

Modelo de regresión lineal simple

Prof. Agustina Gioria
PhD. Fabian Muñoz
Universidad Católica de Córdoba

Temario

- Antecedentes
- Conceptos básicos de correlación
- Conceptos básicos de regresión
- Procedimientos para el cálculo de los coeficientes de regresión y correlación.
- Ejemplos de aplicación



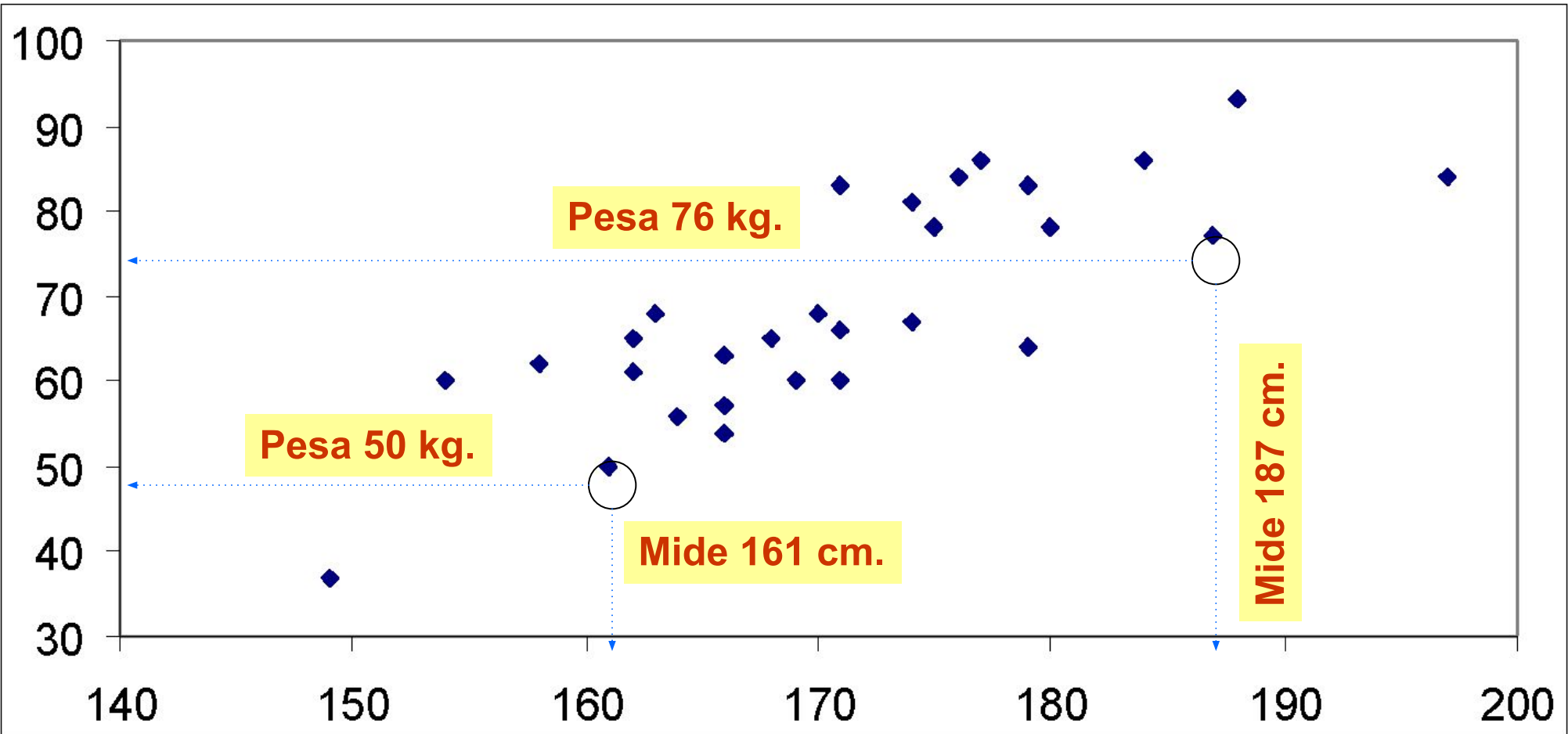
¿Qué es la regresión?

El término regresión fue acuñado por Francis Galton en el siglo XIX para referirse a fenómenos biológicos: los descendientes de progenitores excepcionales son, en promedio, menos excepcionales que los progenitores, y más parecidos a sus ancestros más distantes (Galton utilizó el término reversión al hablar de guisantes en 1877, y regression al referirse a la altura de humanos en 1885)



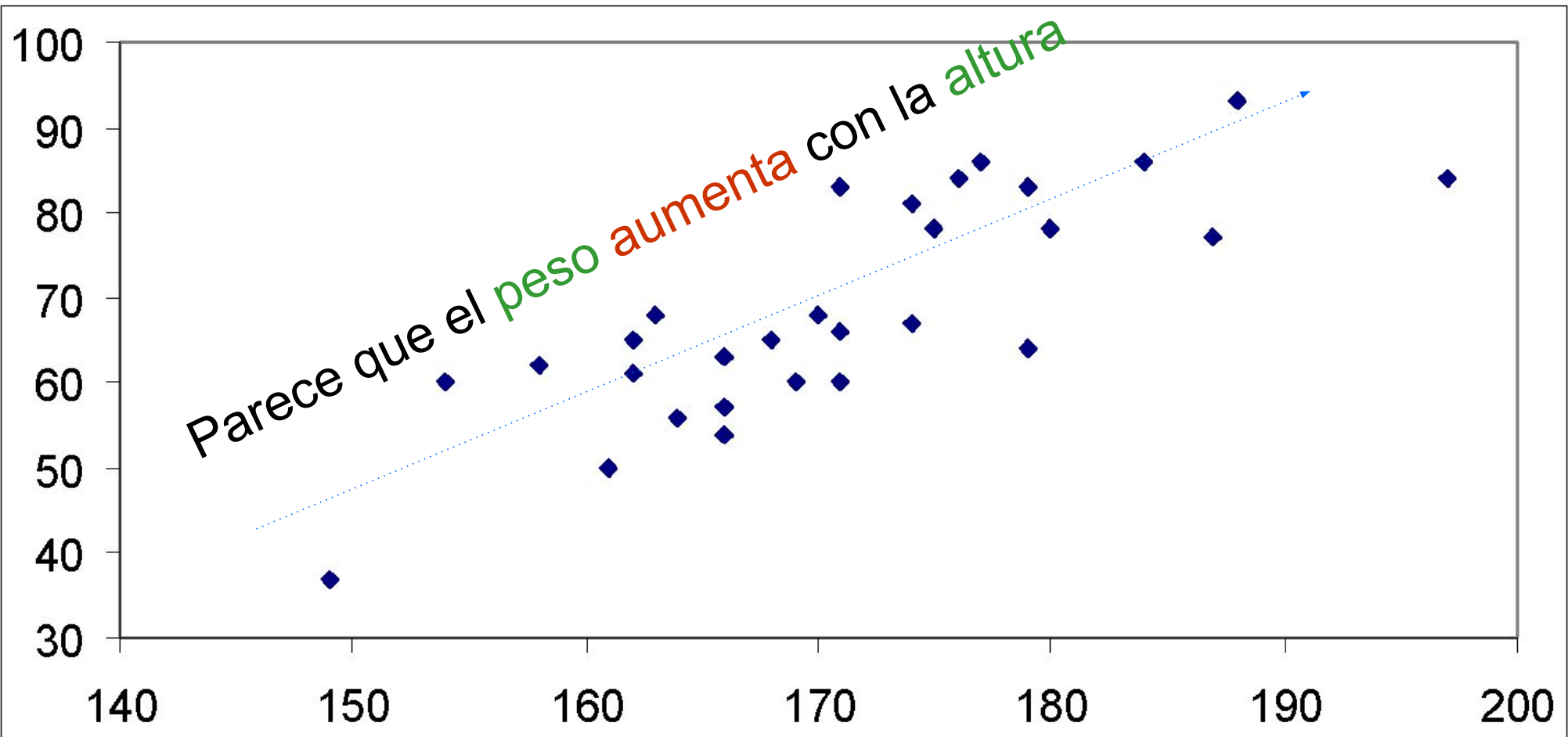
Diagramas de dispersión o nube de puntos

Tenemos las alturas y los pesos de 30 individuos representados en un diagrama de dispersión.



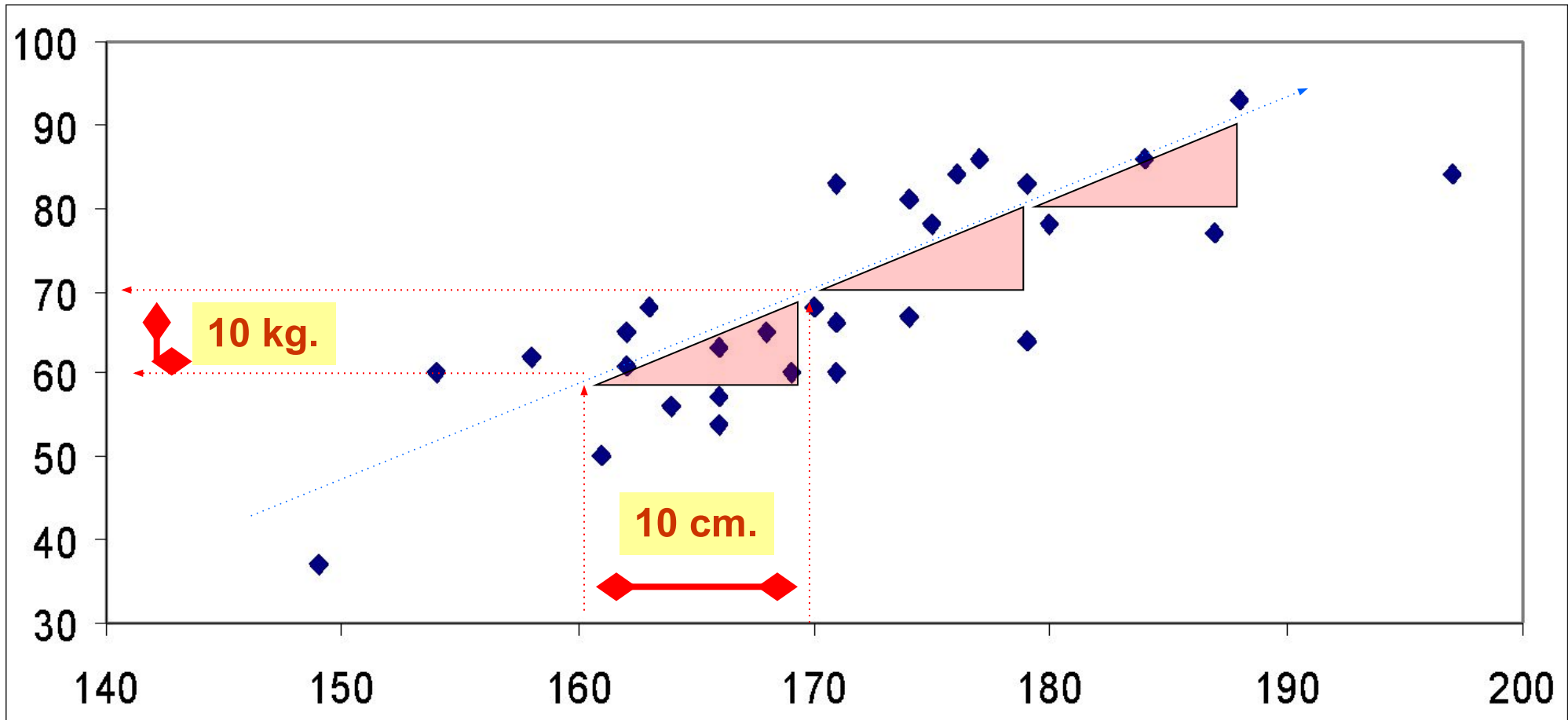
Relación entre variables.

Tenemos las alturas y los pesos de 30 individuos representados en un diagrama de dispersión.



Predicción de una variable en función de la otra.

