores de las variables regresoras de sus medias.

Nota.

- Los métodos que hemos comentado para detectar outliers y observaciones influyentes son necesarios en un análisis de regresión. Sin embargo, no son infalibles, no existe un algoritmo automatizado para la evaluación, ni un criterio único, por lo que requieren del BUEN JUICIO del analista.
- Otro problema es que si dos casos fuertemente influyentes son casi coincidentes, cuando eliminemos uno de ellos, el ajuste prácticamente no se modificara gracias a la presencia del otro!!! Existen extensiones de estos métodos para medir influencia conjunta de los datos tomados de a dos, de a tres, etc.

¿QUÉ HACER CUANDO NUESTROS DATOS NO SATISFACEN LOS SUPUESTOS?

Si el modelo lineal no es adecuado para nuestros datos hay dos caminos posibles:

1. Abandonar el modelo de regresión lineal y usar un procedimiento más apropiado (ver tabla siguiente).
2. Emplear alguna transformación de los datos de modo que el modelo de regresión lineal sea adecuado para nuestros datos.

Cada aproximación tiene ventajas y desventajas.

- Abandonar el modelo lineal, en general significa una mejor comprensión de la naturaleza del problema, pero también implica usar procedimientos más complejo para estimar los parámetros.
- Encontrar transformaciones apropiadas para las variables implica que el método de estimación de los parámetros será simple y en general el modelo contendrá menor número de parámetros, lo que es una ventaja cuando el tamaño de muestra no es muy grande, pero las conclusiones de nuestro

análisis serán válidas para los datos transformados y no siempre es simple encontrar la transformación que solucione los problemas diagnosticados.

A continuación se presenta una Tabla que resume los métodos alternativos para los distintos problemas diagnosticados.

Problema	Remedio
Errores no normales (n pequeño), Outliers, Observaciones influyentes	- Regresión Robusta
Heterogeneidad de varianzas	- Regresión de mínimos cuadrados pesados
Relación no lineal	- Regresión no lineal - Regresión no paramétrica

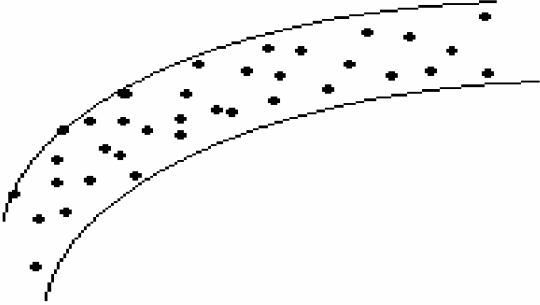
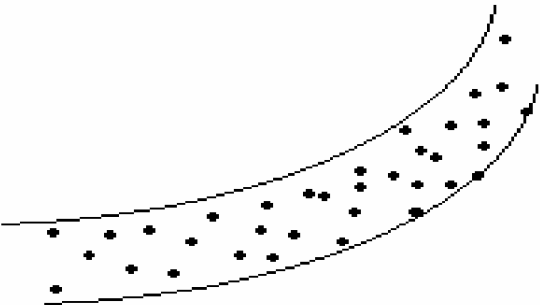
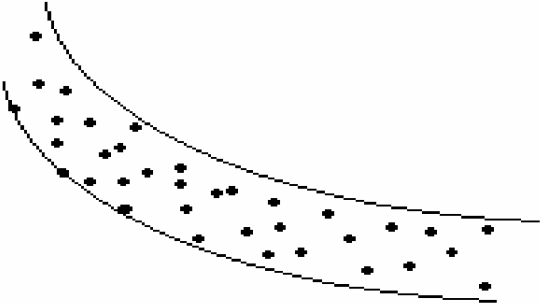
¿Qué hacer si encontramos observaciones outliers o influyentes?

- ❖ Examinar el caso para descartar que pertenezca a una población diferente, o que sea un error de registro. Si es así eliminar el caso y hacer nuevamente el análisis.
- ❖ Si el dato es “correcto”, el paso siguiente es estudiar si el modelo lineal es adecuado. Cuando hay varias variables regresoras, en ocasiones, los outliers ocurren para algunas combinaciones especiales de las variables regresoras, para las cuales el modelo puede resultar inadecuado, lo que podría sugerir la necesidad de incluir términos de interacción.
- ❖ Los datos outliers pueden además estar indicando otros tipos de no adecuación del modelo, por ejemplo la omisión de variables importantes o que la forma funcional propuesta no es correcta. El análisis de los outliers de un conjunto de datos frecuentemente conduce a profundizar el conocimiento del modelo.
- ❖ Una alternativa antes de descartar los datos outliers es usar un procedimiento de estimación diferente al de mínimos cuadrados, en el que se asigna menor peso a las observaciones outliers. Existen diferentes propuestas de este tipo conocidas bajo el nombre de REGRESIÓN ROBUSTA.

¿Qué hacer si el modelo lineal no es apropiado?

La solución es TRANSFORMAR la variable dependiente o la independiente, de modo de “linealizar” la relación. Si la distribución de los residuos es razonablemente normal y la varianza de los errores aproximadamente constante, entonces es conveniente intentar transformar la covariable. La razón es que si transformamos la Y (por ejemplo usamos \sqrt{Y}) puede cambiar sustancialmente la distribución de los errores y afectarse la homogeneidad de varianzas.

