REGRESIÓN Y CORRELACIÓN

José Rodríguez

4 de julio de 2016

1 ***Introducción:*** *Para fijar ideas, pensemos en un plan de ahorro sobre el cual tenemos que tomar la decisión de meternos en una hipoteca. Lo que nos interesa saber es si el plan de ahorro es lo suficientemente confiable para que no corramos el riesgo de regalarle todo al banco. Consideremos primero el caso de una persona que le descuentan su ahorro mensualmente directo por nómina en un empleo seguro. Éste es un caso simple: el plan de ahorro es estadísticamente confiable y podemos predecir cuándo reuniremos para la cuota inicial y después de pagarla dormiremos tranquilos sin preocupaciones de que no tengamos plata para pagar la mensualidad. La seguridad estadística puede disminuir en el caso de una persona que ahorra lo que puede. Si es una señorita muy disciplinada que ya estudió, eso es casi el caso anterior. Pero si está estudiando, lo más seguro es que tanto gasto compita con el ahorro. Pero como además los gastos no dan aviso, el ahorro mensual será muy variable haciendo que la predictibilidad disminuya y que no podamos decir cuándo reuniremos para la cuota inicial. Pensemos ahora en un muchacho que está en la época brincona: lo que hoy ahorró, mañana lo saca para invitar a una amiga. Y si algo le queda, pues es para irse de farra con sus amigos. Desde el punto de vista estadístico, éste plan de ahorro es nulo, lo cual quiere decir que si en un momento dado pareciera que el tipo está ahorrando, al siguiente uno se da cuenta de que está despilfarrando. Todo es muy azaroso, es realista pensar que el muchacho nunca levantará cabeza y que su situación en 20 años será la misma de hoy día: los saldos en rojo no lo abandonarán.*

Formalizando: tenemos un caso de regresión cuando esperamos una relación causa-efecto entre una variable estímulo (el tiempo) y una variable de respuesta de un sistema (el ahorro) y podemos hacer un experimento controlado. Pero además, si la magnitud de los cambios en la respuesta es proporcional a la de los cambios en el estímulo (cada mes una cuota fija de ahorro), tenemos una situación de regresión lineal.

2 ***Nuestro objetivo*** *es estudiar cómo se hace regresión lineal en presencia de ruido aleatorio. Estudiaremos detenidamente cuándo puede decirse que el que ahorra podrá levantar cabeza, es decir, cuándo es posible creer que en medio de los avatares de la vida podremos decir que estamos haciendo crecer el ahorro acumulado. Más exactamente, decidiremos la hipótesis nula que formaliza lo que esperamos del niño brincón. Rechazarla quiere decir que contamos con una disciplina de ahorro que algún día nos permitirá dejar de quedar con saldo en rojo al final de mes. Cuando no se trata de experimentos sino de observaciones, la relación entre dos variables (aleatorias) se estudia con la covarianza y el coeficiente de correlación.*

Nuestro camino comienza desde la teoría básica, para agarrar los conceptos que dan lugar a las aplicaciones y a la discusión, y después iremos a grandes pasos con solamente fórmulas y paquetes. Empezamos con un repaso de la línea recta porque para que podamos decir que tenemos un caso de regresión lineal los datos deben agruparse alrededor de una línea recta. Estarían sobre la línea si no hubiese ruido, pero es por su causa que los datos se escapan de la línea.

1. La línea recta

Hay dos maneras usuales de determinar una línea recta. Veamos la primera:

3 ***Ejemplo*** *Dos puntos determinan una recta.*

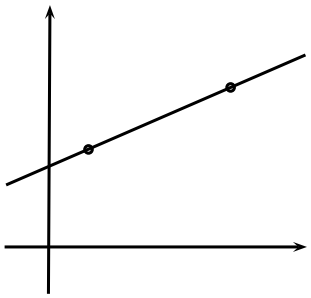

Figura 1. *Por dos puntos distintos pasa una línea y sólo una.*

La línea tiene una ecuación:

4 *La ecuación de una línea que no sea vertical es $y = \alpha + \beta x$. A x se la denomina la* ***variable independiente*** *y a y la* ***dependiente****. La ecuación de una línea vertical es de la forma $x = k$.*

5 ***Ejemplos*** *Hallemos las ecuaciones de algunas líneas.*

a) Pasa por los puntos $(1, 1)$ y $(2, 2)$. Podemos reemplazar los puntos en la ecuación $y = \alpha + \beta x$, obtener 2 ecuaciones y despejar. Reemplazamos:

$1 = \alpha + \beta$
$2 = \alpha + 2\beta$

Restando la de arriba de la de abajo obtenemos:

$1 = \beta$

Reemplazando en la de arriba:

$1 = \alpha + 1$ por lo que $\alpha = 0$. La ecuación es $y = x$.

b) Pasa por los puntos $(1, 5)$ y $(3, 11)$. Obtenemos 2 ecuaciones:

$11 = \alpha + 3\beta$
$5 = \alpha + \beta$

Restando : $6 = 2\beta$ por lo que $\beta = 3$ y por tanto $\alpha = 2$. La ecuación es $y = 2 + 3x$.

c) Pasa por los puntos $(0, 2)$ y $(4, -10)$. Obtenemos del primer punto que $2 = \alpha$ por lo que del segundo

$-10 = 2 + 4\beta$

por lo que $\beta = -3$. La ecuación es $y = 2 - 3x$.

d) Pasa por los puntos $(1, 1)$ y $(1, 5)$. Ésta línea es vertical. Todos sus puntos tienen el mismo valor de $x = 1$. Precisamente, ésta es la ecuación de dicha línea.

La segunda manera de determinar una recta es dando un punto y la pendiente o grado de inclinación.

6 ***Ejemplo*** *Significado geométrico de la pendiente.*

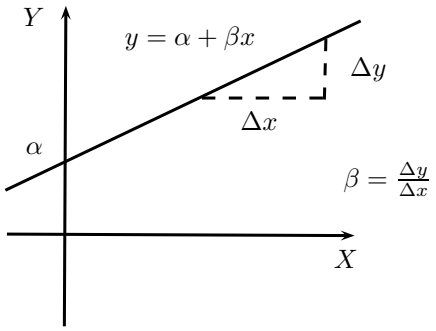

Figura 2. *La ecuación de la recta es* $y = \alpha + \beta x$. *Cuando* $x = 0$, $y = \alpha$ *por lo que* α *es el intersecto con el eje* Y. *Para hallar la pendiente se traza un triángulo rectángulo paralelo a los ejes, cualquiera, y la pendiente* $\beta = \frac{\Delta y}{\Delta x}$ *es igual al quebrado del incremento vertical entre el horizontal. Usamos letras griegas para enfatizar que son valores teóricos, poblacionales. Tenemos aproximadamente:* $\Delta y = 1$, $\Delta x = 1{,}5$ *por lo que* $\beta = \frac{1}{1{,}5} = 0{,}666$. *La pendiente es positiva si la línea es creciente, negativa si es decreciente.*

7 ***Ejemplo.*** *Tengo 100 ahorrados y cada mes logro ahorrar 5. Al cabo del primer mes tendré* $y = 100+5$. *Al cabo del segundo* $y = 100+5+5 = 100+5\times 2$. *Al cabo del tercero* $y = 100+5\times 3$. *Al cabo del mes* x *tendré* $y = 100+5x$. *Por tanto,* y *es el ahorro acumulado hasta el mes* x *inclusive. El intersecto es* 100 *que es el ahorro initial cuando* $x = 0$. *La pendiente es* 5, *el ahorro por mes que acrecienta el ahorro acumulado.*

8 ***Ejemplos*** *Hallemos las ecuaciones de algunas líneas determinadas por la pendiente y un punto.*

a) Pasa por $(1, 1)$ y su pendiente es 1. De la pendiente obtenemos que la ecuación es de la forma $y = \alpha + x$. Como pasa por $(1, 1)$ obtenemos $1 = \alpha + 1$ por lo que $\alpha = 0$. Por tanto, la ecuación es $y = x$.

b) Pasa por $(1, 1)$ y su pendiente es 0. De la pendiente obtenemos que la ecuación es de la forma $y = \alpha$. Como pasa por $(1, 1)$ obtenemos $1 = \alpha$ por lo que $\alpha = 1$. Por tanto, la ecuación es $y = 1$, lo cual dice que como no hay pendiente, no se sube ni se baja, la y permanece simpre la misma, $y = 1$.

2. Regresión lineal

Tenemos un caso de regresión lineal cuando se espera que el cambio en la variable dependiente sea proporcional al cambio en la variable independiente. Es decir, el cambio en la variable dependiente es igual a una constante por el cambio en la variable independiente.

9 ***Ejemplo*** *El carro.*

Entre más gasolina tenga un carro, más anda. En un terreno con subidas y bajadas es difícil decir algo más, pero en terreno plano uno puede ser más específico: el espacio recorrido por un carro es proporcional a la gasolina. Eso se escribe así: si x es la gasolina en galones y y es el espacio recorrido, $y = \beta x$. A x se le denomina la variable independiente, a y la variable dependiente y a β la constante de proporcionalidad. Es usual generalizar diciendo que $y = \alpha + \beta x$, lo cual quiere decir que desde el momento que empieza la observación, el carro ya había recorrido una cierta distancia α con respecto a un origen dado.

10 ***Ejemplo*** *El caminante.*

Uno espera que el espacio recorrido por un caminante sea aproximadamente proporcional al tiempo que el lleva caminando sin importar que a veces camine un poco más rápido o un poco más despacio. Sea x el tiempo que lleva caminando y y el espacio recorrido. Si el caminante fuese a una velocidad uniforme β, el espacio recorrido y se relaciona con el tiempo que lleva caminando x por

$$y = \alpha + \beta x$$

La constante α es el espacio inicial, es decir, la distancia entre el punto donde el caminante inicia su recorrido y un punto de referencia elegido.

11 ***Precaución: Nosotros usamos*** α ***para indicar dos cosas muy diferentes: es, por un lado, el nivel de significancia de una prueba y por otra es el intersecto o valor inicial en la regresión lineal, pero ninguna confusión será posible si uno presta atención al contexto.***

Pero como a veces se camina a una velocidad un poquito mayor que β y a veces un poco menor, podemos proponer que las variables estímulo-respuesta se relacionan mejor por:

$$y = \alpha + \beta x + \epsilon$$

donde ϵ es una v.a., pues uno nunca sabe qué podrá distraer o emocionar al caminante causando una desacelere o acelere en su caminar. Como de costumbre, el ruido ϵ se supone que tiene una distribución normal con media $\mu = 0$ y desviación σ. Una desviación grande indica que el caminante es muy variable en su velocidad.