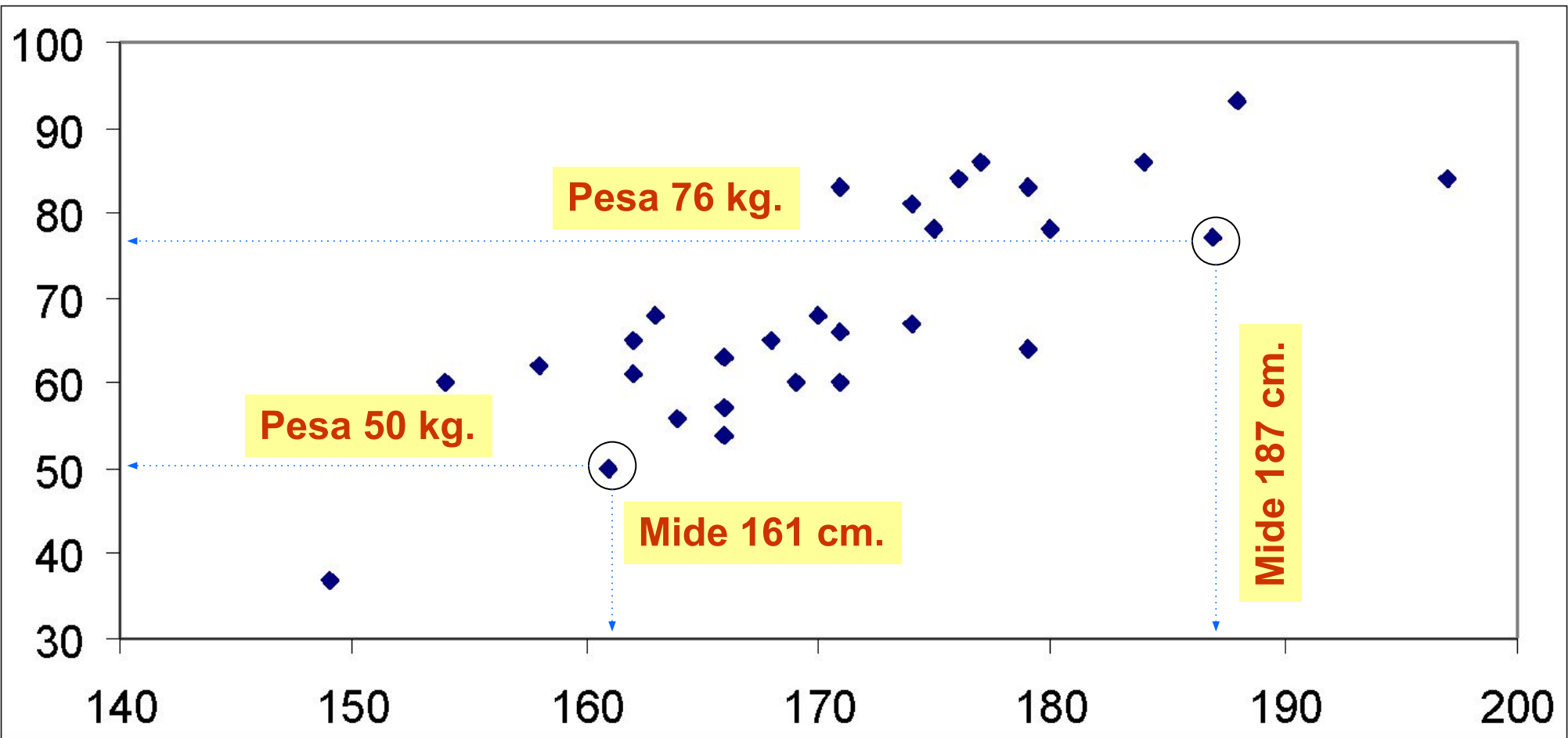
Aparentemente el peso aumenta 10Kg por cada 10 cm de altura... o sea, el peso aumenta en una unidad por cada unidad de altura.



- El objetivo es trazar una línea, que mejor describa la relación entre X y Y .
- Se puede trazar una línea con una regla, que una los puntos, pero es improbable que obtengamos una misma línea y cada una de ellas, da diferente descripción de la relación entre X y Y .

Diagramas de dispersión o nube de puntos

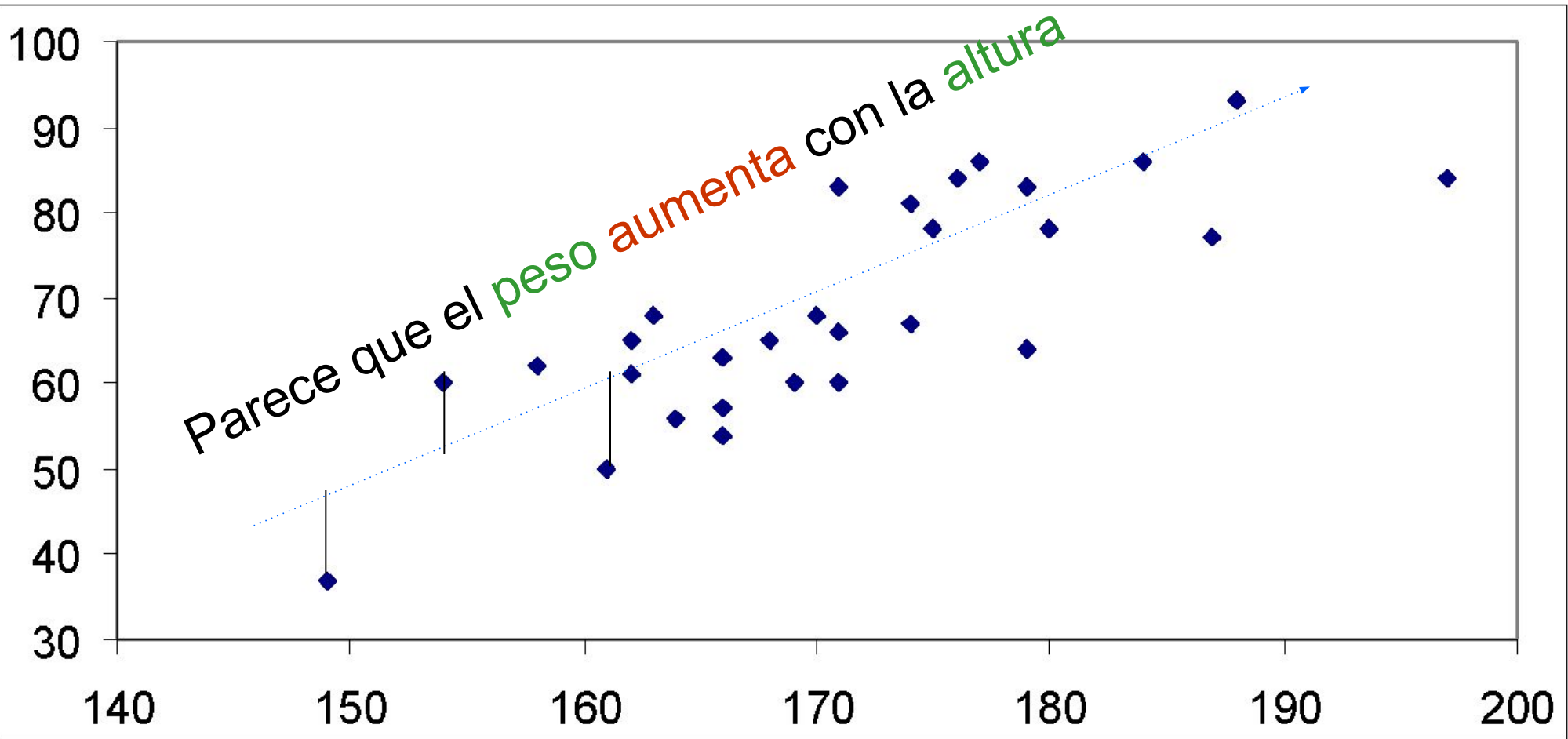
Tenemos las alturas y los pesos de 30 individuos representados en un diagrama de dispersión.




- Cada distancia vertical es la diferencia entre el valor observado para la variable dependiente (en el eje y) y el valor de la línea trazada para el correspondiente valor del eje x.
- La distancia vertical entre los valores observados y los trazados es conocida como residual. Llamamos a cada uno de los residuales e_1 .

Relación entre variables.

Tenemos las alturas y los pesos de 30 individuos representados en un diagrama de dispersión.



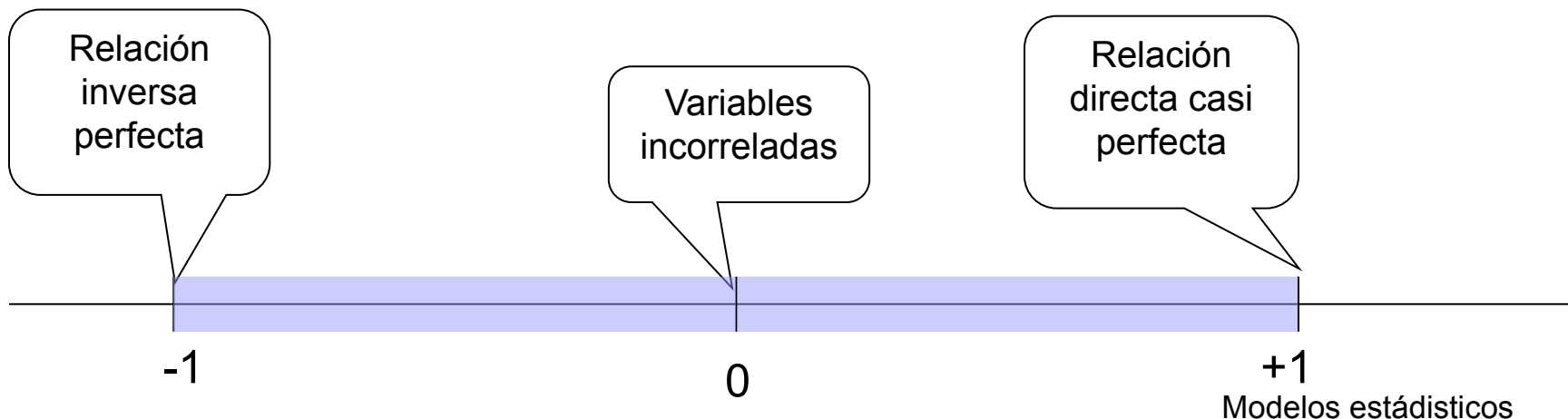
- 
- La línea que mejor traza los datos se le conoce como línea de regresión.
 - Da una estimación del valor promedio de y por algún valor de x . En general decimos que es una regresión de y sobre x .
 - Se puede pensar en la línea de regresión como una línea que une los valores medios de y por cada valor de x .

Coef. de correlación lineal de Pearson

- La coeficiente de correlación lineal de Pearson de dos variables, r , nos indica si los puntos tienen una tendencia a disponerse alineadamente (excluyendo rectas horizontales y verticales).
- El signo de correlación nos dice si el aspecto de la nube de puntos es creciente o no, pero no nos dice nada sobre el grado de relación entre las variables..
- r es útil para determinar si hay relación lineal entre dos variables, pero no servirá para otro tipo de relaciones (cuadrática, logarítmica,...)

Propiedades de r

- Es adimensional
- Sólo toma valores en $[-1,1]$
- Las variables son incorrelacionadas $\Leftrightarrow r=0$
- Relación lineal perfecta entre dos variables $\Leftrightarrow r=+1$ o $r=-1$
- Cuanto más cerca esté r de +1 o -1 mejor será el grado de relación lineal.
 - Siempre que no existan observaciones anómalas.



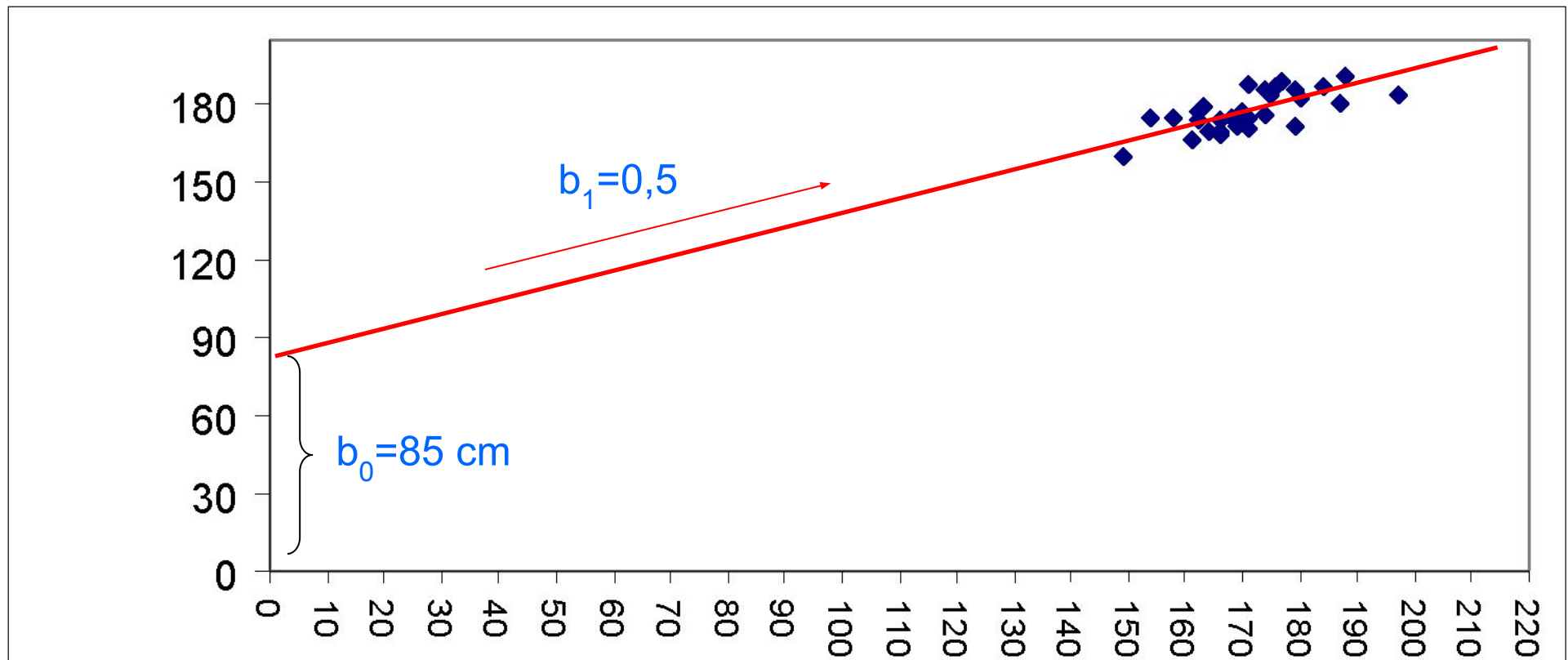
Regresión

- El análisis de regresión sirve para predecir una medida en función de otra medida (o varias).
 - Y = Variable dependiente
 - predicha
 - explicada
 - X = Variable independiente
 - predictora
 - explicativa
 - ¿Es posible descubrir una relación?
 - $Y = f(X) + \text{error}$
 - f es una función de un tipo determinado
 - el error es aleatorio, pequeño, y no depende de X

- En el ejemplo de Pearson y las alturas, él encontró:

- $\hat{Y} = b_0 + b_1 X$

- $b_0 = 85$ cm (No interpretar como altura de un hijo cuyo padre mide 0 cm ¡Extrapolación no plausible!
 - $b_1 = 0,5$ (En media el hijo gana 0,5 cm por cada cm del padre.)



