Universitat de Barcelona
UNIVERSITAT POLITÈCNICA DE CATALUNYA
BARCELONATECH
Departament d'Estadística
i Investigació Operativa

.5. Regresión a la media

Medical Statistics

José Antonio González y Erik Cobo

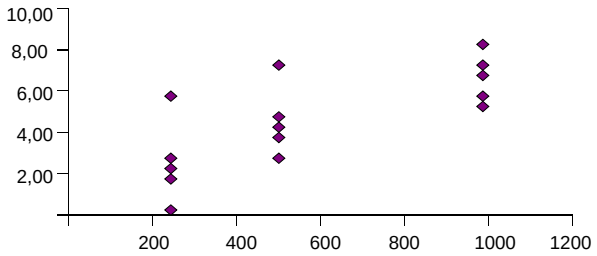
Abril 2015

LAB: Regresión a la media y evolución espontánea

Modelo de regresión tipo I:

$Y_i = \beta_0 + \beta_1 X_i + \varepsilon_i$ $\quad \varepsilon_i$ = desvío (¡vertical!) del caso i con la recta

Ejemplo: Se desea estudiar la concentración en plasma de colesterol tras la administración de 250, 500 y 1000 mgr de cierta sustancia.



Tiene sentido proponer como estimadores de β_0 y β_1, aquellos valores b_0 y b_1 que hagan pequeño $\Sigma_{i=1...n} e_i^2$

Med. Stat: introduction

Un ejemplo *distinto*:

Galton estudiaba la relación existente entre variables antropométricas de los progenitores y de sus descendientes; pero, en la altura de padres e hijos, ¿es realista el modelo anterior?

Ahora ambas variables tienen variación: tanto la altura del progenitor (X_{1i}) como la del descendiente (X_{2i}) vienen influida por factores genéticos (Ψ_i) y ambientales (ε_i):

$$X_{1i} = \beta_0 + \beta_1 \Psi_i + \varepsilon_i$$

$$X_{2i} = \beta_0 + \beta_1 \Psi_i + \varepsilon'_i$$

Asumiendo, en este modelo simplista, que los factores genéticos se transmiten de forma inmutable (Ψ_i no cambia) y que tienen la misma influencia (β_0 y β_1 tampoco cambian), pero sí el ambiente ($\varepsilon_i \neq \varepsilon'_i$).

Med. Stat: introduction

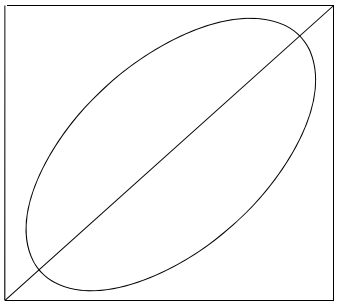
Por simplicidad, asumamos que, para padres e hijos, $\beta_0 = 0$ y $\beta_1 = 1$

$$X_{1i} = \Psi_i + \varepsilon_i$$

$$X_{2i} = \Psi_i + \varepsilon'_i$$

Y que los términos de error tienen idéntica dispersión: $\sigma^2_{\varepsilon\,1} = \sigma^2_{\varepsilon\,2}$

La expresión gráfica de este modelo sería:

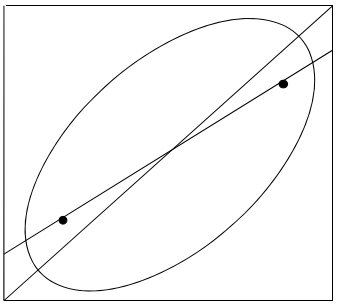


La recta identidad refleja aquellos padres cuya altura se repite en sus hijos.

Med. Stat: introduction

¿Tiene sentido, ahora, minimizar la distancia vertical?

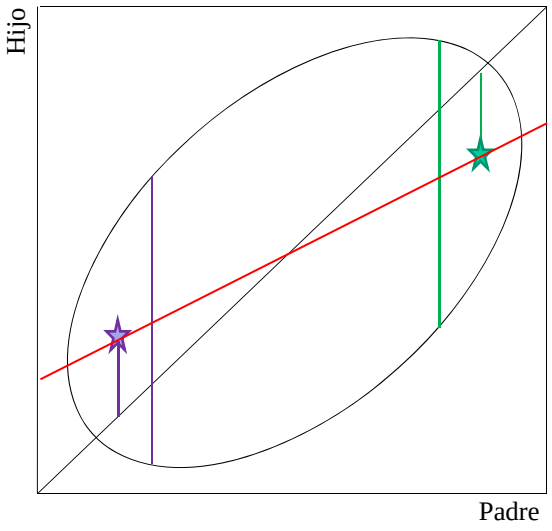
(el punto representa la "media vertical")



Nótese que la recta se aplana, y en consecuencia:
en media, la altura de los hijos de padres muy altos no sería tan alta.
en media, la altura de los hijos de padres muy bajos no sería tan baja.