Este modelo se denomina modelo de **regresión lineal** y se aplica cuando hay una clara relación de proporcionalidad causa-efecto entre la variable x y la variable y. Uno habla de **regresión no lineal** cuando hay relación causa efecto pero no de proporcionalidad. Por ejemplo, esperamos un relación causa-efecto entre el tiempo de estudio y la nota. Ahora bien, si con 1 hora de estudio saqué 1 y con 2 horas de estudio adicional la nota subió 2 puntos (saqué 3), entonces ¿podré esperar que con otras dos horas adicionales (5 en total) sacaré 5 de nota? Probablemente no, en general para sacar una nota promedio no hay que estudiar mucho pero en cambio para ser excelente hay que trabajar muy duro. Así que para sacar 5 en vez de 5 horas seguramente se necesitará trabajar 10.

Cuando nosotros no tenemos ninguna relación de causa-efecto pero observamos una relación lineal entre dos variables, nosotros usamos **correlación**. Por ejemplo, podría haber una correlación positiva entre el precio del dolar y la cantidad de muertes por infarto pero no hay relación causa-efecto. Con todo, uno puede atreverse a decir que, en general, por subir el dolar crecen los costos de muchos elementos críticos por lo que se sufre en demasía y la tensión sicológica adicional crea más infartos.

12 ***Ejemplo*** *La plata en el banco.*

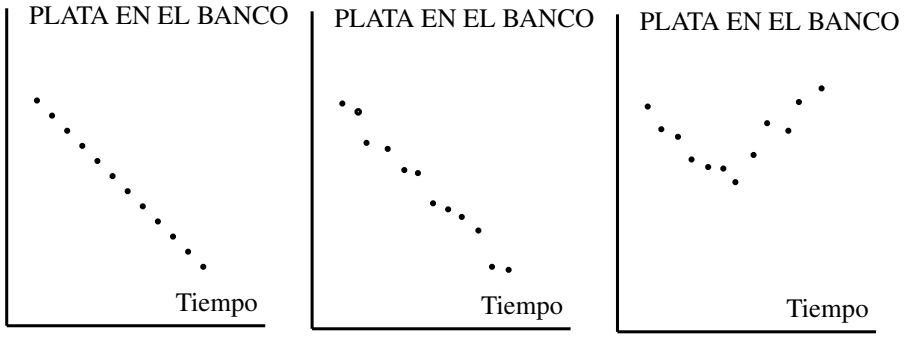

Figura 3. *La gráfica de la izquierda muestra la plata en el banco de una persona que retira y retira pero de manera estrictamente programada, cada mes una suma fija. Se tiene una regresión lineal sin ruido. En la figura central tenemos una persona que tiene un objetivo: retirar pero según la necesidad, pero como no hay mucho ruido ni tampoco otros objetivos, tenemos de todas formas un caso de regresión lineal. La figura de la derecha es una persona que maneja un tienda y ha pasado por un proceso de gaste y gaste el cual es seguido de un segundo estadío en el cual las entradas sobrepasan los gastos y así el puede ahorrar. Tenemos un caso de regresión pero no es lineal, parece más bien una parábola y oficialmente la denominamos regresión polinómica de segundo grado.*

Aprendamos el despliegue operacional de la regresión lineal.

13 ***Ejemplo*** *A veces se observa que el ingreso mensual es proporcional al tiempo de experiencia. En la vida real se nota que el profesional logra una época de maduración y es cuando sus ingresos aumentan generosamente, lo cual demanda unos 25 años de sufrimiento. Veamos si los siguientes datos, para pocos años de experiencia, se ajustan a un modelo lineal. Los datos los reportamos por pares de la forma (tiempo en años, sueldo en unidades arbitrarias): (3,5), (2,5), (2,4), (4,9) (3,7), (5,11).*

Solución:

Paso 1: Nosotros dibujamos los puntos sobre un plano cartesiano, lo cual se denomina un **diagrama de dispersión**.

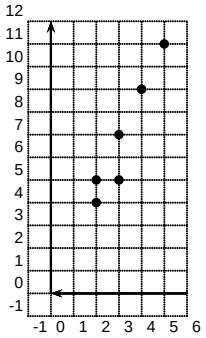

Figura 4. *Los años de experiencia en el eje horizontal, el sueldo en el eje vertical.*

Paso 2: Mirando el diagrama de dispersión, uno se forja una idea sobre aquella línea recta que mejor se ajusta a los datos, es decir, aquella que mejor parece representar los datos.

Cuando la línea viene de unos datos particulares, de una muestra, nosotros usamos para la línea la ecuación con letras latinas:

$y = a + bx$.

Pero cuando nosotros formulamos un modelo, el cual es algo teórico y que debe corresponder a un muestreo con un número infinito de datos, usamos letras griegas

$y = \alpha + \beta x$

En nuestro caso, la línea que mejor se ajusta a nuestros datos parece pasar por el punto más cercano al origen y por el más lejano. Esos puntos son: (2,4) y (5,11), los cuales nos generan una línea (en letras latinas):

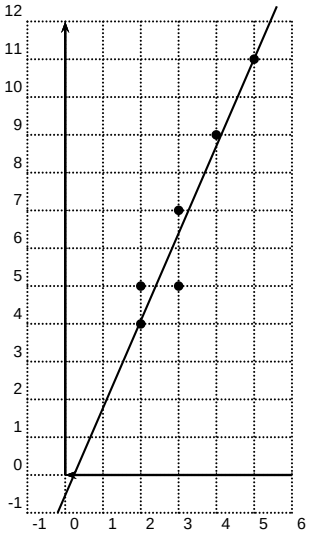

Figura 5. *Estimación visual de la línea que mejor representa los datos, aquella que causa un mínimo de descontento global.*

La pendiente de dicha línea es

$b = \frac{y_2 - y_1}{x_2 - x_1} = \frac{11-4}{5-2} = \frac{7}{3} = 2{,}33$
Por lo tanto, la ecuación de la línea es
$y = a + bx = a + 2{,}33x$
Para hallar a estudiamos un punto cualesquiera, por ejemplo, (2,4):
$4 = a + 2{,}33(2) = a + 4{,}66$
por lo que $a = 4 - 4{,}66 = -0{,}66$
Por consiguiente, nuestra **estimación visual de la línea de regresión** es
$y = -0{,}66 + 2{,}33x$
Mirando la nueva gráfica, vemos que es razonable creer que el ingreso mensual es proporcional a la experiencia, es decir que se justifica un modelo lineal, pues los puntos se ajsustan bien a la línea.

Ahora pasamos de lo intuitivo a los métodos rigurosos para sacar la línea de regresión de mínimos cuadrados. La metodología es como sigue:

Paso 3: Construímos la siguiente tabla:

Regression de ingresos mensuales (y) vs tiempo de experiencia x					
	x	y	xy	x^2	y^2
	3	5	15	9	25
	2	5	10	4	25
	2	4	8	4	16
	4	9	36	16	81
	3	7	21	9	49
	5	11	55	25	121
Sumas	$\sum x$ =19	$\sum y$ =41	$\sum xy$ =145	$\sum x^2 = 67$	$\sum y^2$ =317

Paso 4: Calculamos los valores medios tanto de y como de x. Como en este caso tenemos 6 pares de datos:
$\bar{x} = \sum x/6 = 19/6 = 3{,}17$
$\bar{y} = \sum y/6 = 41/6 = 6{,}83$

Paso 5: Ahora nosotros podemos calcular la línea recta, $y = a + bx$, que mejor se ajusta a los datos, la cual minimiza el descontento cuadrático global de todos y cada uno de los puntos al ser representados por una línea.

14 ***Notación*** *La siguiente notación ayuda a hacer muchos cálculos relacionados con regresión:*

$SS_{xx} = \sum x^2 - n\bar{x}^2$;
$SS_{xy} = \sum xy - n\bar{x}\bar{y}$;

Con esta notación, la línea de regresión (min cuadrados): $y = a + bx$ se determina por

$b = \frac{SS_{xy}}{SS_{xx}} = \frac{\sum xy - n\bar{x}\bar{y}}{\sum x^2 - n\bar{x}^2}$;
$a = \bar{y} - b\bar{x}$.

Para nuestro ejemplo, esto se convierte en

$b = \frac{145 - 6(3{,}17)(6{,}83)}{67 - 6(3{,}17)^2} = \frac{15{,}09}{6{,}71} = 2{,}24$
$a = \bar{y} - b\bar{x}$
$a = 6{,}83 - (2{,}24)(3{,}17) = 6{,}83 - 7{,}10 = -0{,}27$

Así, la línea de regresión de mínimos cuadrados es
$y = -0{,}27 + 2{,}24x$
Comparamos esta respuesta con la que habíamos sacado visualmente: $y = -0{,}66 + 2{,}33x$

Comparando las dos ecuaciones, la sacada por cálculo gráfico y la sacada por fórmulas, vemos que hay concordancia y por tanto podemos confiar en que hemos hecho bien las cosas. Observemos que cuando no hay experiencia uno no gana sueldo sino que pierde plata y tiempo en buscar trabajo y asistir a entrevistas fallidas. Por eso el intersecto es negativo. Pero con la experiencia uno empieza a valorarse y entre mayor experiencia y preparación, más se gana.

Con la ecuación de la línea uno puede atreverse a hacer extrapolaciones. Por ejemplo, para éste ejemplo, podemos decir que cuando x valga 10, y valdrá $y = -0{,}27 + 2{,}24x = -0{,}27 + 2{,}24 \times 10 = -0{,}27 + 22{,}4 = 22{,}13$ Eso sería el sueldo posible de un super-experto. Una predicción es más confiable entre más cerca esté del rango de valores y menos confiable entre más dispersión de los datos fuera de la línea. Por ejemplo, si hay datos para los primeros 5 meses, pronosticar que pasará al mes 6 es algo confiable bastante más que al mes 10. Por otro lado, si los datos están muy dispersos por haber muchos factores azarosos que afectan el resultado, las extrapolaciones son poco confiables así sean de un mes después.

En lo que sigue del capítulo le añadiremos rigor y modernidad a lo dicho.