- El ejemplo del estudio de la altura en grupos familiares de Pearson es del tipo que desarrollaremos en el resto del tema.
 - Altura del hijo = 85cm + **0,5** altura del padre ($Y = 85 + 0,5 X$)
 - Si el padre mide 200cm ¿cuánto mide el hijo?
 - Se espera (predice) $85 + 0,5 \times 200 = 185$ cm.
 - Alto, pero no tanto como el padre. Regresa a la media.
 - Si el padre mide 120cm ¿cuánto mide el hijo?
 - Se espera (predice) $85 + 0,5 \times 120 = 145$ cm.
 - Bajo, pero no tanto como el padre. Regresa a la media.- Es decir, nos interesaremos por **modelos de regresión lineal simple**.


Modelo de regresión lineal simple

- En el modelo de **regresión lineal simple**, dado dos variables
 - Y (dependiente)
 - X (independiente, explicativa)
- buscamos encontrar una función de X *muy simple (lineal)* que nos permita aproximar Y mediante
 - $\hat{Y} = b_0 + b_1X$
 - b_0 (ordenada en el origen, constante)
 - b_1 (pendiente de la recta)
- Y e \hat{Y} rara vez coincidirán por muy bueno que sea el modelo de regresión. A la cantidad
 - $e = Y - \hat{Y}$ se le denomina **residuo** o **error residual**.

FORMALIZACIÓN MODELO DE REGRESION LINEAL SIMPLE.

Para el modelo determinista $y_i = \delta + \beta x_i$, el valor esperado de Y_i es una función lineal de x_i . La generalización apropiada de este modelo supone que el valor esperado de Y_i es una función lineal de x_i . Si denotamos por $E(Y_i/X = x_i)$ a la esperanza de la variable aleatoria Y_i cuando la variable aleatoria X toma el valor específico x_i , entonces, el supuesto de linealidad implica que esta esperanza puede plantearse como:

$$E(Y_i/X = x_i) = \delta + \beta x_i,$$



En la práctica, el valor esperado de Y_i se derivara, casi inevitablemente, de su valor esperado. Si la diferencia se representa mediante la variable aleatoria ϵ_i (que tiene media cero) por lo tanto se tiene que:

$$\epsilon_i = Y_i - E(Y_i/X = x_i) = Y_i - \delta + \beta x_i$$

como se pretende que el error tenga media cero, entonces

$$Y_i = \delta + \beta x_i + \epsilon_i$$

Esta ecuación es llamada recta verdadera (o poblacional) de regresión.



Estimación de los parámetros por mínimos cuadrados.

Es una técnica de Análisis Numérico en la que, dados un conjunto de pares (o ternas, etc), se intenta encontrar la función que mejor se aproxime a los datos (un “mejor ajuste”).

En su forma más simple, intenta minimizar la suma de cuadrados de las diferencias ordenadas (llamadas residuos) entre los puntos generados por la función y los correspondientes en los datos.

Ventajas:

- ✓ Es objetivo, sólo depende de los resultados experimentales.
- ✓ Es reproducible, proporciona la misma ecuación, no importa quién realice el análisis.
- ✓ Proporciona una estimación probabilística de la ecuación que representa a unos datos experimentales.
- ✓ Proporciona intervalos pequeños de error



RESTRICCIONES

- Sólo sirve para ajustar modelos lineales
- Requiere tener, al menos, diez mediciones bajo las mismas circunstancias experimentales.
- Tales resultados deben estar descritos por una distribución de probabilidad conocida. La más común es la distribución normal o gaussiana.
- Se requiere de algún equipo de cálculo, de lo contrario, es muy engorroso.


Estimación de los parámetros por mínimos cuadrados.

Sea la ecuación de $Y_i = \delta + \beta x_i + \epsilon_i$, la estimación de los parámetros δ, β viene dado por:

$$S_{xx} = \sum_{i=1}^n (x_i - \bar{x})^2 = \sum_{i=1}^n x_i^2 - n\bar{x}^2$$

$$S_{yy} = \sum_{i=1}^n (y_i - \bar{y})^2 = \sum_{i=1}^n y_i^2 - n\bar{y}^2$$

$$S_{xy} = \sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y}) = \sum_{i=1}^n x_i y_i - n\bar{x}\bar{y}$$



Una vez realizados estos cálculos se procede a estimar los parámetros de la regresión.

$$\bar{\beta} = \frac{S_{xy}}{S_{xx}}$$

$$\bar{\delta} = \bar{y} - \hat{\beta}\bar{x}$$

Por tanto, la recta de regresión muestral (Estimada por mínimos cuadrados) es entonces aquella cuya ecuación es:

$$y = \bar{\delta} - \bar{\beta}x$$

Teorema de descomposición de la suma de cuadrados

Una ecuación de regresión puede considerarse como un intento de emplear la información proporcionada por una variable independiente X , para explicar el comportamiento de una variable dependiente Y . Como las observaciones de la variable dependiente exhibirán cierta variabilidad en la muestra. Para lograr esto, recordemos que, para los valores muestrales, la recta de regresión estimada puede escribirse como:

$$Y_i = \hat{\delta} + \hat{\beta}x_i + \epsilon_i$$

Donde $\hat{\delta}$ y $\hat{\beta}$ son las estimaciones de mínimos cuadrados del intercepto y de la pendiente de la regresión poblacional, y ϵ_i son los residuos de la recta de regresión ajustada. Sean además sean:

Suma de cuadrados total:

$$SST = \sum_{i=1}^n y_i^2 - n\bar{y}^2 = \sum_{i=1}^n (y_i - \bar{y})^2 = S_{yy}$$

Suma de cuadrado de la regresión:

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

Suma cuadrados residuales

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

siempre se cumple que $SSE = SST - SSR = S_{yy} - \hat{\beta} S_{xy}$

$$SSR = \beta S_{xy}$$

Contraste para la pendiente de la regresión poblacional, usando el procedimiento del análisis de varianza (ANOVA)

Suponga que tenemos n puntos de datos experimentales en la forma acostumbrada (x, y) que se estima la línea de regresión. Ahora suponga que las hipótesis que probaremos para la pendiente de la regresión poblacional β es la siguiente

$$H_0: \beta = 0 \quad \text{Contra} \quad H_1: \beta \neq 0$$

La hipótesis nula dice en esencia que el modelo de $y = \delta$. es decir, la variación en Y resulta del azar o de las fluctuaciones aleatorias que son independiente de los valores de x . Bajo condiciones de esta hipótesis nula se calcula el estadístico de prueba F-Fisher como se muestra en la siguiente tabla Anova.

Análisis de varianza (ANOVA)

Fuente de variación	Suma de cuadrados	Grados de libertad	Cuadrado medio	Razón F
Regresión				
Error				
total				

Entonces, una prueba de hipótesis con un nivel de significancia α para β siendo $F = \frac{SSE}{S_e^2}$ el estadístico de prueba correspondiente y F_α el valor de una variable aleatoria, a la derecha del cual se tiene un área de α en la distribución de F de Fisher con $v_1 = 1$ y $v_2 = n - 2$ grados de libertad, y no aceptamos H_0 al nivel de significancia α cuando $F > F_\alpha(1, n - 2)$

Contraste para el intercepto de la regresión poblacional

Bajo ciertas condiciones, la hipótesis que debe probarse para la pendiente de la regresión poblacional δ es:

$$H_0: \delta = \delta_o \quad \text{Contra} \quad H_1: \delta \neq \delta_o$$

Donde δ_o es cualquier número real, el estadístico de prueba tiene la forma $t = \frac{\hat{\delta} - \delta_o}{s_{\hat{\delta}}}$, la distribución a considerar es la *t - student* con $n-2$ grados de libertad y no se acepta a hipótesis nula si:

$$t \leq -t_{\alpha/2} \quad \text{O} \quad t \geq t_{\alpha/2}$$

De aquí se debe calcular

$$S_{\hat{\delta}}^2 = \frac{(S_e^2) \sum_{i=1}^n x_i^2}{nS_{xx}}$$

Covarianza y coeficiente de correlación

Supongamos que X e Y son un par de variables aleatorias dependientes. Sería deseable disponer, en tal caso, de una medida para la naturaleza de la relación entre ellas. Esto es difícil de conseguir, puesto que pueden estar relacionadas de maneras muy distintas (por ejemplo, lineal, cuadrática, exponencial. etc). Para simplificar, limitemos nuestra atención a la posibilidad de una relación lineal.

Teorema: sean X, Y dos variables aleatorias cualesquiera con varianzas finitas.

$$\text{Corr}(X, Y) = \hat{\beta} \sqrt{\frac{S_{xx}}{S_{yy}}} = \frac{S_{xy}}{\sqrt{S_{xx}S_{yy}}}$$

Interpretación

- a) $-1 \leq \text{Corr}(X, Y) \leq 1$
- b) Si X , Y son independientes, entonces $\text{Corr}(X, Y) = 0$, el recíproco no es cierto.
- c) Para fines descriptivos, la relación se propone como fuerte si $|\text{Corr}(X, Y)| \geq 0.8$, moderada si $0.5 < |\text{Corr}(X, Y)| < 0.8$ y débil si $|\text{Corr}(X, Y)| < 0.5$

Coeficiente de Determinación

El coeficiente de correlación representa la proporción de la variación explicada por el modelo de regresión, es decir, r^2 expresa la proporción de la variación total en los valores de la variable Y que pueden explicar mediante la relación lineal con los valores de la variable aleatoria X.

$$r^2 = \frac{S_{xy}^2}{S_{xx}S_{yy}} = \frac{SSR}{S_{yy}}$$

Supuestos para el modelo de regresión lineal simple.

Denotemos la recta verdadera de regresión por:

$$Y_i = \delta + \beta x_i + \epsilon_i$$

Y asumamos que se disponen de n pares de observaciones. Suelen realizarse al respecto los siguientes supuestos, para el análisis de los residuos (errores).

Linealidad,

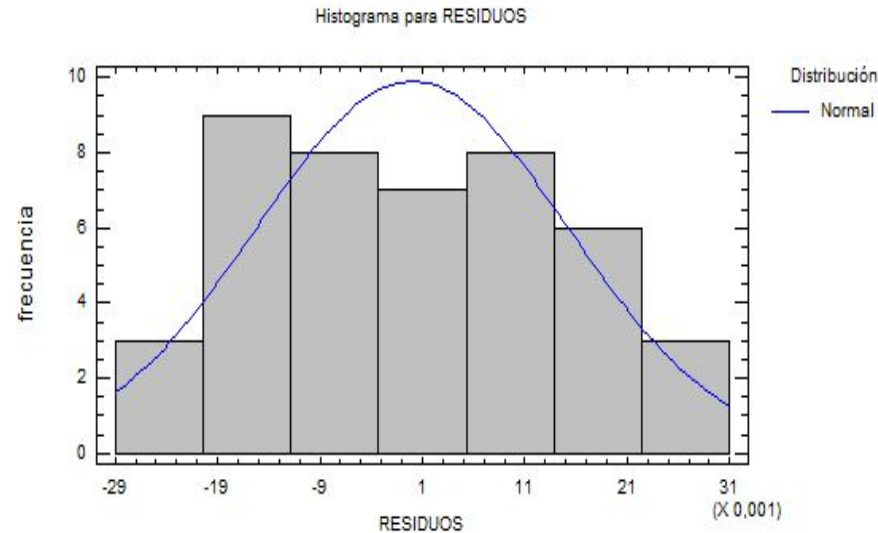
Normalidad,

Homoscedasticidad

Independencia

Análisis de residuos para comprobar supuestos.

Histograma de Residuos



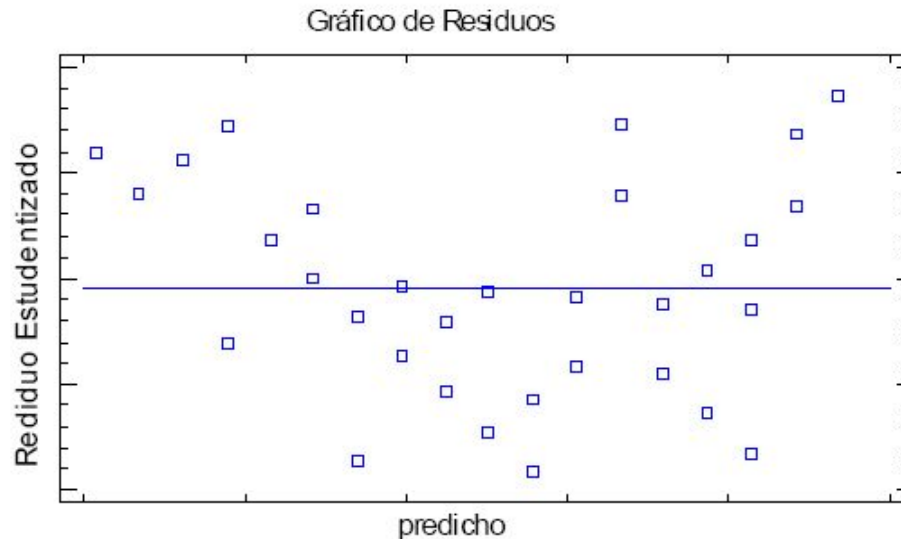
Si el histograma de frecuencias de los residuos no se ajusta al de una campana de Gauss (normal), pueden existir datos atípicos. Eliminado los pares (X_i, Y_i) que producen los valores atípicos, se puede conseguir normalidad en los residuos, no sin antes dar una justificación para tal fin.

swilk residuo

Homocedasticidad

Residuos versus valores de predicción

la grafica cuyos puntos son los pares (Y_i, ϵ_i) y detectamos una tendencia de cualquier tipo en el grafo, puede existir autocorrelación, ya que habrá correlación entre los residuos. También puede haber en este caso heteroscedasticidad, o también falta de linealidad.