Los hijos tienden hacia la media ⇒ En k generaciones, todos iguales.
Debe quedar claro que **es una falacia**, ya que no está bajando la varianza.

Med. Stat: introduction



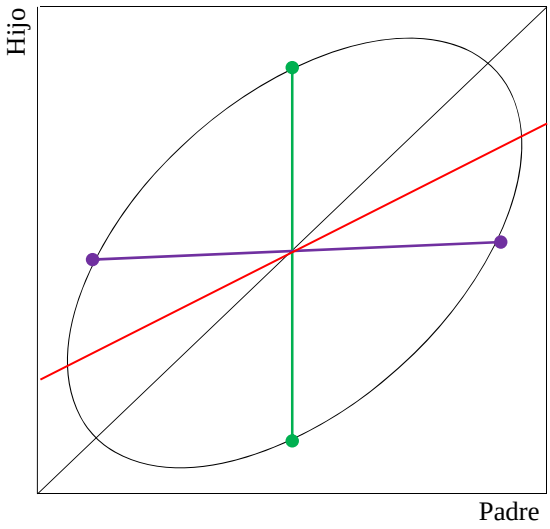
Regresión a la media

Excelentes en 1ª medida
no tan buenos en 2ª

Pésimos en 1ª medida
no tan malos en 2ª

Vuelta a la mediocridad

Med. Stat: introduction



Vuelta a la mediocridad es un fenómeno real

Pero descenso de variabilidad NO es un fenómeno real

RETO: distinguir entre evolución natural y efecto intervención

Med. Stat: introduction

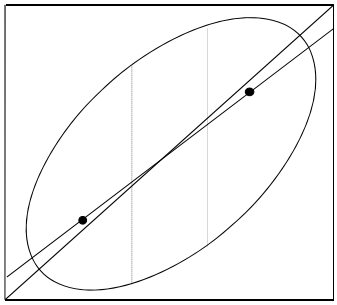
El fenómeno de regresión a la media depende de dos factores:

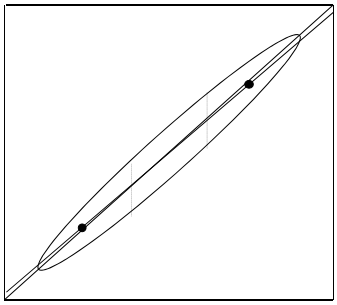
i) Si realizamos una selección 'más amplia' de altos y bajos,

- Disminuye la magnitud del fenómeno
- Disminuye el efecto de tender a la media

ii) Si disminuye la varianza del error:

- La elipse se estiliza.
- Aumenta la correlación.
- Disminuye el efecto de tender a la media.



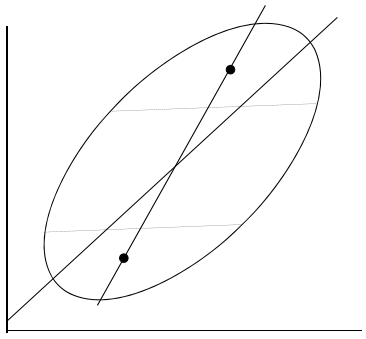


Med. Stat: introduction

Si intentamos predecir la altura de los padres a partir de la de los hijos:

- Se deberán minimizar las distancias horizontales (cambia la definición del error en mínimos cuadrados).
- Cambian los papeles (¡ahora son los padres los que tienden hacia la media!).

(el punto representa la "media horizontal")

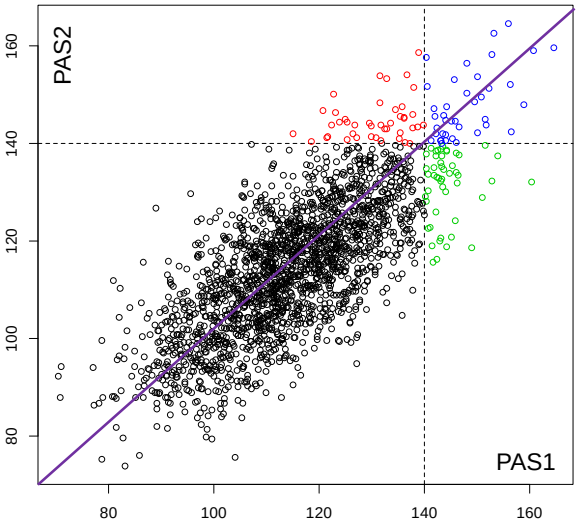


Med. Stat: introduction

Muestreo en dos fases. Se preguntó a 7000 chicos sobre su hábito tabáquico. Para una segunda fase, se escogió una muestra de 100 "fumadores", 100 "experimentadores" y 100 no fumadores, que volvieron a responder cuatro meses después sobre el tema. ¿Y cuáles fueron los resultados?