3. Método de mínimos cuadrados

Hemos visto el cómo de la línea de regresión pero no hemos visto el por qué. Es conveniente saber el por qué para poder entender de qué forma los resultados dependen de un método y así poder dejar la puerta abierta para la investigación de otras metodologías.

Una noción intuitiva de la línea de regresión se puede lograr con el siguiente experimento mental: uno ata con un resorte cada punto a una varilla pero asegurándose de que todos los resortes queden bien tensionados. Después uno libera el sistema a su propio destino hasta que la tensión blobal se minimice. La posición final de la varilla estará globalmente lo más cerca posible de todos los puntos y nos dará una idea de la línea de regresión.

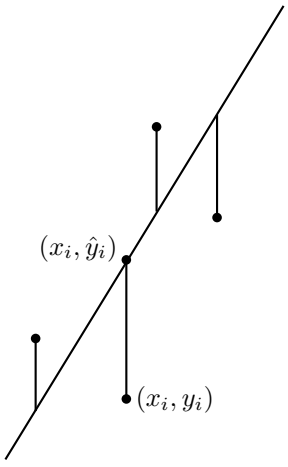

Figura 6. *Método de los resortes para hallar la línea de regresión que mejor se ajuste a un conjunto de puntos dado. Se ata un resorte desde cada punto a una varilla recta. Se suelta el sistema para que la tensión global se minimice. Se tiene cuidado de que los resortes queden verticales en su posición definitiva. La posición final de la varilla representará la línea buscada. La longitud del resorte asociado a un punto se llama el* ***residuo.***

En vez de resortes, podemos hacer una variante que consiste en reemplazar los resortes por hilo y templarlos a mano con el único requisito de que la cantidad total de hilo sea la mínima. Esto nos producirá la línea de regresión por el método del hilo templado. Y como ya tenemos dos metodologías podemos preguntarnos: ¿darán los dos métodos la misma línea en todos los casos? Si uno comete errores pequeños en la metodología de alguno de ellos, ¿existirán grandes cambios en la respuesta?

Hemos, pues, formulado dos preguntas importantes y para poder estudiarlas necesitamos formalizar las metodologías. El método de los resortes tiene como objetivo minimizar la tensión global mientras que el método del hilo templado tiene como objetivo minimizar el hilo total gastado. Sea (x_i, y_i) el punto número i, donde hay n puntos. Sea $y = a+bx$ la ecuación de una línea cualquiera que no sea vertical. El punto sobre la línea que queda en la dirección vertical de (x_i, y_i) es $(x_i, \hat{y}_i) = (x_i, a + bx_i)$. La cantidad de hilo entre (x_i, y_i) y $(x_i, \hat{y}_i)$ es $\mid y_i - \hat{y}_i \mid = \mid y_i - (a + bx_i) \mid = \mid y_i - a - bx_i \mid$.

El hilo total gastado es entonces

$$H = \sum_i^n \mid y_i - a - bx_i \mid$$

La línea de regresión por el método del hilo tiene que minimizar la función H sobre todos las líneas, es decir, sobre todos las dupletas de la forma (a, b). Hay muchos métodos para resolver este problema. Uno de ellos es por ensayo y error: se toma una pareja (a, b), se calcula H. Se toma otra pareja y se vuelve a calcular H. Uno compara cual H es el menor y se queda con el (a, b) correspondiente. Después una prueba otra parjea y así hasta que uno se canse. O quizá uno quiera ayudarse de un programa en Java.

Veamos ahora qué sucede con el método de los resortes. El objetivo en este caso es minimiar la tensión global. La tensión de cada cuerda es la fuerza generada por su elongación que es proporcional a la distancia entre el punto y la línea, pues entre más se elonge un resorte, más fuerza hace. Es decir, la tensión asociada al punto (x_i, y_i) es proporcional a la distancia vertical entre dicho punto y la línea, lo cual da $k \mid y_i - \hat{y}_i \mid = k \mid y_i - (a + bx_i) \mid = k \mid y_i - a - bx_i \mid$. Por ende, la tension global es

$$T = k \sum_i^n (signo) \mid y_i - a - bx_i \mid$$

El problema con el signo es el siguiente: la fuerza es un vector con una dirección y la suponemos que siempre se dirige hacia la línea. Así, si un punto está encima de la línea, la tensión va hacia abajo y se está abajo de ella, la tensión va hacia arriba. El objetivo de la regresión lineal pr el método de los resortes es minimizar la tensión global T. Aunque este problema sea más laborioso que el de los hilos, también puede resolverse por ensayo y error o uno podría ayudarse con un algoritmo genético, que simula la evolución biológica con el ánimo de resolver problemas de matemáticas.

El problema con los métodos del hilo y de los resortes es que involucran al valor absoluto y signos que indican dirección. Para hacer una especificación del valor absoluto hay que ver si se trata de un número positivo y dejarlo igual o si de uno negativo y cambiarle el signo. Todo eso es muy engorroso. Existe un truco que siendo simple permite liberarse de los inconvenientes del valor absoluto y es tomar los cuadrados en vez del valor absoluto. Físicamente, éso corresponde a guiarse por la energía potencial en el sistema de resortes y no por la fuerza o por la cantidad de hilo gastada. Formalzando:

El método de regresión de los mínimos cuadrados asociado a un conjunto de n puntos de la forma (x_i, y_i) produce una función E que mide el error cuadrático dado por

$$E = \sum_i^n (y_i - a - bx_i)^2$$

y su objetivo es hallar la pareja (a, b) que minimice dicha función. A la línea correspondiente $y = ax+b$ se la llama **línea de regresión por el método de mínimos cuadrados.** A los matemáticos les gusta el método del error cuadrático más que el de los resortes y el del hilo porque la función E puede minimizarse por derivadas. Para hallar un máximo o mínimo local de una función derivable definida en un intervalo abierto, sin los extremos, lo primero que hay que hacer es derivar e igualar a cero, pues la derivada da la pendiente de la línea tangente al punto. Cuando uno tiene un máximo o un mínimo, la línea tangente es horizontal:

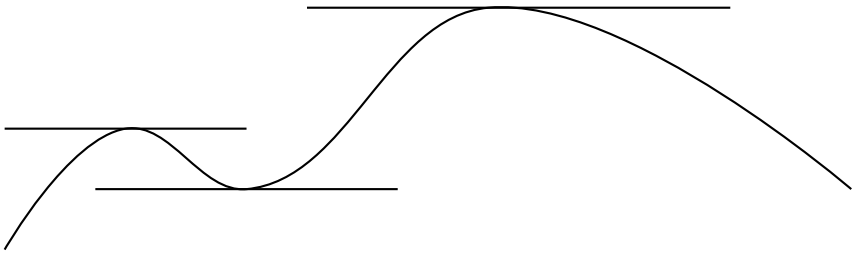

Figura 7. *Sobre una cruva suave, los máximos y mínimos que no estén en los extremos de la curva se hallan derivando e igualando a cero, pues la derivada da la pendiente de la línea tangente, y sobre un máximo o mínimo dicha línea es horizontal, con pendiente cero.*

Cuando uno tiene una función de una variable, la derivada es una derivada ordinaria:

$f(x) = 3x^2 + 5x + 7$
$f'(x) = 6x + 5$

Pero cuando uno tiene una función de dos variables, la derivada es una derivada parcial que significa que estamos no en una curva sino en una montaña en un mundo de tres dimensiones y ya no se toman línea tangetes horizontales sino planos tangentes horizontales.

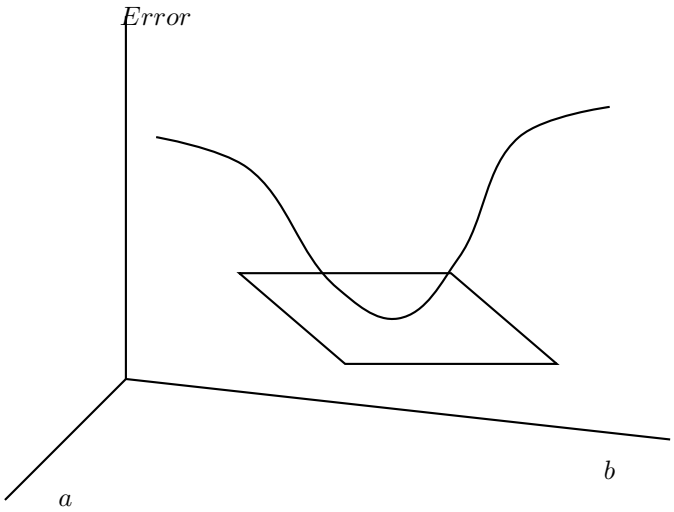

Figura 8. *La función Error depende de dos variables* (a, b)*. Su gráfica es como un valle en un espacio tridimensional. Para hallar los mínimos hallamos los puntos donde el plano tangente sea horizontal, lo cual implica que las derivadas parciales sean iguales a cero.*

La derivada parcial de una función de varias variables con respecto a una de ella se calcula como si fuese una derivada ordinaria pero asumiendo que todas las demás variables tienen un valor fijo o constante. Ejemplo: si $f(x, y) = 3x^2y + 2xy - x^2 + y^3 - 8$ entonces la derivada parcial de f con respecto a x se nota $\partial f/\partial x$ y es

$\partial f/\partial x = 6xy + 2y - 2x + 0 + 0$

En efecto, la derivada parcial con respecto a x de $3x^2y$ es $3y(2x) = 6yx = 6xy$ pues $3y$ se toma como constante. De igual modo, la derivada parcial con respecto a x de y^3 da cero, lo mismo que la de 8, pues la derivada de una constante es cero.

Similarmente, si $f(a,b) = (3a+4b)^2$

$\partial f/\partial a = 2(3a+4b)(3)$

pues hay que tener en cuenta la derivada interna que con respecto a a es 3.

Minimicemos ahora la función E por derivación parcial igualada a cero.