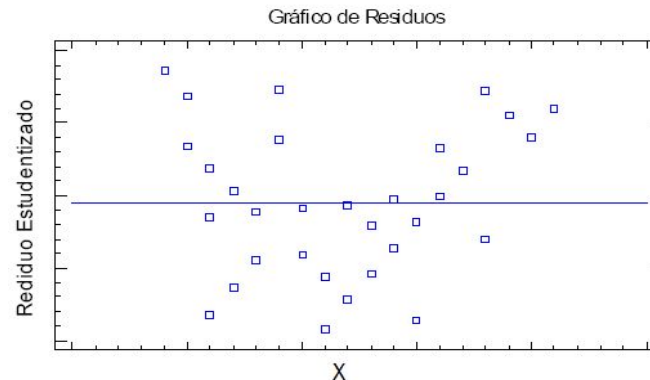
estat hettest

Breusch-Pagan / Cook-Weisberg test for heteroskedasticity
Ho: Constant variance

Linealidad - Independencia

Residuos versus X

Si graficamos los valores residuales versus los valores de la variable independiente, ósea los valores de ϵ_i en el eje vertical y los valores de X_i sobre el eje horizontal. Esta gráfica es útil para visualizar la necesidad de un modelo curvilíneo.



Note que existen intervalos en que los residuos yacen abajo de 0 (mostrado en la línea horizontal). Dentro de este rango, la línea recta sobreestima los valores estimados. Así mismo, existen intervalos que tiende a subestimar la predicción de los datos.

rvpplot (var independiente)



PRACTICA CON R

Se tomaron los datos del club de salud de 20 estudiantes de posgrado de Uninorte. Se desea saber si es posible predecir el número de pulsaciones por minutos después de realizar una actividad física. Realice el modelo de regresión y verifique los supuestos. (base de datos en el primer libro del archivo Excel “Lineal 1”)

Y: Pulsaciones por minuto en reposo

X: tiempo en correr 1 milla (dado en segundos)



REGRESIÓN LINEAL MÚLTIPLE

El modelo de dos variables, con frecuencia es inadecuado en la práctica. Es el caso del ejemplo consumo-ingreso, en donde se supuso implícitamente que solamente el ingreso X afecta el consumo Y . Pero la teoría económica rara vez es tan simple, ya que, además del ingreso, existen muchas otras variables que probablemente afectan el gasto de consumo.

Por consiguiente, se necesita ampliar el modelo simple de regresión con dos variables para considerar modelos que contengan más de dos variables.

La adición de variables conduce al análisis de los modelos de regresión múltiple, es decir, a modelos en los cuales la variable dependiente, o regresada, Y , depende de dos o más variables explicativas, o regresoras.

Modelo multiple

Generalizando la función de regresión se puede escribir:

$$Y_i = \beta_1 + \beta_2 X_{2i} + \beta_3 X_{3i} + \dots \beta_n X_{ni} \dots + \mu_i$$

donde Y es la variable dependiente, X_2, X_3 y... X_n las variables explicativas (o regresores). μ_i es el término de perturbación estocástica, e i la i ésima observación.

Los coeficientes se denominan coeficientes de regresión parcial.

Análisis de varianza (ANOVA)

Fuente de variación	Suma de cuadrados	Grados de libertad	Cuadrado medio	Razón F
Regresión				
Error				
total				

Se establece la siguiente prueba de hipótesis:

$$H_o = \beta_1 = \beta_2 \dots = \beta_n = 0$$

$$H_o = \text{alguno diferente de } 0$$

Modelos estadísticos

Coeficiente de Determinación

El coeficiente de correlación representa la proporción de la variación explicada por el modelo de regresión, es decir, r^2 expresa la proporción de la variación total en los valores de la variable Y que pueden explicar mediante la relación lineal con los valores de la variable aleatoria X.

R-cuadrado

$$R^2 = \frac{SSR}{SSR + SSE}$$

R-cuadrado Ajustada

$$R^2 = 1 - \left(\frac{n-1}{n-p} \right) \frac{SSR}{SSR + SSE}$$

Estadístico de Durbin-Watson

El **estadístico de Durbin-Watson**, desarrollado por el reputado economista Watson, es un estadístico de prueba que se utiliza para detectar la presencia de autocorrelación en los residuos (errores de predicción) de un análisis de la regresión.

$$D = \frac{\sum_{i=2}^n (e_i - e_{i-1})^2}{\sum_{i=1}^n e_i^2}$$

Si $n > 500$, entonces

$$D' = \frac{|D - 2|}{\sqrt{4/n}}$$



Criterios de información, sobre la elección de un buen modelo.

La idea general es esta: necesitamos una manera de elegir (formalmente) entre una serie de modelos estadísticos que difieren en complejidad y en calidad de ajuste (o capacidad de explicar nuestros datos).

Generalmente, cuanto más complejo es un modelo (porque tiene muchas variables, interacciones etc) mejor describirá el proceso que estamos analizando, pero si es demasiado complejo perderá generalidad (será muy bueno prediciendo los datos para los que está entrenado, pero nada más allá de eso). Así que tenemos un compromiso: queremos un modelo lo suficientemente complejo para que sea una buena descripción de nuestro sistema de estudio, pero no tan complejo como para que no tenga validez general.

Criterio de información de Akaike (AIC) - estadístico

Es una medida de la calidad relativa de un modelo estadístico, para un conjunto dado de datos. Como tal, el AIC proporciona un medio para la selección del modelo.

$$AIC = 2k - \ln(L)$$

donde k es el número de parámetros en el modelo estadístico, y L es el máximo valor de la función de verosimilitud para el modelo estimado.

Dado un conjunto de modelos candidatos para los datos, el modelo preferido es el que tiene el valor mínimo en el AIC. Por lo tanto AIC no solamente recompensa la bondad de ajuste, sino también incluye una penalidad, que es una función creciente del número de parámetros estimados.


Si lo que estamos haciendo es quitar parámetros para simplificar el modelo, podemos establecer como criterio dejar de quitar variables cuando el decremento de $AIC < 2$, lo que indica cambios no significativos en la bondad de ajuste del modelo.

Estas son las razones por las que, al comparar modelos, elegimos el que tiene menor AIC y establecemos que no son “significativos” los cambios de AIC menores que 2.

Criterio de información bayesiano (BIC) - estadístico

o el más general **criterio de Schwarz** (SBC también, SBIC) es un criterio para la selección de modelos entre un conjunto finito de modelos. Se basa, en parte, de la función de probabilidad y que está estrechamente relacionado con el Criterio de Información de Akaike (AIC).

Dadas dos modelos estimados, el modelo con el menor valor de BIC es el que se prefiere. El BIC es un aumento de la función de σ_ε^2 y una función creciente de k . Es decir, la variación no explicada en la variable dependiente y el número de variables explicativas aumentan el valor de BIC. Por lo tanto, menor BIC implica un número menor de variables explicativas, mejor ajuste, o ambos. La fuerza de la evidencia en contra del modelo con el mayor valor de BIC se puede resumir de la siguiente manera



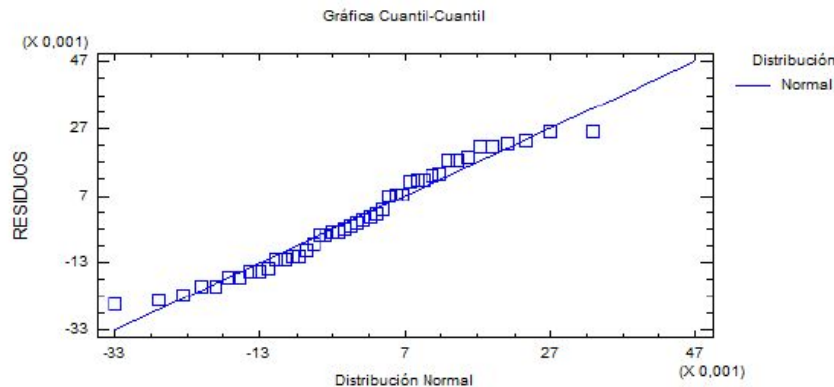
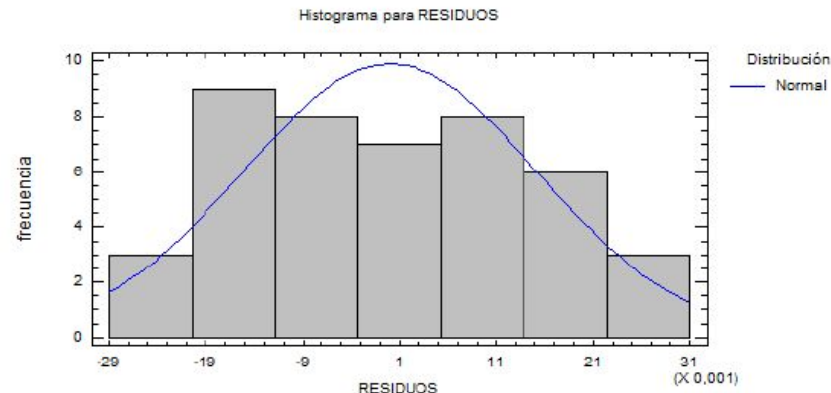
ΔBIC	Evidencia contra un BIC alto
0 - 2	No vale la pena más que una simple mención
2 - 6	Positivo
6 -10	Fuerte
>10	Muy fuerte

El BIC generalmente penaliza parámetros libres con más fuerza que hace el criterio de información de Akaike, aunque depende del tamaño de n y la magnitud relativa de n y k .

Es importante tener en cuenta que el BIC se puede utilizar para comparar los modelos estimados sólo cuando los valores numéricos de la variable dependiente son idénticos para todas las estimaciones que se están comparando. Los modelos que se comparan no tienen que ser anidados , a diferencia del caso cuando los modelos se comparan utilizando un F o prueba de razón verosimilitud .

Normalidad de residuos

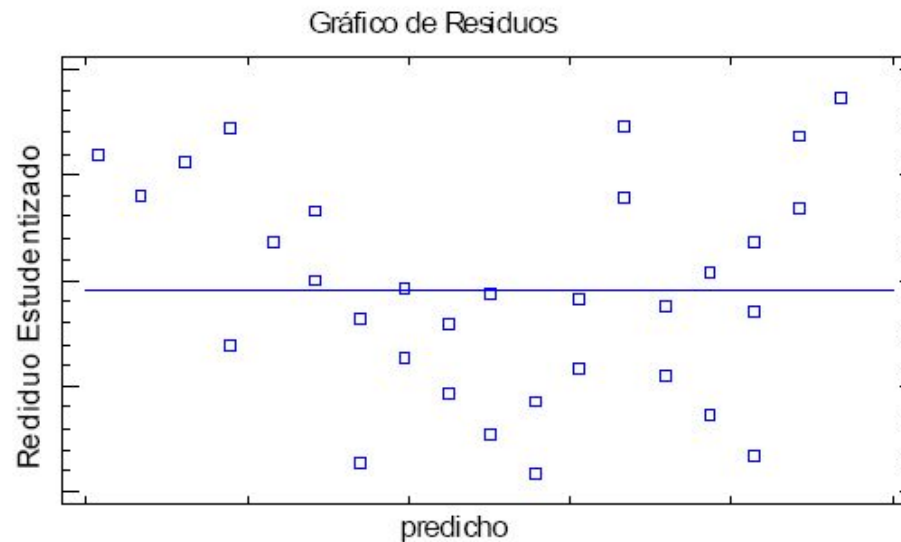
Si el histograma de frecuencias de los residuos no se ajusta a una campana de Gauss, puede ser indicio de datos atípicos, con justificación plausible o estadística se puede eliminar las parejas que producen datos atípicos.



Homocedasticidad

Residuos versus valores de predicción

la grafica cuyos puntos son los pares (Y_i, ϵ_i) y detectamos una tendencia de cualquier tipo en el grafo, puede existir autocorrelación, ya que habrá correlación entre los residuos. También puede haber en este caso heteroscedasticidad, o también falta de linealidad.