La Figura siguiente muestra algunas transformaciones típicas (llamamos X' a la variable transformada) que permiten “linealizar la relación entre X e Y.

Imagen	Transformaciones posibles para X
	$X' = \sqrt{X}$ o en general X^A con $0 < A < 1$ $X' = \log(X)$
	$X' = X^2$ o en general X^A con $A > 1$ $X' = \exp(X)$
	$X' = \frac{1}{X}$ o en general $\frac{1}{X^A}$ con $A \neq 0$ $X' = \exp(-X)$

- ❖ A veces es necesario sumar una constante a los datos para poder aplicar la transformación. Por ejemplo, si algunos datos son cercanos a 0 y se desea usar la transformación $1/X$, basta con correr el origen usando $1/(X+c)$, donde c es una constante apropiada.
- ❖ Después de seleccionar la o las transformaciones a utilizar debe ajustarse nuevamente el modelo sobre la o las variables transformadas y estudiar los residuales para decidir si el modelo resulta adecuado.

¿Qué hacer si falla el supuesto de distribución normal de los errores y/o de homogeneidad de varianzas?

Estos dos problemas frecuentemente aparecen juntos. La solución en general es una transformación de la variable Y , ya que lo que necesita modificarse es la forma de la distribución de Y y su varianza. En ocasiones esta transformación puede ayudar además a “linealizar” la relación de Y con las covariables.

Las transformación más simples son las transformaciones de potencia del tipo:

$$Y' = Y^\lambda$$

En esta familia de transformaciones están incluidas las siguientes transformaciones simples:

$\lambda = 3$	$Y' = Y^3$
$\lambda = 2$	$Y' = Y^2$
$\lambda = 1$	Datos crudos
$\lambda = 1/2$	$Y' = \sqrt{Y}$
$\lambda = 1/3$	$Y' = \sqrt[3]{Y}$
$\lambda = 0$	$Y' = \log(Y)$ por definición
$\lambda = -0.5$	$Y' = -\frac{1}{\sqrt{Y}}$
$\lambda = -1$	$Y' = \frac{1}{Y}$
$\lambda = -2$	$Y' = \frac{1}{Y^2}$

Nuevamente aquí puede ser necesario agregar constantes positivas antes de transformar porque algunas funciones no están definidas en ciertos rangos de valores. Por ejemplo, la función logaritmo o la raíz cuadrada sólo están definidas para valores mayores que cero.

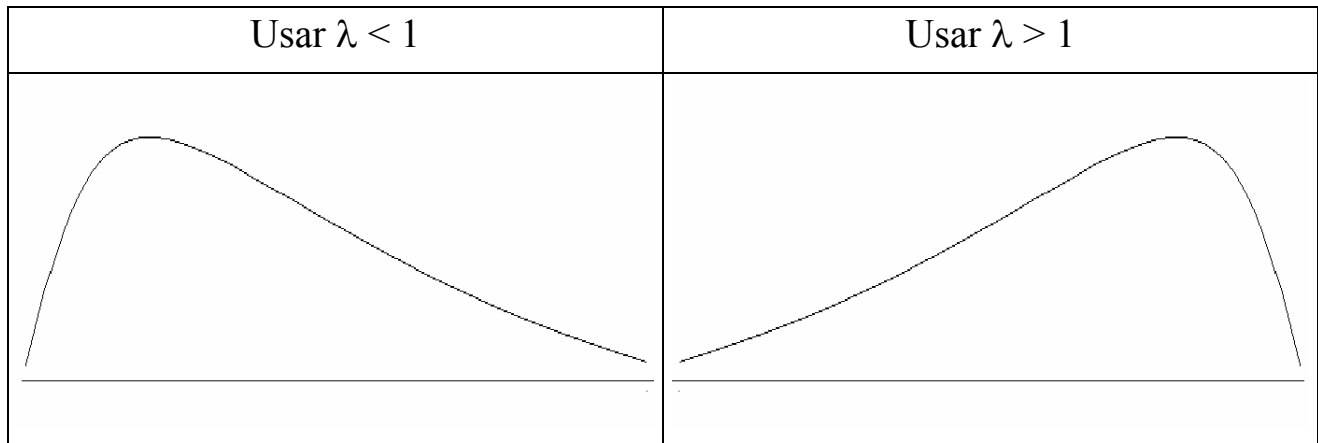
¿Cuándo usarlas?

$\lambda > 1$ sirve para distribuciones con asimetría izquierda (cola larga izquierda)

$\lambda = 1$ tengo los datos originales

$\lambda < 1$ sirve para distribuciones con asimetría derecha (cola larga derecha).

La situación más frecuente son datos con asimetría derecha, y varianza que aumenta con el valor predicho. En base al box-plot o al histograma de los residuos del modelo uno decide usar:



¿Cómo seleccionar la transformación adecuada. es decir el valor de λ ?

Ensayo y error. Existe un procedimiento automático para seleccionar la potencia de la transformación, el MÉTODO DE BOX-COX. Este método estima el parámetro λ de la transformación más apropiada para la regresión de la variable Y versus X_1, \dots, X_k . El procedimiento devuelve un intervalo de confianza para λ , de modo que el investigador elige dentro de ese intervalo un valor para λ que conduzca a una transformación de la variable Y , que en lo posible sea de simple interpretación.

Como en el caso de transformaciones de las variables regresoras, una vez que se ha seleccionado una transformación, debemos ajustar el nuevo modelo y chequear la validez de los supuestos.