$E = \sum_i^n (y_i - a - bx_i)^2$

Derivemos con respecto a a:

$\partial E/\partial a = \sum_i^n 2(y_i - a - bx_i)(-1) = 2\sum_i^n (y_i - a - bx_i)(-1) = 2\sum_i^n (-y_i + a + bx_i) = 0$

dividiendo por dos y reorganizando:

$\sum_i^n -y_i + \sum_i^n a + b\sum_i^n x_i = 0$

$a\sum_i^n 1 + b\sum_i^n x_i = \sum_i^n y_i$

$na + b\sum_i^n x_i = \sum_i^n y_i$

Como $\bar{x} = (\sum_i^n x_i)/n$, y también $\bar{y} = (\sum_i^n y_i)/n$, entonces $(\sum_i^n x_i) = n\bar{x}$ y $(\sum_i^n y_i) = n\bar{y}$, y por tanto:

$na + nb\bar{x} = n\bar{y}$

Dividiendo por n

$a + b\bar{x} + = \bar{y}$

y despejando a obtenemos:

$a = \bar{y} - b\bar{x}$

Derivemos con respecto a b:

$\partial E/\partial b = \sum_i^n 2(y_i - a - bx_i)(-x_i) = 2\sum_i^n (y_i - a - bx_i)(-x_i) = 2\sum_i^n (-x_i y_i + ax_i + b(x_i)^2) = 0$

dividiendo por dos y reorganizando:

$\sum_i^n -x_i y_i + a\sum_i^n x_i + b\sum_i^n (x_i)^2 = 0$

$a\sum_i^n x_i + b\sum_i^n (x_i)^2 = \sum_i^n x_i y_i$

$an\bar{x} + b\sum_i^n (x_i)^2 = \sum_i^n x_i y_i$

$(\bar{y} - b\bar{x})n\bar{x} + b\sum_i^n (x_i)^2 = \sum_i^n x_i y_i$

$n\bar{x}\bar{y} - nb(\bar{x})^2 + b\sum_i^n (x_i)^2 = \sum_i^n x_i y_i$

$b[-n(\bar{x})^2 + \sum_i^n (x_i)^2] = \sum_i^n x_i y_i - n\bar{x}\bar{y}$

$b = \frac{\sum_i^n x_i y_i - n\bar{x}\bar{y}}{-n(\bar{x})^2 + \sum_i^n (x_i)^2} = \frac{\sum_i^n x_i y_i - n\bar{x}\bar{y}}{\sum_i^n (x_i)^2 - n(\bar{x})^2} = \frac{SS_{xy}}{SS_{xx}}$

donde

$SS_{xy} = \sum_i^n x_i y_i - n\bar{x}\bar{y}$
$SS_{xx} = \sum_i^n (x_i)^2 - n(\bar{x})^2$

Y ésta es la historia de la línea de regresión por el método de mínimos cuadrados. Todo lo demás de este capítulo se edifica sobre esta misma metodología, la cual se aplica para una variable lo mismo que para muchas.

4. Test para la relación funcional

La gráfica siguiente nos puede ayudar a entender qué es lo que queremos hacer.

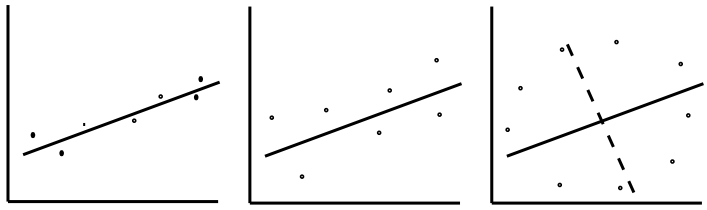

Figura 9. *En la gráfica de la izquierda vemos que los datos permiten visualizar una línea creciente como la línea de mejor ajuste. Decimos que el ruido es pequeño porque el desajuste de los puntos a la línea es mínimo. Como el ruido es pequeño, tenemos poder predictivo basándonos en la línea. En la gráfica central el ruido es mayor, el desajuste de los puntos con respecto a la línea es más grande, pero no queda exagerado decir que tenemos una línea de pendiente positiva. Pero en la gráfica de la derecha el ruido es tan grande que si bien la línea en trazo sólido parece que es el mejor ajuste lineal de los datos, uno también podría pensar que los datos también se ajustan a la línea punteada. Y cualquier otra pendiente también sonaría bien. En un caso así, cuando el ruido no permite oír la información contenida en una línea, lo mejor es suponer que el mejor modelo lineal está dado por una línea horizontal, lo cual daría a entender que la variable de salida no se ve afectada por la de entrada y que por tanto cuando la variable de entrada varía, la de salida permanece constante. Como el ruido es tan alto, al repetir el experimento unas veces dará una línea de pendiente positiva y otras negativa: no hay predecibilidad sobre el signo de la pendiente. Eso es lo que significa que la pendiente de la línea de regresión sea cero.*

El modelo de regresión lineal $y = \alpha + \beta x + \epsilon$ asume una dependencia funcional de y con respecto a x. Por consiguiente, la hipótesis nula natural en un estudio de regresión es que y de hecho no depende de x. Eso quiere decir que cuando x cambia, la y parece no oír, es decir, cuando la x cambia, la y permanece como estaba, excepto los cambios debidos al ruido aleatorio. Eso implica que la y permanece constante, horizontal. Pero como hay ruido, la y irá una vez arriba y otra vez abajo de su promedio. Tenemos, pues, una disyuntiva:

$H_o : \beta = 0$, (x no afecta linealmente a y. Aunque x cambie, y no cambia de forma consistentemente lineal).
$H_a : \beta \neq 0$, (y cambia proporcionalmente a los cambios en x)

Hay dos procedimientos para probar esta H_o, por una t y por una F, que es el vamos a ver. La necesidad de hacer la prueba se justifica porque el ruido puede causar que aparezcan relaciones espurias, que no vuelven a aparecer cuando se repita el experimento. ¿Pero si el ruido puede crear relaciones espurias, cómo podemos asegurarnos de que una relación funcional existe? Lo podemos hacer porque el ruido tiende a autoaniquilarse: la tendencia sistemática tiende a verse más nítida entre más datos haya, y para un cierto volumen de datos, el ruido puede ser tolerable y no impedir que se vea el efecto de la **variable independiente o experimental** sobre la **variable dependiente o respuesta**.

Aprendamos a usar una F, la que nos sirvió para comparar varianzas, para dilucidar si $\beta = 0$ o no.

La forma como nosotros sabremos que y depende de x es moviendo el valor de x y observando el efecto sufrido por y. Si y de forma consistente varía al variar la x, nosotros podremos concluir que una relación funcional existe, pero si la variación observada de y pudiese atribuírse al efecto del ruido, no habría necesidad de suponer entonces que una relación funcional entre x y y existe. La variación de y se da por:

La suma total de cuadrados (total Sum of Squares) = SS Total = $\sum(y_i - \bar{y})^2 = \sum y_i^2 - \frac{(\sum y_i)^2}{n}$

Ahora denotamos por $\widehat{y}$ el valor $\alpha + \beta x_o$ que es el valor de la línea en x_o. Observando que

$0 = -\widehat{y} + \widehat{y}$

podemos insertar ese cero en el SS total como sigue:

SS total= $\sum(y_i + 0 - \bar{y})^2 = \sum(y_i - \widehat{y} + \widehat{y} - \bar{y})^2 == \sum((y_i - \widehat{y}) + (\widehat{y} - \bar{y}))^2$
$= \sum(y_i - \widehat{y})^2 + 2\sum(y_i - \widehat{y})(\widehat{y} - \bar{y}) + \sum(\widehat{y} - \bar{y})^2$

15 **Ejercicio** *Pruebe que siempre se tiene que:* $2\sum(y_i - \widehat{y})(\widehat{y} - \bar{y}) = 0$

Aplicando este resultado, obtenemos:

16 ◇ ***Teorema:*** *SS Total* $= \sum(y_i - \bar{y})^2 = \sum(y_i - \widehat{y})^2 + \sum(\widehat{y} - \bar{y})^2$.

Esta ecuación nos permite desentrañar lo que es información de lo que es ruido. Pero cada modelo tiene su definición específica de ruido. En nuestro caso, ruido es cualquier cosa que cause que un evento se desvíe del modelo de regresión lineal, es decir, el efecto del ruido coincide con el término $(y_i - \widehat{y})^2$. Por otra parte, $(\widehat{y} - \bar{y})^2$ representa la información pues dice que el modelo lineal se aparta de ser constante e igual al promedio de y. Recordemos que el promedio de y es el bastión de la H_o: la y no escucha los cambios de x, sino que permanece constante e igual a su promedio, aparte de los cambios debidos al ruido.

17 ◇ ***Teorema:*** *En nuestro modelo, la información y el ruido son ortogonales, i.e., el ruido no interfiere con la información y, por tanto, ambos conceptos están nítidamente definidos.*

Vemos que el término $\sum(\widehat{y} - \bar{y})^2$ es la variación atribuible a la dependencia lineal. Así que, la idea es que esta variación sea lo suficientemente grande como para ser considerada como relevante. ¿Pero, grande con respecto a que? Pues a la variabilidad causada por el ruido que se da por $\sum(y - \widehat{y})^2$. Para facilitar los cálculos, nosotros podemos usar el próximo resultado.

18 **Definición y teorema** *Si definimos SS de regresión mediante la identidad*

SS de regresión= $\sum(\widehat{y} - \bar{y})^2$*, lo que se aparta consistentemente del promedio.*

y si definimos el SS del ruido como

SS del ruido $= \sum(y - \widehat{y})^2$*, lo que se aparta de la línea de regresión,*

entonces

a) SS Total= $\sum y^2 - \frac{(\sum y)^2}{n}$

b) SS de regresión $= \frac{(\sum xy - \frac{\sum x \sum y}{n})^2}{\sum x^2 - \frac{(\sum x)^2}{n}}$

c) SS del ruido= SS total - SS de regresión.

Atención: para efectos operacionales es mejor utilizar

b') SS de regresión $= \frac{(SS_{xy})^2}{SS_{xx}}$

en donde:

$SS_{xy} = \sum xy - n\bar{x}\bar{y}$
$SS_{xx} = \sum (x)^2 - n(\bar{x})^2$

lo cual se demuestra así:

SS de regresión $= \frac{(\sum xy - \frac{\sum x \sum y}{n})^2}{\sum x^2 - \frac{(\sum x)^2}{n}}$
$= \frac{(\sum xy - n\frac{\sum x}{n}\frac{\sum y}{n})^2}{\sum x^2 - n\frac{\sum x}{n}\frac{\sum x}{n}}$
$= \frac{\sum xy - n\bar{x}\bar{y})^2}{\sum (x)^2 - n(\bar{x})^2}$
$= \frac{(SS_{xy})^2}{SS_{xx}}$

Ahora bien, bajo la hipótesis nula, la variable independiente x no influye para nada sobre la variable dependiente y, sino que toda posible variación es nada más que un efecto de los factores no controlados, que genéricamente se denomina azar o ruido y que se modelan por la v.a. ϵ que tiene media 0 (pues es azar y por tanto no es ni fu ni fa) pero varianza σ^2. Por consiguiente, bajo la hipótesis nula, la única fuente de variación con respecto a la media global es el azar. Esto nos permite decir que tenemos dos tipos de variación, una registrada por *SS de regresión* y la otra por *SS de ruido*. Pero ambas, bajo la H_o, tienen una única fuente el ruido, σ^2. La gran idea es ahora transformar estas dos variaciones en varianzas. Curiosamente, todo lo que se necesita para obtener varianzas es dividir dichas variaciones por números adecuados, llamados grados de libertad.