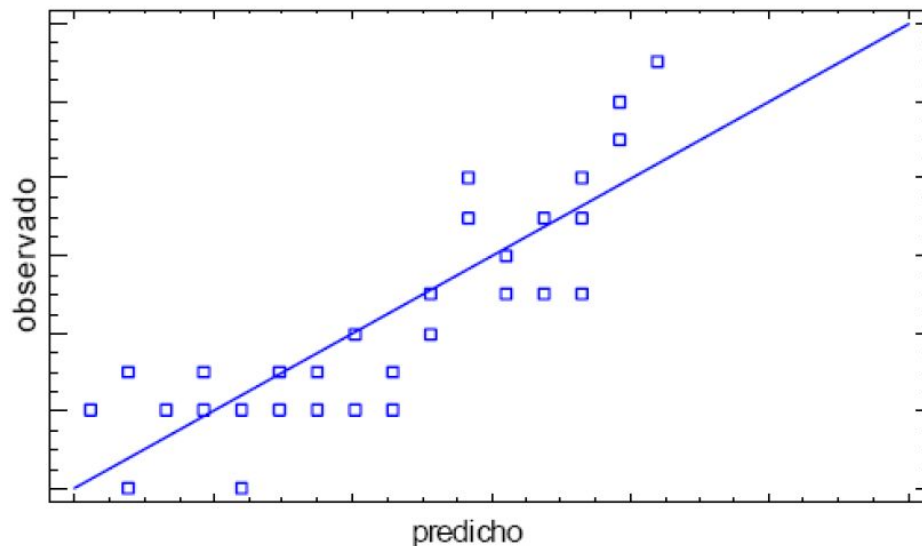


imtest, white

White's test for H_0 : homoskedasticity
against H_a : unrestricted heteroskedasticity

Relación lineal

Si graficamos los valores observados versus los valores de predicción, ósea los valores observados Y en el eje vertical y los valores de predicción \hat{Y} sobre el eje horizontal

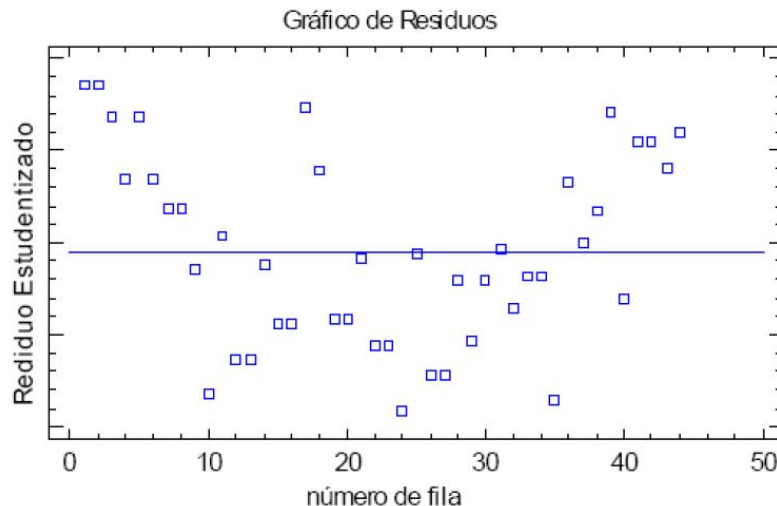


Si el modelo ajusta bien, los puntos deben estar colocados aleatoriamente alrededor de la línea diagonal. Es posible observar algunas veces curvatura en esta grafica, lo cual indica la necesidad de un modelo curvilíneo mas que uno lineal. En la grafica de arriba, la variabilidad parece ser un poco constante, sin embargo, alguna evidencia de curvatura esta presente

twoway (scatter RY prediccion) (lfit RY prediccion)

Independencia

Al realizar la grafica de los residuos estandarizados y el orden de los datos, cualquier tendencia en los valores, podría indicar una influencia externa (dependencia),



Se usa también el estadístico de Durbin – Watson, con estos valores, se puede adoptar la regla no demasiado rigurosa que si $DW \approx 0$ hay autocorrelación perfecta positiva, si $DW \approx 2$ no hay autocorrelación (Independencia) y si $DW \approx 4$ existe autocorrelación perfecta negativa.

La presencia de autocorrelación puede aliviarse introduciendo variables Dummy al modelo.


estat durbinalt

Problema de la Multicolinealidad y su detección

En el modelo lineal $Y_i = \delta + \beta_1 X_{1t} + \beta_2 X_{2t} + \beta_3 X_{3t} \dots + \beta_k X_{kt} + \epsilon_t$ suponemos una serie de hipótesis en las que se encontraban que las variables X_1, X_2, \dots, X_k son linealmente independientes, es decir, no existe relación lineal exacta entre ellas. Esta hipótesis se denomina de independencia, y cuando no se cumple, decimos que el modelo presenta multicolinealidad.

Como síntomas más comunes de multicolinealidad tenemos los siguientes casos.

- Valores altos en modulo en la matriz de correlaciones.
- Poca significatividad de las variables X y a la vez R^2 alto
- Gran significatividad conjunto del modelo (gran rechazo de $R^2 = 0$)
- Valores de índice de condición inferiores a 30.



Algunas soluciones más comunes para la multicolinealidad sea:

- Ampliar la muestra
- Trasformar las variables adecuadamente
- Suprimir algunas variables con justificación estadística
- Sustitución de las variables explicativas por sus componentes principales mas significativas.

Manos a la obra

El objetivo del estudio es encontrar la mejor relación entre las variables de interés con el resto de posibles variables explicativas o regresoras. Realice el modelo de regresión lineal múltiple y analice el cumplimiento de supuestos.

Y	R	"índice de criminalidad, número de delitos conocidos por la policía por cada millón de habitantes"
X1	Age	"distribución de la edad, número de varones de edad 14-24 por cada mil de toda la población del estado"
X2	S	"variable binaria que distingue entre estados del sur (S=1), (S=2) el resto del país"
X3	ED	"nivel educativo, número medio de años de escolarización"
X4	EX	"gasto per cápita en protección policial relativa"
X5	LF	"proporción en participación en trabajos de fuerza por cada mil hombres con edad 14-24"
X6	M	"Número de varones por mil mujeres"
X7	N	"Tamaño de la población del estado en cien mil"
X8	NW	"El número de personas de raza no blanca por 1000 habitantes"
X9	U1	"Razón de desempleo entre hombres de edad 14-24, por cada mil"
X10	U2	"Razón de desempleo entre hombres de edad 35-39, por cada mil"
X11	W	"Riqueza medida por el ingreso familiar"
X12	X	"Desigualdad en ingresos, el número de familias por mil que ganan por debajo de la mitad de la mediana de ingresos"

Modelos de regresión Poisson

La distribución de Poisson fue derivada por SIMEON DENIS POISSON, quien en 1837 publicó un trabajo de Investigación en el que se presentaba una nueva distribución para el cálculo de probabilidades aplicado al ámbito penal. Poisson encontró que cuando el tamaño de una muestra es grande y la probabilidad de ocurrencia de un evento es pequeña, el valor esperado $\mu = np$ tiende a una constante.



APLICACIONES DE LA VARIABLE DE POISSON

Como ya se mencionó anteriormente, un conteo es el número de veces en que cierto evento ocurre en una misma unidad de observación durante un determinado periodo de tiempo o espacio. Ejemplos de tales eventos o conteos pueden ser:

Conteos en el tiempo:

Número de accidentes de tráfico en un tramo de cierta carretera en un mes.
Número del registro de partículas de una desintegración radioactiva por segundo.
Número de mutaciones en una población de animales durante 5 años.



Conteos en el espacio:

- Número de accidentes de tráfico que se originan en el cruce de 2 carreteras.
- Número de organismos infecciosos propagados en una placa.
- Número de células sanguíneas en una muestra de sangre (el espacio es igual al volumen en centímetros cúbicos)
- Número de árboles infectados por hectárea en un bosque.

Distribución Poisson

Esta distribución ha sido empleada extensamente en biología y medicina como modelo de probabilidad.

Si x es el numero de ocurrencias de algún evento aleatorio en un intervalo de espacio o tiempo determinado , la probabilidad de que x ocurra es dada por:

$$P(X = k) = \frac{e^{-\lambda} \lambda^k}{k!}$$

Donde el parámetro λ es la media de eventos ocurridos en determinado tiempo o espacio.

El valor esperado viene dado por:

$$E(X) = \lambda$$

La varianza

$$V(X) = \lambda$$

Ejemplo:

$$P(X \leq a) = P(X = a) + P(X = a - 1) + \dots + P(X = 0)$$

$$P(X < a) = P(X \leq a - 1), P(X > a) = P(X \geq a + 1)$$

$$P(X \geq a) = 1 - P(X \leq a - 1)$$

$$P(a \leq X \leq b) = P(X \leq b) - P(X \leq a - 1)$$

En un estudio de suicidas, los científicos encontraron que la distribución mensual de adolescentes suicidas en el Córdoba, entre 2018 y 2019 siguió una distribución de Poisson con parámetro $\lambda = 2.75$. Encuentre la probabilidad de que un mes seleccionado aleatoriamente:

A. Se suicide 3 adolescentes


Sea X v.a que determina el numero de suicidios mensuales.

$$P(X = 3) = \frac{e^{-2.75} 2.75^3}{3!} = 0.2215$$

B. Se suicide a lo sumo 3 adolescentes.

$$P(X \leq 3) = P(X = 3) + P(X = 2) + \dots + P(X = 0) = 0.7030$$

$$P(X \geq 3) = 1 - P(X \leq 2) = 1 - (P(X = 2) + \dots + P(X = 0))$$



C, ¿Cuál es la probabilidad de que un mes seleccionado aleatoriamente sea uno en el que ocurrirán tres o cuatro suicidios?

D. Estime la probabilidad de que en un determinado mes, más de 3 suicidios ocurran.

E. Estime el numero promedio de suicidio ocurrido en un mes e interprete su desviación estándar

Regresión Poisson

Este modelo está diseñado para ajustar un modelo de regresión en el cual la variable dependiente Y consiste de conteos. El modelo de regresión ajustado relaciona Y con una o más variables predictoras X , que pueden ser cuantitativas o categóricas.


El procedimiento ajusta un modelo usando máxima verosimilitud o mínimos cuadrados ponderados.

En todos los casos de una regresión de Poisson los valores de la variable son discretos, digamos $0, 1, 2, \dots$ sin un límite superior; sesgados hacia la izquierda e intrínsecamente heterocedásticos, es decir con una varianza que se incrementa paralelamente con la media



La variable respuesta

La variable respuesta se asume que tiene una distribución de probabilidad Poisson, en la cual la variable aleatoria se define como el número de eventos que ocurren en un intervalo de tiempo, cuya ocurrencia es aleatoria, independiente en el tiempo y con una tasa constante de ocurrencia. La distribución Poisson es usada para modelar eventos por unidad espacial como también por unidad de tiempo.



La distribución de Poisson se ha aplicado a diversos eventos, como el número de soldados muertos a patadas por caballos en el ejército prusiano (von Bortkiewicz 1898); el patrón de impactos de las bombas de zumbido lanzadas contra Londres durante la Segunda Guerra Mundial (Clarke 1946); conexiones telefónicas a un número incorrecto (Thorndike 1926); e incidencia de enfermedades, típicamente con respecto al tiempo, pero ocasionalmente con respecto al espacio.



Los supuestos básicos son los siguientes:

1. Existe una cantidad llamada tasa de incidencia que es la tasa a la que ocurren los eventos. Algunos ejemplos son 5 por segundo, 20 por 1000 personas-año, 17 por metro cuadrado y 38 por centímetro cúbico.
2. La tasa de incidencia se puede multiplicar por la exposición para obtener el número esperado de eventos observados. Por ejemplo, una tasa de 5 por segundo multiplicada por 30 segundos significa que se esperan 150 eventos; una tasa de 20 por 1000 personas-año multiplicada por 2000 personas-años significa que se esperan 40 eventos; etcétera.
3. En exposiciones muy pequeñas, la probabilidad de encontrar más de un evento es pequeña en comparación con.
4. Las exposiciones que no se superponen son mutuamente independientes.

Modelo Estadístico

El modelo estadístico asumido para los datos es que los valores de la variable dependiente Y siguen una distribución Poisson de la forma:

$$p(Y_i) = \frac{e^{-\lambda_i t_i} (\lambda_i t_i)^{Y_i}}{Y_i!}$$

donde λ_i es el parámetro de la tasa Poisson en los valores de las variables predictoras correspondientes a la i -ésima observación. Se supone además que la tasa se relaciona con las variables predictoras a través de una función de enlace log-lineal de la forma

$$\log \lambda = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \dots + \beta_n X_n$$

La distribución Poisson viene determinada completamente por su media. Esto impone la restricción de $E(Y/X) = V(Y/X)$ la cual no siempre se cumple en las aplicaciones empíricas.

Modelo de regresión de Poisson


- Nada impide que usemos MCO en una variable que tome valores (0, 1, 2, ...)
- Pero, ¿Qué pasa si queremos usar un modelo log-nivel?
- No es posible tomar el logaritmo de una variable de conteo porque toma el valor cero (y usualmente el grupo que toma ese valor en una porción no es despreciable).
- Por ello se modela $E(y / x_1, x_2, \dots x_n)$ como

$$E(y/x_1, x_2, \dots x_n) = E(y/x) = \exp(\beta_0 + \beta_1 X_1 + \beta_2 X_2 + \dots + \beta_n X_n) > 0$$

Por lo tanto se tiene

$$\text{Log}(E(y/x x_1, x_2, \dots x_n)) = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \dots + \beta_n X_n$$

$$\log \lambda = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \dots + \beta_n X_n$$



*Los valores de B se pueden interpretar como **semielasticidad (simil a OR)** de Y respecto a X . Se interpreta como un incremento de 1 unidad en X es asociado a un cambio en Y de $(100 \cdot \beta_1)\%$. Se interpreta como un incremento del 1% en X es asociado a un cambio en Y de $B_1\%$.*

Se debe usar la aproximación exacta para la interpretación cuando los betas estimados son grandes (> 0.1)

$$(\exp(\widehat{\beta}_k) - 1) * 100$$

Dado que la distribución Poisson viene determinada completamente por su media. Esto impone la restricción de $E(Y/X) = V(Y/X)$. El modelo exhibe heterocedasticidad

Modelo Estimado de Regresión:

Estimaciones de los coeficientes del modelo de regresión, con errores estándar y las Razones de Momios. Las razones de momios se calculan a partir de los coeficientes del modelo $\hat{\beta}_j$ por medio de

$$\text{razón de momios} = \exp(\hat{\beta}_j)$$

La razón de momios representa el incremento porcentual en el momio de los eventos por unidad de incremento en X.