Second occasion	First occasion				Total
	>1/day	>1/week	occasional	never	
>1/day	15	5	2	0	22
>1/week	12	25	11	1	49
occasional	6	32	65	25	138
never	0	2	10	72	84
Total	33	64	98	98	293

Med. Stat: introduction



Variabilidad de PAS en 2 visitas

PAS1=PAS2

Siempre Hipertenso

Siempre Normotenso

Normotenso visita 1
Hipertenso visita 2

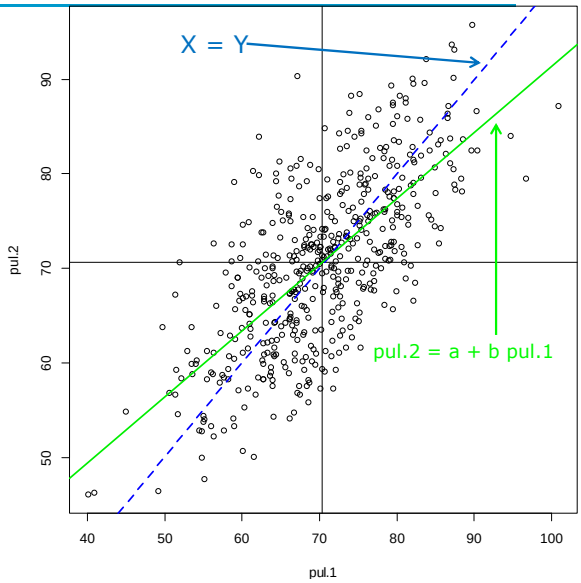
Hipertenso visita 1
Normotenso visita 2

Med. Stat: introduction

Ejemplo del pulso

El gráfico muestra las observaciones obtenidas por dos personas que miden el pulso cardíaco a N estudiantes. Las medidas no coinciden, pero fluctúan alrededor de una línea de identidad.

Sin embargo, las medias del pulso 2 condicionadas por el pulso 1 difieren de la identidad.



12

Med. Stat: introduction

Sea la repetición de una medida: X_1 y X_2.
Sea $A>E(X_1)$, entonces, $E(X_2|X_1>A) < E(X_1|X_1>A)$
Consecuencia de la fiabilidad de una variable: $X_1 = T + e$
Si seleccionamos a los casos por $X_1 > A$, no implica que $T > A$.

Asumiendo distribución normal bivariante con varianzas iguales:

$$E[(X_2-X_1)|X_1=a] = (a-\mu_{X1})(\rho-1)$$

Nótese que el fenómeno será mayor:
cuanto mayor sea la distancia del punto de corte A de la media y
cuanto menor sea la repetitividad de la variable en estudio.

Por **ejemplo**, sea $a=\mu_{X1}+2\sigma$, y $\rho=0'75$ $(=R_X)$, entonces
$E[(X_2-X_1)|X_1=a] = (\mu_{X1}+2\sigma-\mu_{X1})(0'75-1) = -\sigma/2$
El valor esperado de la próxima repetición será ¡media sigma inferior!

Soluciones: - ¡Grupo control!
- Aumentar la fiabilidad, haciendo medias de repeticiones.
- Cuantificar el cambio respecto a una segunda valoración.

Med. Stat: introduction

¿Es científicamente necesario aleatorizar?

Referencia interna: comparación "después frente a antes"

Retos:

- Evolución natural del proceso
- Otras intervenciones
- Difícil enmascarar
- Regresión a la media

Med. Stat: introduction

Regression toward the mean *refers to the fact that those with extreme scores on any measure at one point in time will probably have less extreme scores the next time they are tested for purely statistical reasons. Scores always involve a little bit of luck. Many extreme scores include a bit of luck that happened to fall with or against you depending on whether your extreme score is extremely high or extremely low.*

For example, imagine we want to examine students who score a perfect 100% correct and a perfect 100% wrong on some standardized test. Both of these scores are hard to get. To get 100% wrong, you would have to be wrong 100% of the time and even when you guess the answer, your guesses are wrong 100% of the time.

1. *Some of those people who scored 100% right guessed on some of the questions and guessed right. Conversely, some of those who scored 100% wrong guessed on some of the questions and guessed wrong.*

2. *The odds of all your guesses being right or wrong is pretty slim. By chance alone, the next time you flip a coin to decide which answer to pick, you will probably not guess 100% right or wrong again. If you guessed right 100% of the time last time, you probably won't be so lucky the next time (on average) and your score will drop a bit (towards the average, or mean). In fact, luck can't take you any higher than 100% right. If luck moves you at all, it moves you down (towards the mean). Similarly, if you scored 100% wrong the first time, you probably won't be so unlucky the next time. Again, you can't do worse than 100% wrong. So, by chance alone your score will go up the next time you take the test.*