Ahora, podemos definir un estadígrafo para comparar la información contra el ruido. Definimos por tanto:

19 ♣ ***Definición.*** $R_{exp} = \frac{\frac{SS\ de\ Regresión}{1}}{\frac{SS\ de\ ruido}{n-2}}$

Debemos ahora predecir lo que se espera que valga R_{exp} bajo la H_o. Cuando no hay más que azar, no hay otras fuentes de variación. Por tanto, las varianzas que entran en la R_{exp} deben estar ambas relacionadas con el azar. Se puede demostrar que dichas varianzas son ambas estimadores (insesgados) de la varianza del azar, σ^2. Por tanto, bajo la H_o, y si hubiese un número infinito de datos, el valor esperado o promedio de R_{exp} sería 1. Podemos ahora comparar entonces lo que se ve, R_{exp} con lo que se cree, que $R = 1$.

20 ◇ ***Teorema:*** *Cuando la hipótesis nula, $\beta = 0$, es correcta el estadígrafo*

$F_{exp} = \frac{R_{exp}}{R} = \frac{R_{exp}}{1} = \frac{\frac{SS\ de\ Regresión}{1}}{\frac{SS\ de\ ruido}{n-2}}$

que relaciona lo que se ve con lo que se cree bajo la Ho, se distribuye como una F con 1 g.l. para el numerador y $n-2$ g.l. para el denominador. Si los grados de libertad son infinitos en número, el estadígrafo F_{exp} toma bajo la H_o el valor uno. Por lo tanto, la región de aceptación de la H_o contendrá el UNO. Por otro lado, cuando la variable independiene x afecta la variable dependiente y, su efecto se mide a través de SS de regresión: entre más pronunciado sea el efecto de x sobre y, más grande será este término y la F_{exp} será significativamente más grande que UNO. Por ésa razón, la H_o se rechaza con una cola, la cola superior.

21 ***Ejemplo*** *Imaginemos que dos pueblos hermanos forman colonias separadas. Sus formas de hablar divergirán e iran formando lenguas diferentes. La diferencia se cuantifica por una distancia: se le da un texto de un lengua a un representante de la otra población y se le examina sobre el grado de comprensión obtenido. Entre menos comprensión, más distancia. Los siguientes pares de datos (1,1)(2,2)(3,1)(4,2)(5,3)(6,2) (7,2) (8,2) indican el tiempo, en centurias, y la distancia entre dos lenguas. Si las dos lenguas van cada una por su lado, la distancia entre ellas crecería con el tiempo pero si las dos lenguas están ligadas debido a una estrecha relación entre sus hablantes, es posible que ellas evolucionen pero de forma sincronizada y la distancia no aumentará, sino que cambiará erráticamente alrededor de un promedio. Nuestra H_o es que con el tiempo no habrá separación: $\beta = 0$. Decidamos con $\alpha = 0{,}02$ y dos colas si esto es verdad o no. Observemos que estamos discutiendo un proecso histórico en el cual el tiempo está dado por un reloj y por lo tanto no es aleatorio en tanto que la distancia entre las lenguas puede depender de muchos factores no controlados y por tanto se trata de una variable aleatoria. Tenemos el ambiente perfecto para el estudio de correlación.*

Solución: Hagamos primero el diagrama de dispersión:

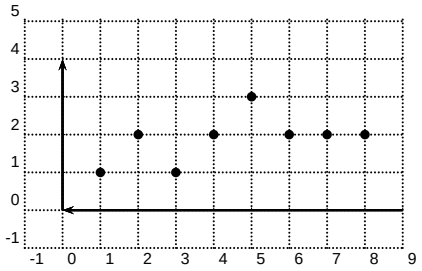

Figura 10. *La línea que mejor representa la H_o es, estimada visualmente, $y = 2$. Como esa línea es horizontal, la H_o se acepta, la pendiente β es cero.*

A ojímetro vemos del diagrama de dispersión que los valores giran erráticamente alrededor de la línea horizontal $y = 2$. Por lo tanto, no habría dependencia funcional de la variable dependiente de la independiente. Veamos ahora cómo se procede a tomar el veridcto con el rigor de la metodología moderna. Construyamos la tabla de regresión:

Regresión de distancia de separación y vs tiempo x					
	x	y	xy	x^2	y^2
	1	1	1	1	1
	2	2	4	4	4
	3	1	3	9	1
	4	2	8	16	4
	5	3	15	25	9
	6	2	12	36	4
	7	2	14	49	4
	8	2	16	64	4
Sumas	$\sum x$ =36	$\sum y$ =15	$\sum xy$ =73	$\sum x^2 = 204$	$\sum y^2$ =31

Esta tabla nos da la siguiente numerología:

```
n = 8
MediaX = 4.5
MediaY = 1.875
SumX2 = 204.0
SumY2 = 31.0
SumXY = 73.0
SSxx = sumX2 - n * mediaX*mediaX
SSxx = 42.0
```

```
SSxy = sumXY - n * mediaX*mediaY
SSxy = 5.5
b = SSxy/SSxx
b = 0.131
a = mediaY - b * mediaX  = 1.28
```

Así, la línea de regresión de mínimos cuadrados es

$y = 1{,}28 + 0{,}131x$

Lo que haremos es decidir si la pendiente registrada de 0,131 representa un efecto real de x sobre y ó si es una relación espúrea creada por el azar y que seguramente no se repitirá con un nuevo experimento. Veamos.

Las sumas de cuadrados son:
SS Total= $\sum y^2 - \frac{\left(\sum y_i\right)^2}{n} = 31 - \frac{(15)^2}{8} = 2{,}875$
SS de regresión= $\frac{(\sum xy - \frac{\sum x \sum y}{n})^2}{\sum x^2 - \frac{(\sum x)^2}{n}} = \frac{73 - \frac{(36)(15)}{8}}{204 - \frac{(36)^2}{8}} = \frac{30{,}25}{42} = 0{,}7202$
SS de ruido= SS Total - SS de regresión= 2.875-0.7202= 2.1548.
Los grados de libertad: para regresión, 1; para ruido, 8-2=6.

$F_{exp} = \frac{\frac{SS\ de\ regresión}{1}}{\frac{SS\ de\ ruido}{n-2}} = \frac{\frac{0{,}7202}{1}}{\frac{2{,}1548}{6}} = \frac{0{,}7202}{0{,}3591} = 2{,}005$

Para $\alpha = 0{,}02$ y dos colas, el valor crítico de F con una cola y con 1 y 6 g.l. es 9.87. Nuestro estadígrafo dio 2.005, lo cual dice que la discrepancia entre lo que se espera en la H_o y lo que se ve es pequeña. Concluimos por tanto que aquí no tenemos datos suficientes para pretender que hay una regresión lineal justificada.

Veredicto: Las lenguas estudiadas seguramente están cambiando, pero la distancia entre ellas es constante en el tiempo. Es decir, las lenguas co-evolucionan, van como dos hermanitas cogidas de la mano a pesar de que son un poco distintas, lo cual se sabe pues la línea de regresión no es el eje X con ecuación $y = 0$ sino que queda un poco por encima.

Nuestro veredicto dice que la mejor manera de explicar los datos de la muestra es diciendo que la distancia entre las lenguas es constante a través del tiempo y de 1,875, pero debido a que el análisis literario de la muestra no cubrió todos las facetas del lenguaje a toda hora, por puro azar se generó la impresión de que había una pequeña divergencia lineal creciente dada por $b = 0{,}131$.

22 ***Interpretación general del veredicto***

Tenemos en general que cuando se acepta la hipótesis nula de que la pendiente de la línea de regresión es horizontal, lo que estamos diciendo es que al repetir muchas veces el experimento podemos esperar que la nueva pendiente sea unas veces positiva y otras negativa. Es decir, no tenemos poder predictivo en cuanto al signo de la pendiente. Pero cuando se rechaza la hipótesis nula con una cola y se infiere que, por ejemplo, la pendiente es positiva, entonces tenemos base para decir que al repetir muchas veces el experimento, la pendiente será una muy buena proporción de veces positiva. Acá tenemos poder predictivo.

5. Predicciones y el error estandar

La primera utilidad de la línea de la regresión está en la predicción del valor que tomará la función para un valor determinado.

23 ***Ejemplo*** *Si un modelo de mínimos cuadrados predice una relación causa (x) -efecto (y) dada por y:*

$y = -0.27 + 2.24x$

y deseamos saber el valor de la respuesta para $x = 2.5$*, entonces nuestra predicción dice que:*

$y(2.5) = -0.27 + 2.24(2.5) = -0.27 + 5.6 = 5.33$

Nos preguntamos: ¿qué tan seriamente debemos tomar esta predicción?