Porcentaje de Desviación: dada por el modelo, se calcula por medio de:

$$R^2 = \frac{\delta(\beta_1, \beta_2 \dots + \beta_n / \beta_0)}{\delta(\beta_0)}$$

$$R^2_{adj} = \frac{\delta(\beta_1, \beta_2 \dots + \beta_n / \beta_0) - 2p}{\delta(\beta_0)}$$

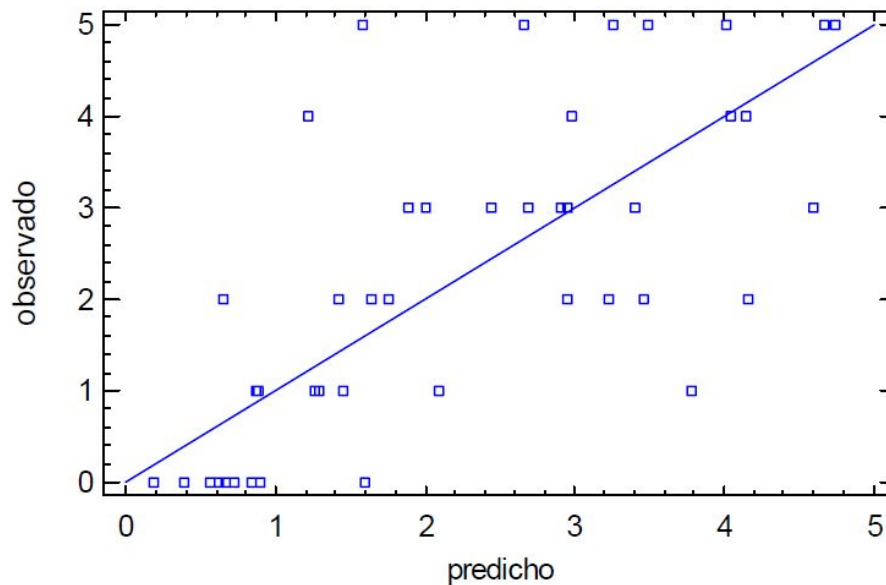
donde p es igual al número de coeficientes en el modelo ajustado, incluyendo al término constante. Es semejante a la estadística R-cuadrada ajustada en que compensa el número de variables en el modelo.

El modelo ajustado para los datos viene dado por:

$$\hat{\lambda} = \exp(\beta_0 + \beta_1 X_1 + \beta_2 X_2 + \dots + \beta_n X_n)$$

Observados Versus Predichos

El gráfico *Observados versus Predichos* muestra los valores observados de Y en el eje vertical y los valores medios predichos $\hat{\lambda}_i t_i$ en el eje horizontal.



estat gof

$H_0 = \text{Datos no siguen una distribucion poisson}$



En 2020 se realizó un estudio para estimar el número de veces que un grupo de personas han estado en la cárcel,

Estime un modelo Poisson para estimar los factores asociados para la tasa de delitos cometidos en un periodo determinado.

Base de datos: Crime1

Estudios de Supervivencia

- Investigaciones en las cuales una muestra de unidades es observada por un período de tiempo durante el cual se producen uno o varios acontecimientos.
- Un estudio ideal comienza con la totalidad de la muestra y permanece con ella hasta que todas las unidades alcancen el objetivo determinado (el acontecimiento de interés).
- Sin embargo, la mayoría de los estudios no son ideales. Dos importantes excepciones caracterizan estos estudios:
 - 1 Hay investigaciones que necesitan muchas unidades o que investigan acontecimientos raros. Estos deben ir añadiendo unidades durante meses o años.
 - 2 Hay estudios en que unidades que se pierden o desaparecen del estudio o que no han sufrido el acontecimiento de interés antes de que el estudio termine. Estas observaciones se dicen que son censuradas.

Ejemplo: Tiempo hasta la progresión

Sin datos censurados

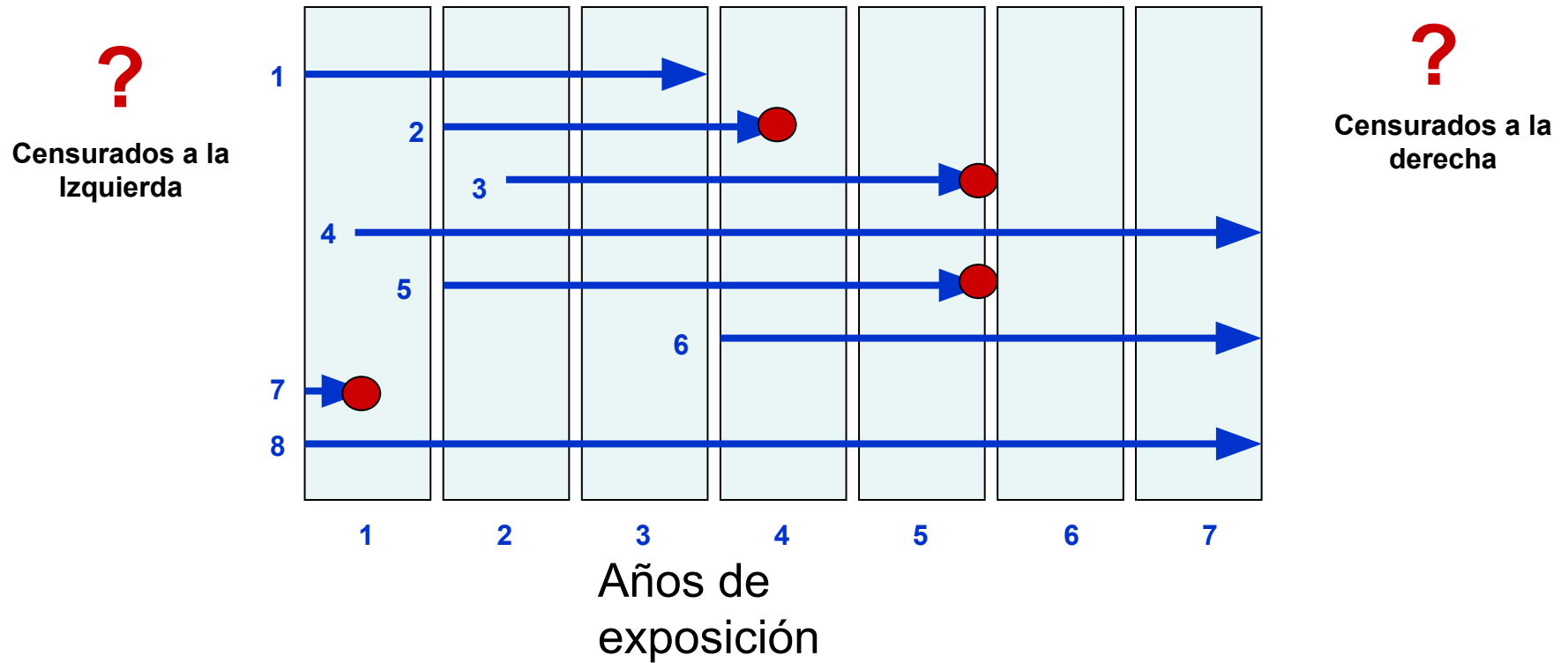
Si tenemos los siguientes tiempos de progresión de una enfermedad de 6 pacientes: 14, 13, 12, 23, 23 y 24 meses.

¿Cuál es la proporción de pacientes que sobreviven sin progresión mas de 24 meses?

Con datos censurados

Si tenemos los siguientes tiempos de progresión de una enfermedad de 6 pacientes: 14, 13, 12, 23+, 23+ y 24 meses.

Análisis de supervivencia de Kaplan-Meier





Modelos de Supervivencia

Existen muchas situaciones en las se desea examinar la distribución de un período entre dos eventos, como la duración del empleo (tiempo transcurrido entre el contrato y el abandono de la empresa). Sin embargo, este tipo de datos incluye generalmente algunos casos censurados. Los casos censurados son casos para los que no se registra el segundo evento (por ejemplo, la gente que todavía está trabajando en la empresa al final del estudio). El procedimiento de Kaplan-Meier es un método de estimación de modelos hasta el evento en presencia de casos censurados. El modelo de Kaplan-Meier se basa en la estimación de las probabilidades condicionales en cada punto temporal cuando tiene lugar un evento y en tomar el límite del producto de esas probabilidades para estimar la tasa de supervivencia en cada punto temporal.



APLICACIONES

Gobierno:

Determinar qué tan largo es el periodo de desempleo en Colombia

Financiero:


Determinar que tan largo es el periodo sin fusionarse de una empresa

Educación:

Estimar la función de supervivencia (no abandono) de los estudiantes de educación superior.

Salud:

Determinar cuál es la proporción de pacientes que viven por cierto tiempo después de un tratamiento

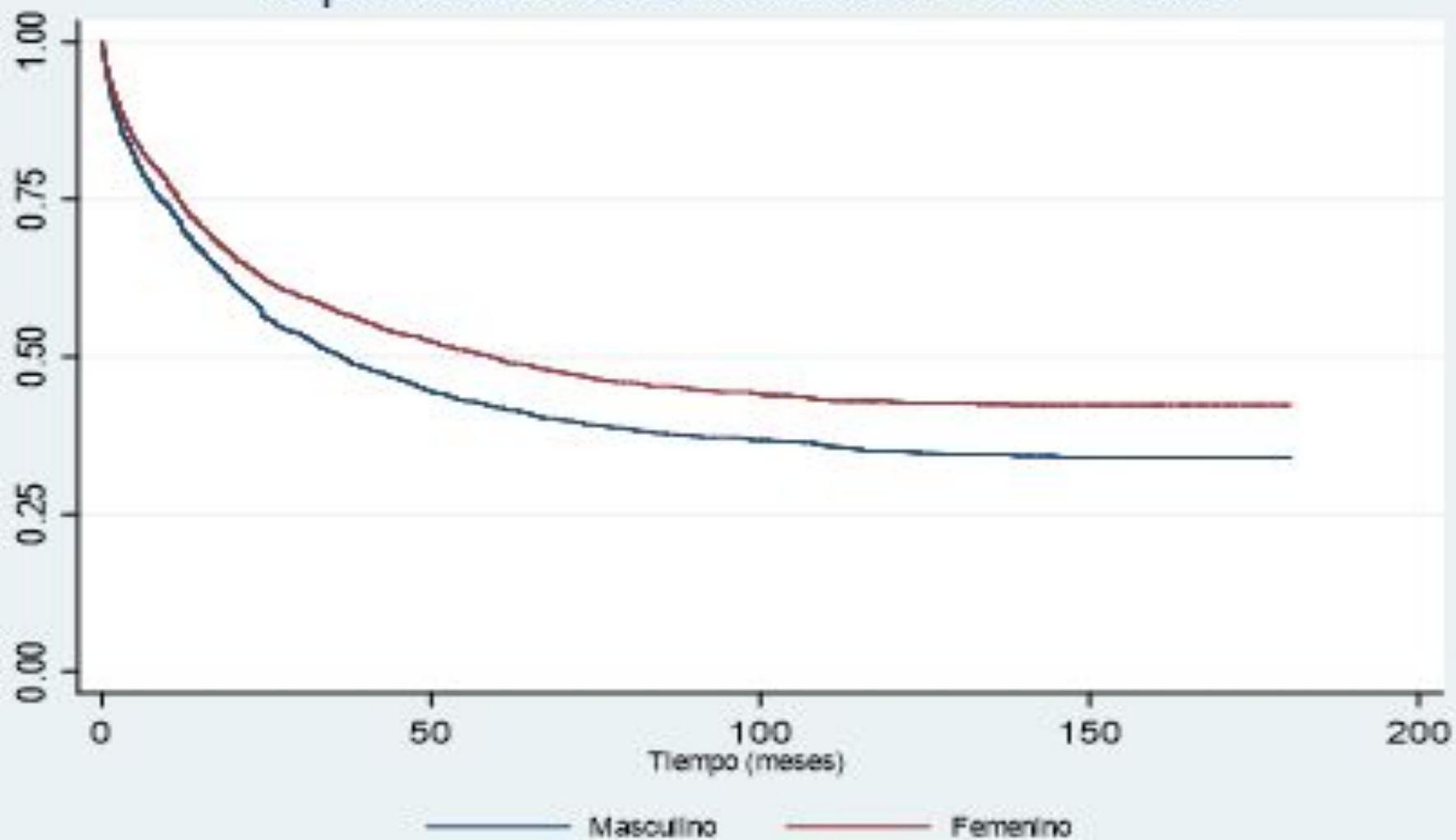


Ejemplo. ¿Posee algún beneficio terapéutico sobre la prolongación de la vida un nuevo tratamiento para el SIDA Se podría dirigir un estudio utilizando dos grupos de pacientes de SIDA, uno que reciba la terapia tradicional y otro que reciba el tratamiento experimental.

Al construir un modelo de Kaplan-Meier a partir de los datos, se podrán comparar las tasas de supervivencia globales entre los dos grupos, para determinar si el tratamiento experimental representa una mejora con respecto a la terapia tradicional. Si desea obtener información más detallada, también es posible representar gráficamente las funciones de riesgo o de supervivencia y compararlas visualmente.

Ejemplo: Identificar las tasas de supervivencia general de algún tipo de cáncer entre hombre y mujer.

Kaplan-Meier survival estimates - Cáncer Colon



Supervivencia y riesgo

Los datos de supervivencia se puede analizar con dos tipos de probabilidades diferentes: supervivencia y riesgo.

Función de supervivencia $S(t)$ “Probabilidad de supervivencia”

$$\hat{S}(t) = \prod_{t_i \leq t} \frac{n_i - d_i}{n_i}$$

n_i = Número de individuos en riesgo en el instante t .

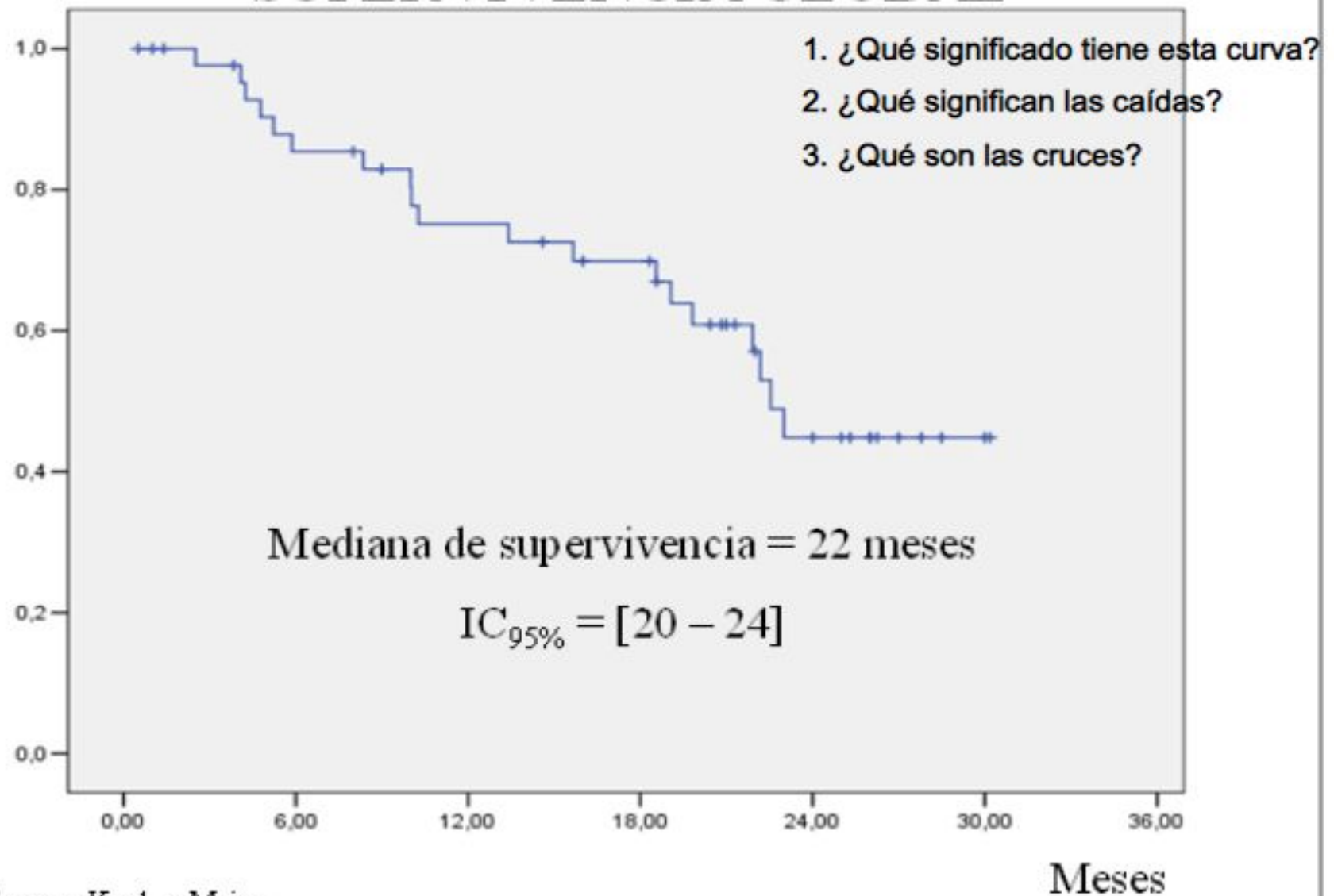
d_i = Número de veces que se materializa el evento en el instante t .

- Probabilidad de que un individuo sobreviva desde la fecha de entrada en el estudio hasta un momento determinado en el tiempo t .
- Se centra en la “no ocurrencia del evento”

Función de Riesgo (Hazard Ratio) $h(t)$.

- Probabilidad de que un individuo que esta siendo observado en el tiempo t le sucede el evento de interés en ese preciso momento
- Se centra en la “ocurrencia del evento”

SUPERVIVENCIA GLOBAL



Gráfica por Kaplan-Meier



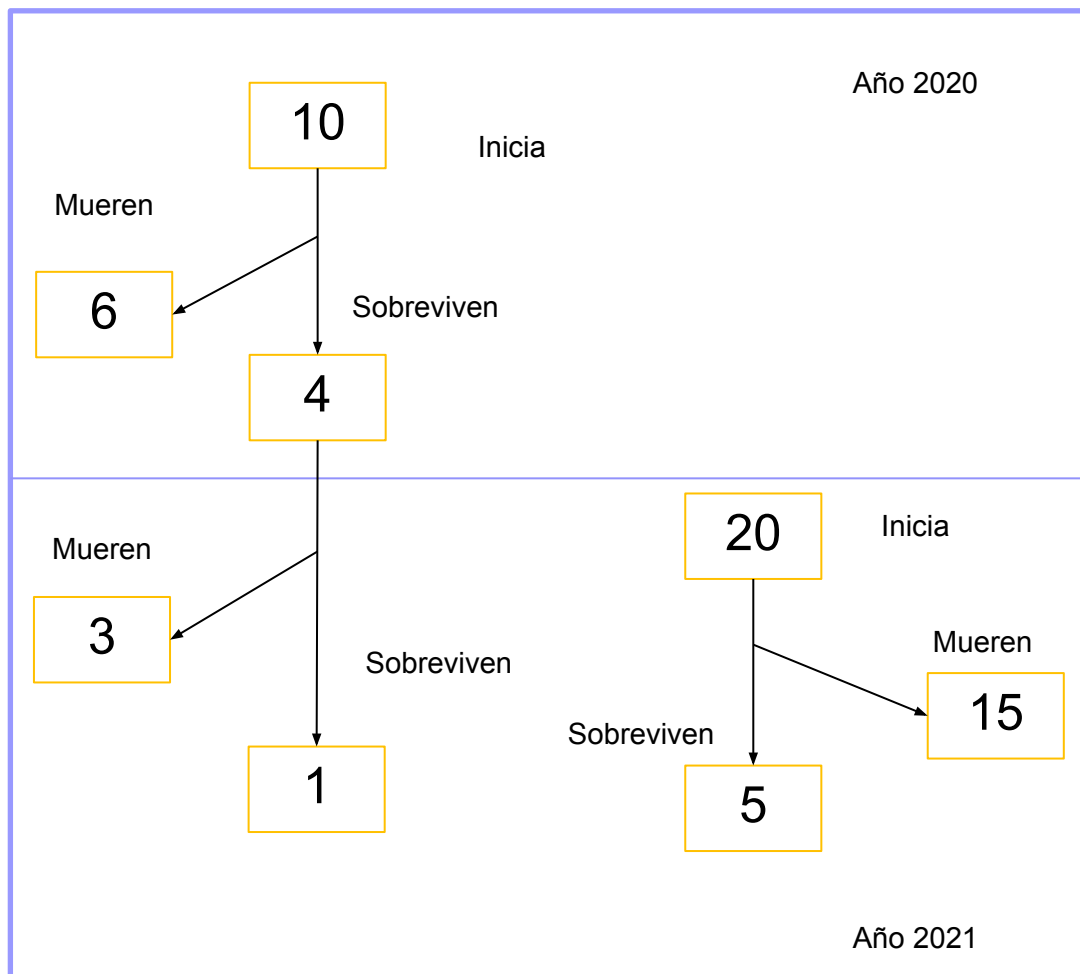
Estimación de la función de Supervivencia: Método de Kaplan – Meier

Aprovecha de información “censurada”

Calcula la supervivencia cada vez que un paciente presenta el evento

Se basa en el concepto de probabilidad condicional

Estimación de la función de Supervivencia: Método de Kaplan – Meier



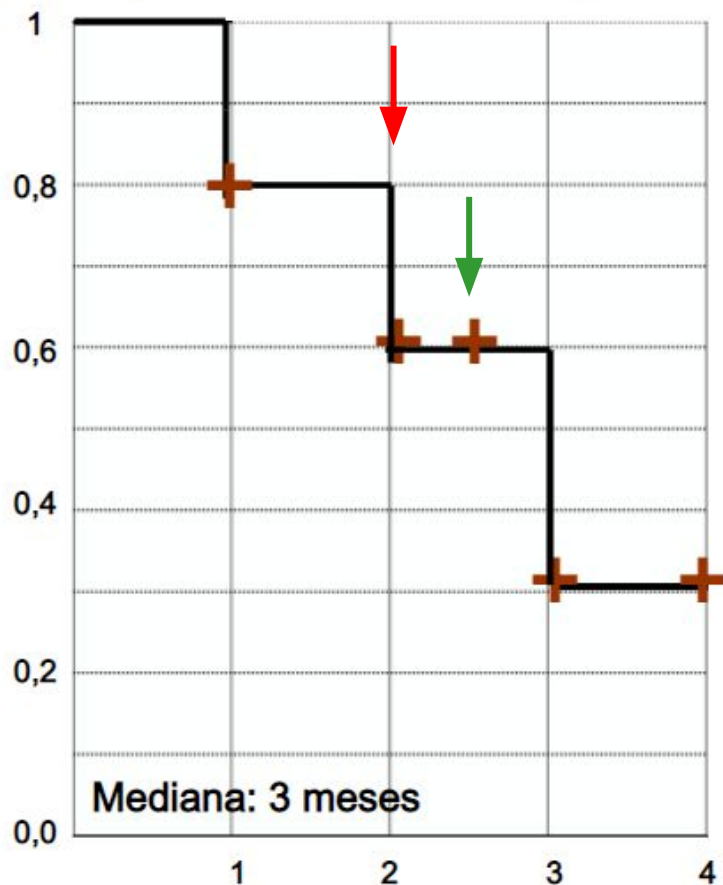
Alternativas:

1: Ignorar a los segundos pacientes que ninguno ha estado dos años en observación

$$\hat{S}(2) = 1/10$$

2: Los pacientes que sobreviven dos años son aquellos que sobreviven uno y después otro.