Para responder la pregunta debemos tener en cuenta que nuestro sistema experimental no está aislado sino que sufre el efecto del medio, el cual se describe por lo que hemos llamado ruido y que aparece en el modelo lineal como:

$y = \alpha + \beta x + \epsilon$

donde ϵ es el registro del ruido. Suponemos que para un único dato ϵ es una v.a. (variable aleatoria) normalmente distribuída. ¿Cuál es su media? Es cero pues de no serlo sería un efecto sistemático y no sería ruido. La varianza si puede ser cualquier cosa y se puede notar σ^2, y su desvación σ, y suponemos que no dependen de x. Por tanto, como los datos se basan en una muestra que tiene varianza de ruido no nula, las predicciones no podrán ser exactas sino que tendrán un margen de error dado por un intervalo de confianza. El cuadro global se ilustra como sigue:

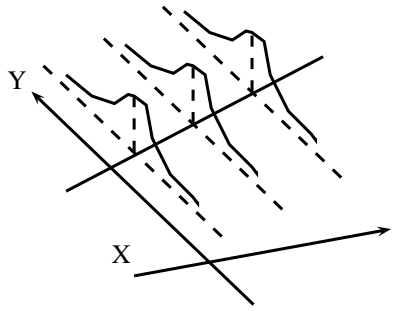

Figura 11. *En esta gráfica tridimensional, el plano XY está en el piso. El ruido hace que los datos observados no sean los predichos por la línea de regresión sino que estén desajustados un poquito. Los desajustes pequeños son más probables que los grandes. Más oficialmente, el desajuste está dado por una distribución normal de desviación* σ*. Esta desviación es válida para un único dato. Cuando hay varios datos, uno debe involucrar el TLC. Esto hace que una predicción de cuánto valdrá y para un cierto x deba ser enunciada no como el valor sobre una línea sino más bien como un intervalo de confianza. Pero atención, la confianza de una predicción será mayor cerca de la media tanto de y como de x pues uno tiene el soporte de los datos alrededor. En tanto que lejos de las medias, hacia los extremos, uno ni siquiera sabe si el modelo se extrapola linealmente o nó. Por eso, los IC de las predicciones deben ser más amplios hacia los extremos.*

La incertidumbre en las predicciones se acostumbra a dar en términos de lo que se denomina el error estandar. Pero eso sí, todo se edifica sobre el sentido común: si el modelo del ejemplo anterior cuantifica la relación entre la cantidad de vitaminas ingeridas y la fortaleza del sistema inmunológico, entonces hay que tener presente que algunas vitaminas pueden ser tóxicas si se ingieren en gran cantidad. Eso quiere decir que las extrapolaciones carecen de autoridad entre más lejos estén del rango de datos.

24 *El error estandar*

El sentimiento de confiabilidad de una predicción se cuantifica mediante el **error estándar** de la predicción, el cual se nota s_y, el cual se define mediante el siguiente grupo de fórmulas:

$\bar{x} = \sum x/n.$
$\bar{y} = \sum y/n.$
$SS_{xx} = \sum x^2 - n\bar{x}^2$
$SS_{xy} = \sum xy - n\bar{x}\bar{y};$
$SS_{yy} = \sum y^2 - n\bar{y}^2;$
$b = \frac{SS_{xy}}{SS_{xx}};$
$SSE = SS_{yy} - bSS_{xy}$

$s = \sqrt{SSE/(n-2)}$, acá está la desviación del desajuste con el TLC.

$c = \sqrt{1 + \frac{1}{n} + \frac{(x_o - \bar{x})^2}{SS_{xx}}}$, éste es el efecto de posición, estar cerca o lejos de la media.

$s_y = sc$, la incertidumbre total es proporcional a ambos factores, el desajuste innato y el efecto de posición.

Supongamos ahora que nosotros hemos escogido una significancia α. Entonces, el valor predicho de y para un x_o dado se da por el intervalo de confianza:

$$[(a + bx_o) - t_{\alpha/2}s_y, (a + bx_o) + t_{\alpha/2}s_y]$$

donde t se toma con n-2 grados de libertad. Nosotros perdemos 2 grados de libertad porque calculamos a y b del conjunto de datos.

25 ***Ejemplo*** *Calculemos el valor predicho de y para* $x_o = 10$ *y* $\alpha = 0{,}05$ *si hay* $n = 6$ *pares de datos con*

$\sum x$ =*19,* $\sum y$ =*41,* $\sum xy$ =*145,* $\sum x^2$ = *67,* $\sum y^2$ =*317*

Solución: ejecutamos las fórmulas:
$\bar{x} = \sum x/6 = 19/6 = 3{,}16$.
$\bar{y} = \sum y/6 = 41/6 = 6{,}83$.
$SS_{xx} = \sum x^2 - n\bar{x}^2 = 67 - 6(3{,}16^2) = 6{,}7$
$SS_{xy} = \sum xy - n\bar{x}\bar{y} = 145 - 6(3{,}16)(6{,}83) = 15{,}50$;
$SS_{yy} = \sum y^2 - n\bar{y}^2 = 317 - 6(6{,}83)^2 = 37{,}10$;
$b = \frac{SS_{xy}}{SS_{xx}} = 15{,}50/6{,}7 = 2{,}31$;
$a = \bar{y} - b\bar{x} = 6{,}83 - (2{,}31)(3{,}16) = -0{,}47$
$SSE = SS_{yy} - bSS_{xy} = 37{,}10 - 2{,}31(15{,}50) = 1{,}295$
$s = \sqrt{SSE/(n-2)} = \sqrt{1{,}295/4} = 0{,}57$
$c = \sqrt{1 + \frac{1}{n} + \frac{(x_o - \bar{x})^2}{SS_{xx}}}$
$c = \sqrt{1 + \frac{1}{6} + \frac{(10-3{,}16)^2}{6{,}7}} = \sqrt{1 + 0{,}167 + \frac{(6{,}83)^2}{6{,}7}}$
$= \sqrt{1 + 0{,}167 + \frac{46{,}6}{6{,}7}} = = \sqrt{1{,}167 + 6{,}96} = 2{,}85$
$s_y = sc$
$s_y = 0{,}57(2{,}85) = 1{,}62$

En consecuencia, para 4 g.l., $\alpha = 0{,}05$ y dos colas, $t = 2{,}77$, por lo que el intervalo de confianza para la predicción de y para $x_o = 10$ es:

$[(a + bx_o) - t_{\alpha/2}s_y, (a + bx_o) + t_{\alpha/2}s_y]$
(-0.47+2.31(10) - 2.77(1.62), -0.47+2.31(10)+ 2.77(1.62)
(-0.47+23.1 -4.48, -0.47+23.1 +4.48) = (18.15, 27.11)

Observemos que los intervalos son muy grandes, lo cual denota que hacer en regresión predicciones precisas es algo muy costoso en términos del número de datos a reunir. Conviene tener en cuenta que la respuesta definitiva es sensible a las aproximaciones que uno haya utilizado a lo largo de los cálculos. Que sea ese otro motivo para recordar que *Excel*, *Gnumeric* y *R* pueden hacer cálculos con una precisión del orden de 10 cifras decimales.

El intervalo de confianza también es útil para probar una hipótesis nula sobre una predicción. Digamos, nosotros no podemos exigir, con un significancia de 0.05, que para $x_o = 10$ el valor de y pueda ser 35 o 15.

26 ***Ejercicio*** *Invente, invente, invente muchas telenovelas y deles un buen final.*

6. Covarianza y correlación

La regresión asume que hay una relación causa-efecto entre dos variables, a la causa se llama indepediente y al efecto dependiente. Utilizamos regresión en experimentos o observaciones programados en las cuales uno tenga control sobre X, la variable independiente, es decir, Y es aleatoria pues depende del ruido pero X no pues es controlada. Con todo, también puede encontrarse que dos variables ambas aleatorias varíen de manera conjunta, que ambas aumentan, que ambas disminuyan o que mientras la una aumente, la otra disminuya y sin embargo uno no pueda decir que hay una relación causa-efecto. Estos casos se estudian mejor por la covarianza y la correlación.

27 ***Definición*** *Sean X, Y dos variables aleatorias que se registran conjuntamente como valores (x_i, y_i). La covarianza entre X y Y se define como:*

$$COV(X;Y) = \sum \frac{(x_i-\bar{x})(y_i-\bar{y})}{n-1}$$

¿Qué mide la covarianza? Ella mide para cada (x_i, y_i) las desviaciones de cada variable con respecto a su media y si dichas desviaciones tienen ambas el mismo signo, produce un aporte positivo a la covarianza total, pero si tienen signo diferente, produce un aporte negativo. Si hay consistencia entre todos los valores, entonces, una covarianza positiva dice que ambas variables varían con igual signo con respecto a su media, es decir, ambas se desvían hacia arriba de sus medias o ambas hacia abajos. Por tanto, la gráfica de la relación (x_i, y_i) tenderá a aglomerarse sobre una función creciente.

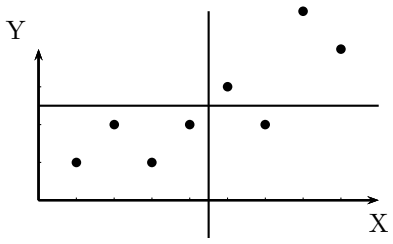

Figura 12. *La media de X está en 4.5, la de Y en 2.5. Se observa la siguiente tendencia: cuando X sube de su media, Y también lo hace. Si X baja de su media, Y hace lo mismo con la suya: la covarianza es positiva y los datos se ajustan a una función creciente.*

Pero si la covarianza tiene signo negativo, éso significa que hay una dominancia de los puntos en que las desviaciones de sus variables con respecto a sus medias tienen signo contrario: la x crece pero la y decrece. En ése caso la gráfica de la relación será una función decreciente:

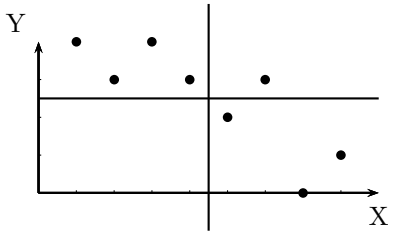

Figura 13. *La media de X está en 4.5, la de Y en 2.5. Se observa la siguiente tendencia: cuando X sube de su media, Y baja de la suya. Si X baja de su media, Y sube de la suya: la covarianza es negativa y los datos se ajustan a una función decreciente.*

Una covarianza cercana a cero indica que no hay una tendencia definida y se tienen abundantes puntos en todos los cuatro cuadrantes generados por las medias como ejes.

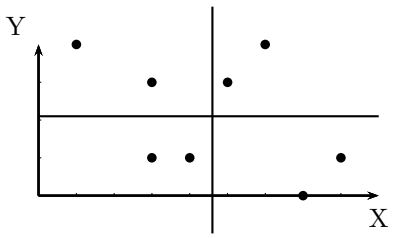

Figura 14. *La media de X está en 4.6, la de Y en 2.1. No se observa una tendencia dominante, cada quien se mueve por su lado. La covarianza es cercana a cero.*

Ejemplos: si uno mide el peso y la altura de los niños, uno encontrará una covarianza positiva. Si uno mide la altura de los niños menores de cinco años y el tiempo que pasan durmiendo, uno encuentra una covariana negativa. Si uno mide la altura de los niños de la misma edad y su destreza para las matemáticas, uno encontrará una covarianza cercana a cero.