$$\hat{S}(2) = \left(\frac{4 + 5}{10 + 20} \right) 1/4$$



↓ Evento
↓ Censura

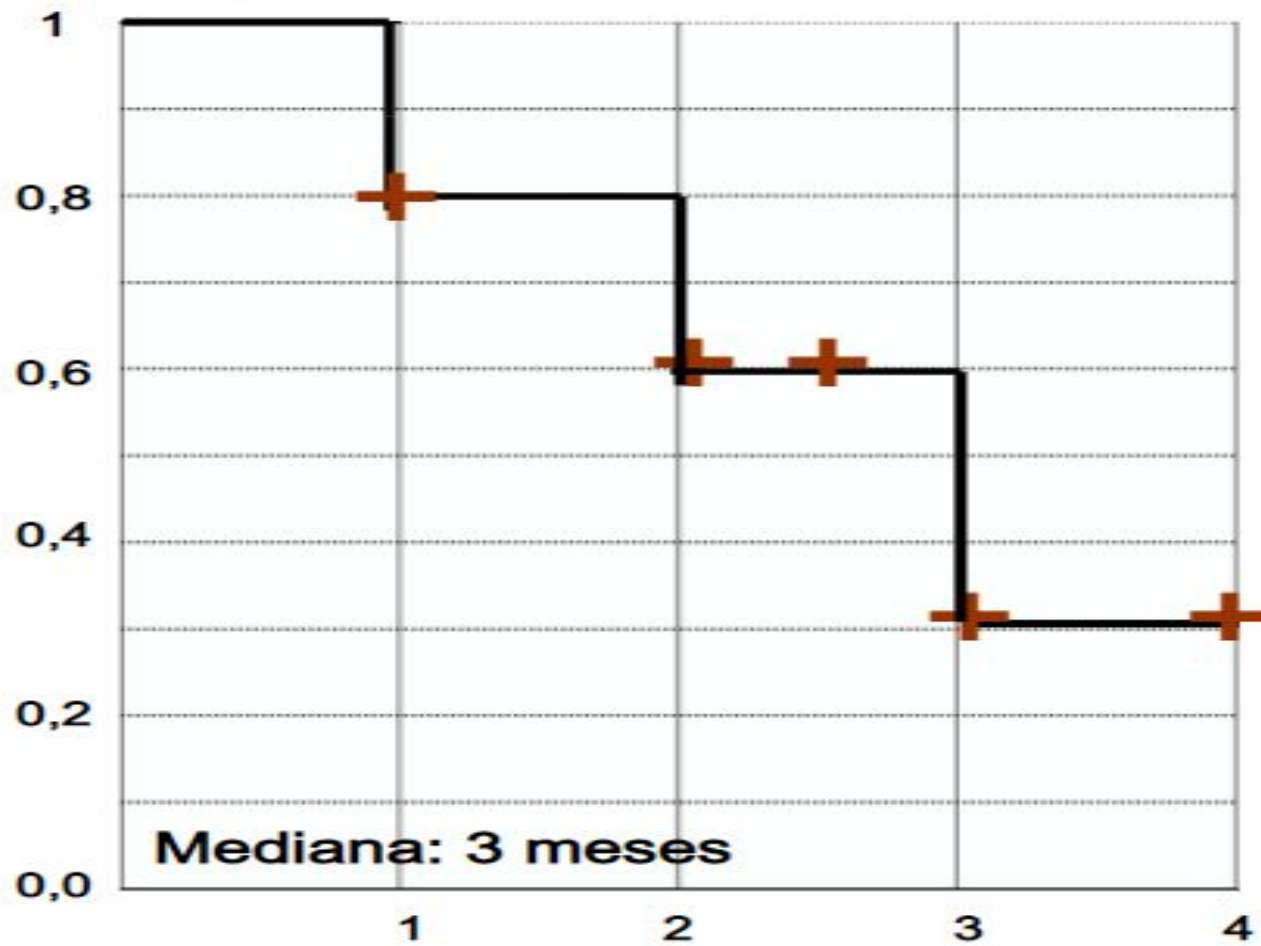
- Los saltos se dan solo cuando ocurre un evento.
- Cada dato censurado influye disminuyendo el denominador, con lo que, un individuo censurado no provoque un salto, si provoca una mayor magnitud en el tamaño del siguiente salto.

¿Cómo construir la curva de K-M?

Suponga que al inicio de un estudio se observan 5 pacientes por 4 meses.

Tiempo (meses)	Estatus	Fallecidos durante intervalo	Vivos al inicio del intervalo	Probabilidad de sobrevivir ese intervalo	Probabilidad acumulada de sobrevivir
1	F	1	5	$4/5$	$4/5=0.8$
2	F	1	4	$3/4$	$0.8 * 3/4=0.6$
2,5	C				
3	F	1	2	$1/2$	$0.6*1/2=0.3$
4	C				

Curva de K-M



Comparación de dos o más curvas de Supervivencia


Debido a que los tiempos de supervivencia no presentan una distribución normal, es necesario hacer una prueba no paramétrica denominada:

Log-Rank Test

Ho: La supervivencia de los grupos que se comparan es la misma

Hi: Al menos uno de los grupos tiene una supervivencia diferente

El estadístico usado es Chi-cuadrado con $k-1$ grados de libertad, siendo k el numero de grupos (número de curvas que se comparan)



Tiene en cuenta la evolución completa de la curva de supervivencia de ambos grupos, es decir, es capaz de detectar diferencias “persistentes” a lo largo del tiempo en la supervivencia

Otorga la misma ponderación a todos los tiempos de seguimiento

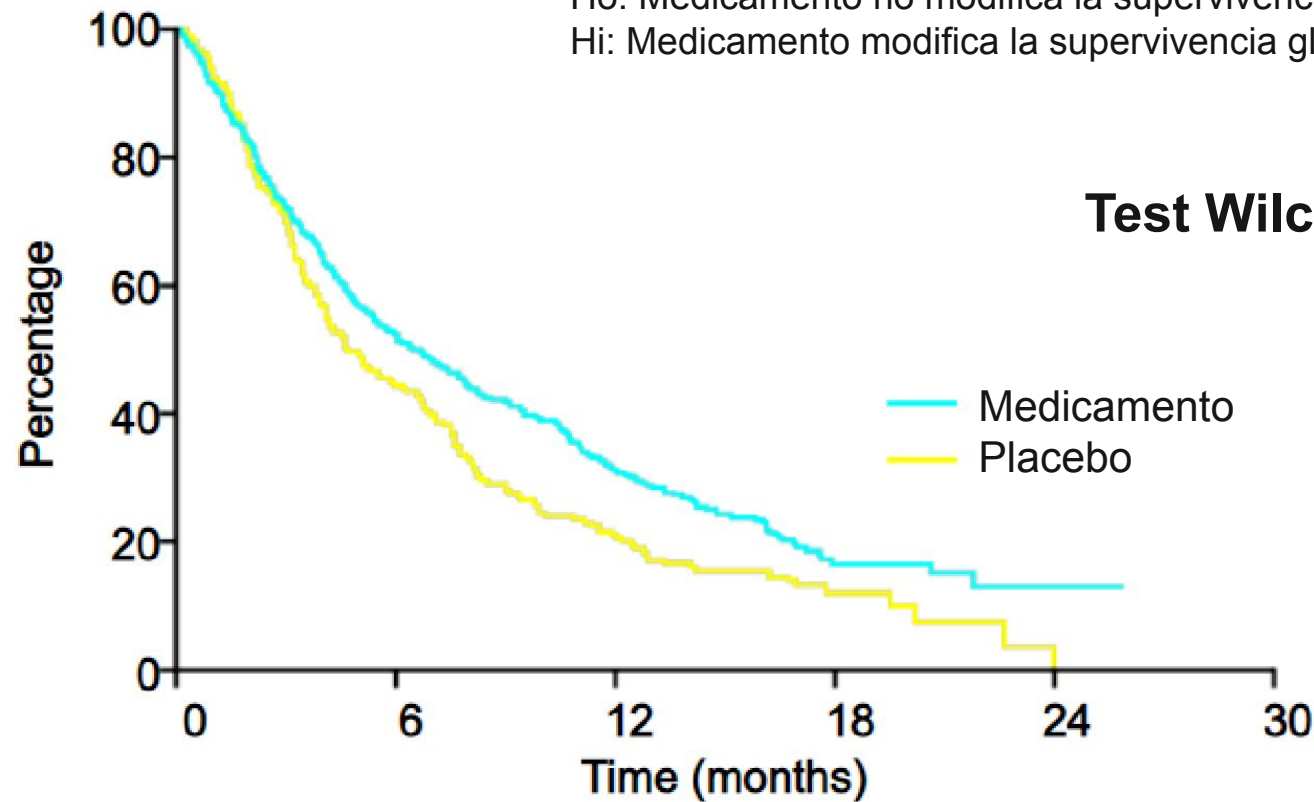
Es útil cuando el evento es poco frecuente o si las curvas son aproximadamente paralelas (no se cruzan)

Si las curvas se cruzan es más útil el test de Wilcoxon generalizado, también llamado Test de Breslow (Gehan). Otorga una mayor ponderación a los tiempos iniciales, que tienen mas observaciones.

Comparación de un tratamiento contra VHI vrs. Placebo

Ho: Medicamento no modifica la supervivencia global en comparación al placebo.
Hi: Medicamento modifica la supervivencia global en relación al placebo.

Test Wilcoxon P:value: 0.001





Problemas con test de igualdad de supervivencia

- No valora o cuantifica la diferencia (En caso de haberla)
- No estudia el posible efecto de otras covariables (variables pronósticas o predictoras)

Variables de confusión.

- Su presencia produce sesgos entre la variable dependiente y la independiente: La posible solución es realizar ajustes estadísticos con análisis estratificados o con técnicas de análisis multivariante.

Variables de interacción o modificadoras de efecto

- Sus valores cambian la intensidad o el sentido de la relación entre el factor de estudio (exposición) y la variable dependiente (respuesta)



Modelo de riesgos proporciones Regresión de COX

Identificada Factores asociados al desenlace de interés

Identifica Hazard Ratio.

HR: Razón entre las funciones de riesgo de los grupos en comparación.

- Es una tasa más que una probabilidad
- El termino Hazard corresponde a una tasa instantánea, que conceptualmente solo requiere una duración de tiempo mínimo.

Identifica Hazard Ratio.

Hazard: Se calcula dividiendo los sucesos ocurridos en un instante de tiempo, entre el total de sujetos en riesgo.

Probabilidad (P)	Intervalo de tiempo	Hazard
1/5 (Grupo A)	2 años	0.2 (a 2 años)
2/6 (Grupo B)	2 años	0.33 (a 2 años)

Hazard Ratio: Es la razón de Hazard (razón entre dos funciones de riesgo).

$$\text{HR: } 0.33/0.2 = 1.7$$

La velocidad con que ocurre el fenómeno es 1.7 veces superior en el grupo B que en el grupo A, al hacer la comparación a los 2 años



Modelo de riesgos proporciones Regresión de COX

La estimación final de HR global que produce el análisis de la regresión de COX viene a ser algo similar al promedio de los HR parciales.

Promedia de manera pondera las HR de los diversos momentos en los que se produce un evento, dando lugar a un HR global

La regresión de COX asume que la relación de las tasas instantáneas es constante con el tiempo (proportional hazards model)



Supuesto: Los Hazard son constante con el tiempo

test of proportional-hazards assumption

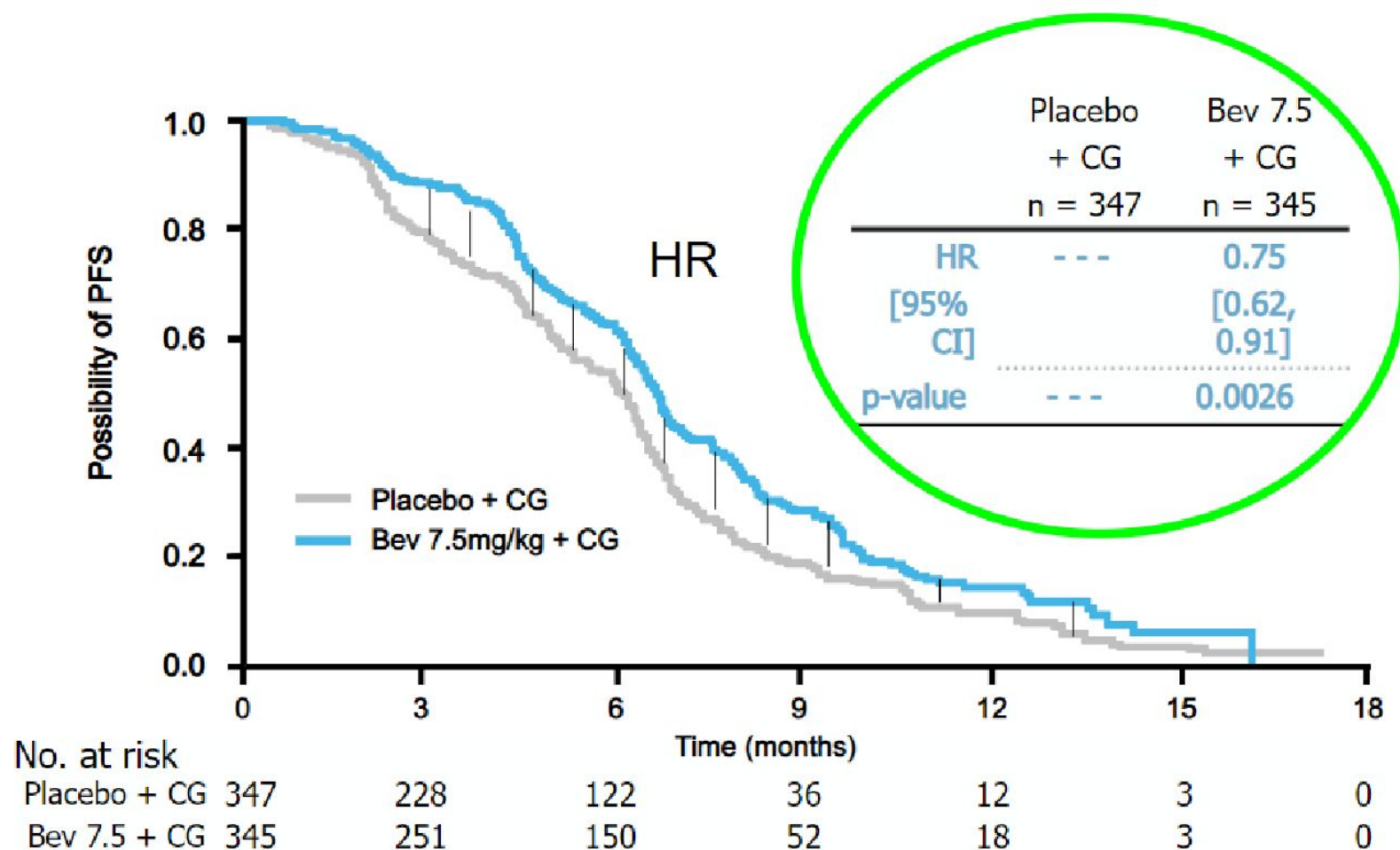
Estat phtest

Si este supuesto no se cumple, no es correcto estimar un HR ya que se podrían anular,

Para suplir este inconveniente se requiere emplear un modelo de riesgo no proporcionales.

Una forma alternativa y sencilla es interpretar las estimaciones del HR fragmentadas en el tiempo.

Ensayo de superioridad



Cis + Gem vs Cis+ Pem: Overall Survival