28 ***Fórmula operativa.*** *Para los cálculos se pueden usar las siguientes identidades:*

$$COV(X;Y) = \sum \frac{(x_i-\bar{x})(y_i-\bar{y})}{n-1} = \frac{\sum xy - \frac{\sum x \sum y}{n}}{n-1} = \frac{\sum xy - n\bar{x}\bar{y}}{n-1}$$

29 ***Ejemplo*** *Si los datos del ejemplo 21 de la página 14 sobre las lenguas correspondiesen a dos variables aleatorias X, Y, tendríamos $\sum xy = 73$, $\sum x = 36$, $\sum y = 15$, por tanto*

$$cov = \frac{\sum xy - \frac{\sum x \sum y}{n}}{n-1} = \frac{73 - \frac{36\times 15}{8}}{7} = 0{,}786.$$

Nos queda mal interpretar el tiempo X en un proceso histórico de divergencia de lenguas como aleatorio y por tanto no estaríamos en el ámbito de la covarianza. Para estarlo, podemos cambiar la telenovela un poquito: Y representa la distancia entre las lenguas en tanto que X es el número de contracciones introducidas en el lenguaje que suceden aleatoriamente. Por ejemplo: en una conversación conmigo, un muchacho usó la expresión `toes` *para reemplazar la palabra* `entonces`*. Aunque le entendía perfectamente, aparte de él a nadie más le he oído tal contracción. El francés hablado es muy distinto del escrito pues está lleno de contracciones que se comen vocales y terminaciones a partir de una lengua antigua impuesta por el latín. El alemán tiene relativamente pocas vocales lo cual indica que cayeron en contracciones que tuvieron lugar de muy antiguo. En general, los especialistas alegan que un elemento básico de la evolución de toda lengua está compuesto de secuencias de contracciones y expansiones: cuando una contracción merma la comprensión, se produce una expansión aclaratoria. Su producto es sometido a una contracción de otro tipo buscando que la comunicación sea más rápida.*

¿Cómo decidir si una covarianza es grande o pequeña? Para poder hacer eso, necesitamos normalizar la covarianza lo cual produce el coeficiente de correlación cuyos valores van desde -1 hasta 1.

30 ***Definición y teorema.*** *Sean X, Y dos variables aleatorias que se registran conjuntamente como valores (x_i, y_i). La correlación, o coeficiente de correlación, r entre X y Y se define como:*

$$r = \frac{\sum(x_i-\bar{x})(y_i-\bar{y})}{\sqrt{\sum(x_i-\bar{x})^2 \sum(y_i-\bar{y})^2}} = \sqrt{\frac{a\sum y + b\sum xy - n\bar{y}^2}{\sum y^2 - n\bar{y}^2}}, \text{ pero con su signo.}$$

La correlación toma valores entre -1 y 1. Los paquetes pueden asociar nombres muy oficiales al coeficiente de correlación, por ejemplo: `sample Pearson product-moment correlation`*.*

Para juzgar una coeficiente de correlación se relaciona con +1 y -1. Es -1 cuando la covarianza es negativa y los puntos de la gráfica se ajustan perfectamente a una línea decreciente. Es +1 cuando la covarianza es positiva y los puntos de la gráfica se ajustan perfectamente a una línea creciente. Una correlación cercana a cero indica que los puntos están desperdigados sin un desequilibrio consistente.

A r^2 se le llama ***r-cuadrado (r-squared)*** *y también* ***coeficiente de determinación****, toma valores entre 0 y 1. Indica que tan bueno es el ajuste de los datos al modelo y tambien da la proporción de la varianza explicada por el modelo por lo que puede darse en porcentaje. Un valor absoluto cercano a 1 indica un ajuste fuerte y un alto grado de predicibilidad pues hay poco ruido y por tanto la varianza de los datos y la varianza del modelo son casi iguales. En ese caso el p-value es pequeño. Un valor cercano a cero se acompaña con un p-value grande, se debe a que el ruido es alto y por tanto la varianza explicada por el modelo es baja. No hay predecibilidad.*

Cuando la muestra es muy pequeña, hay que tener en cuenta los efectos finitarios o de muestreo los cuales se expresan con el ***coeficiente r-cuadrado-ajustado*** *que es:*

$$r\text{-}cuadrado\text{-}ajustado = 1 - \frac{n-1}{n-2} \times (1 - r^2)$$

Si n es grande, no debemos esperar efectos finitarios. En efecto, el quebrado en la expresión anterior es casi 1 y tenemos:

$$r\text{-}cuadrado\ ajustado = 1 - (1 - r^2) = r^2$$

El r-cuadrado-ajustado puede tomar valores negativos para muestras pequeñas y también cuando r^2 es pequeño. Puede entenderse como una protesta en contra de tener muy pocos datos.

31 ***Ejemplo*** *Para los datos del ejemplo 21 de la página 14 sobre lenguas (1,1)(2,2)(3,1)(4,2)(5,3)(6,2) (7,2) (8,2) tenemos* $\sum xy = 73$, $\sum x = 36$, $\sum y = 15$, $\sum x^2 = 204$, $\sum y^2 = 31$, $\bar{y} = 1{,}875$, $b = 0{,}131$, $a = 1{,}28$ *por tanto*

$$r = \sqrt{\frac{a\sum y + b\sum xy - n\bar{y}^2}{\sum y^2 - n\bar{y}^2}}$$
$$r = \sqrt{\frac{1{,}28 \times 15 + 0{,}131 * 73 - 8 * (1{,}875)^2}{31 - 8 * (1{,}875)^2}}$$

$r = 0{,}25$.

El coeficiente de determinación es $r^2 = 0{,}0625$ *que es extremadamente bajo: aparte de un 7 % por ciento, la variabilidad de los datos se explica por azar, por las variables no controladas.*

32 ***Test de correlación y el IC correspondiente.***

El azar puede crear correlaciones que no existen, es decir, que cuando se repite el muestro, ya no aparecen. Para filtrar el efecto del azar formulamos

$H_o : \rho = 0$ donde ρ es la correlación poblacional. Se asume que se conoce r, la correlación de la muestra.

O también, alguien puede pretender que conoce el valor de la correlación y nuestro objetivo es estudiar su pretensión. Necesitamos probar:

$H_o : \rho = \rho_o$, dado un valor r obtenido de una muestra al azar.

La realidad es que no existe un método exacto para llevar a cabo los test solicitados. Pero existe un método que usa aproximaciones a una normal y es como sigue:
El coeficiente de correlación no está definido antes de -1 ni después de +1. Por lo tanto, dicho coeficiente está demasiado aglomerado hacia cero con respecto a una campana normal que se extienda de largo a largo. Lo que hay que hacer es estirar las puntas del intervalo procurando dejar el intermedio tan igual como se pueda. Para ello usamos un invento de Fisher que para muestras grandes produce una Z, una normal. Fisher nos propone observar que la función $f(x) = 0{,}5ln(\frac{1+x}{1-x})$ hace lo que se necesita:

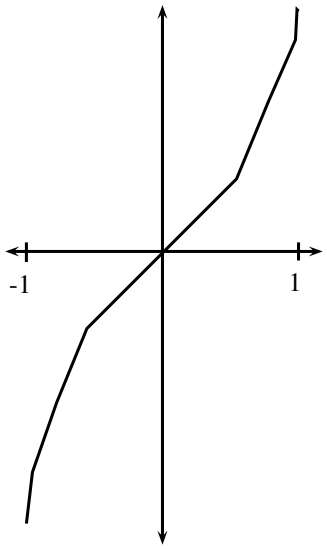

Figura 15. *La función* $f(x) = 0{,}5ln(\frac{1+x}{1-x})$ *toma el intervalo horizontal (-1,1) y lo estira para que llene todo el eje vertical de largo a largo pero modifica muy poco los valores que están cerca de cero.*

Cuando aplicamos esta función al coeficiente de correlación r se denomina la transformada de Fisher de la correlación:

$ZFisher = 0{,}5ln(\frac{1+r}{1-r})$;

Como variable aleatoria, $ZFisher$ tiene media cero y desviación

$\sigma_{ZFisher} = \sqrt{\frac{1}{n-3}}$.

Por tanto, el intervalo de confianza para $ZFisher$ es $ZFisher(r) \pm 1{,}96\sigma_{ZFisher}$. El IC de r se halla a partir del de $ZFisher$ con la transformada inversa a la de Fisher:

$ZFisher = 0{,}5ln(\frac{1+r}{1-r})$
$2ZFisher = ln(\frac{1+r}{1-r})$
$e^{2ZFisher} = \frac{1+r}{1-r}$
$(1-r)e^{2ZFisher} = 1+r$
$e^{2ZFisher} - re^{2ZFisher} = 1+r$
$e^{2ZFisher} - 1 = re^{2ZFisher} + r$
$e^{2ZFisher} - 1 = r(e^{2ZFisher} + 1)$
$r = \frac{e^{2ZFisher}-1}{e^{2ZFisher+1}}$.

El paquete R hace todo automáticamente, pero hay otros paquetes que no y a uno le toca hacer la programación. Por ejemplo, en Java el segmento de código correspondiente podría ser como sigue:

```
//r y n ya han sido inicializadas
double ZFisher = 0.5*(Math.log(1+r)-Math.log(1-r));
double sigmaZFisher = 1/Math.sqrt(n-3);
//Significancia de trabajo = 0.05,
//z-crítico  con dos colas = +/- 1.96
double ZFisherMinus = ZFisher-1.96*sigmaZFisher;
double ZFisherPlus = ZFisher+1.96*sigmaZFisher;
```

```
double rhoMinus =(Math.exp(2*ZFisherMinus)-1)
   /(Math.exp(2*ZFisherMinus)+1);
double rhoPlus=(Math.exp(2*ZFisherPlus)-1)
   /(Math.exp(2*ZFisherPlus)+1);
System.out.println("Intervalo de confianza de r
 =  ("+ rhoMinus + " , " + rhoPlus + ")" );
```

Este segmento de código ya ha sido incluido en el programa en Java que acompaña al Volumen `Basic statistics version 18` de ésta serie.

Como sería conveniente ver un ejemplo a mano, hagamos uno.

33 ***Ejemplo*** *Para los datos del ejemplo 21 de la página 14 (1,1)(2,2)(3,1)(4,2)(5,3)(6,2) (7,2) (8,2) tenemos* $\sum xy = 73$, $\sum x = 36$, $\sum y = 15$, $\sum x^2 = 204$, $\sum y^2 = 31$, $\bar{y} = 1{,}875$, $b = 0{,}131$, $a = 1{,}28$ *por tanto la correlación de la muestra es*

$r = 0{,}25$.

Nos hacemos dos preguntas. Primera: ¿Puede esta correlación se espúrea, creada por azar? Segunda: ¿Puede decirse que los datos no excluyen que la población tenga una correlación alta, digamos 0.9?

Solución: comenzamos haciendo la transformada de Fisher:

$ZFisher = 0{,}5ln(\frac{1+r}{1-r})$

Esta transformada da una distribución aproximadamente normal con media cero y desviación

$\sigma_{ZFisher} = \sqrt{\frac{1}{n-3}} = \sqrt{\frac{1}{5}} = 0{,}447.$

Podemos usarla de forma tradicional para poner a prueba

$H_o : \rho = 0$

Nuestro coeficiente de correlación vale 0.25 y su transformada es:

$ZFisher(0{,}25) = 0{,}5ln(\frac{1+0{,}25}{1-0{,}25}) = 0{,}55$

Si la correlación es cero su transformada es cero.

$ZFisher(0) = 0{,}5ln(\frac{1+0}{1-0}) = 0$

Podemos comparar estos dos valores, 0,55 y 0 teniendo en cuenta que ZFisher se distribuye como una Z, por lo que al normalziar por la desviación estamos en una z, la normal estándar:

$z = \frac{0{,}55-0}{0{,}447} = 1{,}23$

este valor es normal para todas las significancias de interés. Por lo tanto, aceptamos la H_o y nos sentimos apoyados para creer la idea de que la correlación es espúrea. Pero ésto no excluye que la correlación poblacional sea alta en verdad, digamos 0.9, y que tenemos un valor muestral muy bajo por puro azar. Este tipo de peleas se dirimen mejor con intervalos de confianza. Primero hallamos el el intervalo de confianza para $ZFisher$ que es $ZFisher(r) \pm 1{,}96\sigma_{ZFisher}$:

$ZFisherMinus = ZFisher - 1{,}96\sigma_{ZFisher} = 0{,}55 - 1{,}96(0{,}447) = -0{,}326$
$ZFisherPlus = ZFisher + 1{,}96\sigma_{ZFisher} = 1{,}4$

A estos valores tenemos que hallarles la transformada inversa:

$r = \frac{e^{2ZFisher}-1}{e^{2ZFisher+1}}$.

Tenemos que los valores extremos del coeficiente de correlación poblacional son:

$\rho_{Minus} = \frac{e^{2ZFisherMinus}-1}{e^{2ZFisherMinus+1}} = \frac{e^{2(-0,326)}-1}{e^{2(-0,326)+1}} = -0{,}32$
$\rho_{Plus} = \frac{e^{2(1,4)}-1}{e^{2(1,4)+1}} = 0{,}9$

Por tanto, el IC del 95 % de confianza del coeficiente de correlación es (-0.32, 0.9), el cual no está en contra de creer que la correlación poblacional podría ser tan alta como 0.9. Tenemos demasiada libertad de creencias, es decir, demasiada incertidumbre. La forma adecuada de salir de tal verguenza es reunir más datos, pues así disminuiríamos la desviación asociada a ZFisher y por tanto los intervalos serían más pequeños.

7. Resumen de fórmulas

Regresión (causa-efecto) y correlación (no necesariamente causa y efecto) con n pares de datos.

I. Regresión. Causa $=x$, efecto $= y$. Modelo: $y = \alpha + \beta x + \epsilon$; ϵ es el ruido.

1. Hacer dibujo y tabla con x, y, xy, x^2, y^2. Calcular $\bar{x}$ y $\bar{y}$; $SS_{xx} = \sum x^2 - n\bar{x}^2$; $SS_{xy} = \sum xy - n\bar{x}\bar{y}$; $SS_{yy} = \sum y^2 - n\bar{y}^2$; $SSE = SS_{yy} - bSS_{xy}$

2. Línea de regresión (min cuadrados): $y = a + bx$; $b = \frac{SS_{xy}}{SS_{xx}}$; $a = \bar{y} - b\bar{x}$.

3. Test t para $\beta = 0$. Ponga $\beta_o = 0$ en 4. Se toman $n - 2$ g.l.

4. Test t para $\beta = \beta_o$: $s = \sqrt{SSE/(n-2)}$; $s_b = s/\sqrt{SS_{xx}}$; $t = \frac{b-\beta_o}{s_b}$

5. Intervalo de confianza para β: $(b - ts_\beta, b + ts_\beta)$ con $n - 2$ g.l.

6. test F (anova) para $\beta = 0$: $TotalSS = \sum y_i^2 - \frac{(\sum y_i)^2}{n}$; $RegreSS = \frac{(S_{xy})^2}{S_{xx}}$; $RuidoSS = totalSS - regreSS$; $Fisher = \frac{\frac{RegreSS}{1}}{\frac{RuidoSS}{n-2}}$ con 1 y $n - 2$ g.l.

7. Prediccion de y dado x_o: $\widehat{y} = a + bx_o$.

8. IC para la predicción de y dado x_o: $\widehat{y} \pm t_{\alpha/2} s_y$ con $s_y = s\sqrt{1 + \frac{1}{n} + \frac{(x_o - \bar{x})^2}{SS_{xx}}}$

9. IC para la predicción del valor medio de y dado x_o: $\widehat{y} \pm t_{\alpha/2} s\sqrt{\frac{1}{n} + \frac{(x_o - \bar{x})^2}{SS_{xx}}}$

10. Dado x_o, $H_o : y = y_o$. Se usa $t = \frac{a + bx_o - y_o}{s_y}$ con n-2 g.l con s_y de 8.

11. IC para α dado a: se pone $x_o = 0$ en 8.

12. Dado a, test $H_o : \alpha = \alpha_o$: $t = \frac{a - \alpha_o}{s_y}$ con $n - 2$ g.l. con s_y de 8.

13. Para comparar dos pendientes: $ruido = \sum y^2 - \frac{(\sum xy)^2}{\sum x^2}$

$s_p^2 = \frac{ruido_1 + ruido_2}{n_1 + n_2 - 4}$; $s_{(b_1 - b_2)} = \sqrt{\frac{s_p^2}{(\sum x^2)_1} + \frac{s_p^2}{(\sum x^2)_2}}$

$t = \frac{b_1 - b_2}{s_{(b_1 - b_2)}}$ con $n_1 + n_2 - 4$ g.l.

14.Si se sabe que $b_1 = b_2$, todos los datos se reunen y dan una sola $b = b_c$:

$b_c = \frac{(\sum x^2)_1 b_1 + (\sum x^2)_2 b_2}{(\sum x^2)_1 + (\sum x^2)_2}$

15. Para comparar 2 predicciones de 2 modelos se usa el s_p^2 de 13:

$y_1 = a_1 + b_1 x_0$, $y_2 = a_2 + b_2 x_0$,

$s_{(y_1 - y_2)} = \sqrt{s_p^2 [\frac{1}{n_1} + \frac{1}{n_2} + \frac{(x_o - \bar{X}_1)^2}{(\sum x^2)_1} + \frac{(x_o - \bar{X}_2)^2}{(\sum x^2)_2}]}$

$t = \frac{y_1 - y_2}{s_{(y_1 - y_2)}}$; g.l$= (n_1 - 2) + (n_2 - 2) = n_1 + n_2 - 4$.

II. Correlación. Se mide con r, coeficiente de correlación, y con r^2, coeficiente de determinación.

14. $r = \frac{\sum (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum (x_i - \bar{x})^2 \sum (y_i - \bar{y})^2}} = \sqrt{\frac{a \sum y + b \sum xy - n\bar{y}^2}{\sum y^2 - n\bar{y}^2}}$, pero con su signo.

16. Dado un r, para probar $H_o : \rho =$ correlación poblacional $= \rho_o$:

$\omega = \omega(r) = 0{,}5 ln(\frac{1+r}{1-r})$; $z = \frac{\omega(r) - \omega(\rho_o)}{\sigma_\rho}$ donde $\sigma_\rho = \sqrt{\frac{1}{n-3}}$

18. El IC para $\omega(\rho)$ es $\omega(r) \pm 1{,}96\sigma_\rho$. El IC de r se halla a partir del de ω con $r = \frac{e^{2\omega} - 1}{e^{2\omega + 1}}$.

34 *Verificación de supuestos*

La línea de regresión de mínimos cuadrados no tiene ningún prerequisito. Se puede calcular para cualquier conjunto de puntos. Pero cuando se trata de hacer inferencias de la muestra a la población, cuando se corre algún test, el desarrollo matemático siempre descansa sobre ciertas condiciones o supuestos. En particular, los teoremas matemáticos para la regresión lineal univariada que hemos visto tiene los siguientes supuestos:

1. Existe una clara división entre la variable que se controla o independiente x, y las demás variables que pueden influir sobre la variable respuesta o dependiente.

2. No hay error en el proceso de control.

3. Cada variable independiente tiene que aportar información no redundante. Es decir, la correlación entre pares de variables no puede ser muy cercana a 1 o a -1.

4. De no ser por las variables no controladas, el sistema siempre respondería lo mismo ante los mismos valores del estímulo x.

5. Toda fuente de incertidumbre se debe a las variables no controladas, al ruido, cuyo efecto se ve como una v.a. normalmente distribuida con media cero y varianza σ^2 que se superpone al valor promedio de la variable respuesta. Se asume que dicha varianza no depende del valor de x ni tampoco de ningún otro factor externo.

Cuando se llenan los prerequisitos, todo es limpio. Lamentablemente este concepto es difuso. Por ejemplo: ¿Existirá alguien que rechace un estudio de regresión porque encontró una correlación de 0.9 pero que acepte otra donde la correlación fue 0.89? Por eso, es natural preguntarse:

¿Qué tanto peso debe concedérsele a cada uno de los requisitos? En general, cuando se habla de rigor se tiene en mente la naturaleza del ruido, que da origen a los **residuos**: que son las distancias verticales entre cada punto y la línea, que se notan $y_i - \hat{y}_i$. Los residuos deben estar normalmente distribuidos, con media cero (no se chequea porque es cierto por construcción) y varianza que no depende de x. Para poder chequear la homogeniedad de varianzas, uno debe tener por lo menos dos datos de la **variable respuesta o dependiente** por cada dato de la **variable independiente o estímulo o experimental**.

Lo que se desea es que los residuos no puedan refutar la idea de que el ruido sea realmente aleatorio, que se distribuya normalmente, con media cero, y que su varianza sea independiente de x. Y como puede pasar que el rigor exiga abandonar el modelo, entonces uno quisiera saber qué remedio existe que esté por ahí a la mano.

Lo que más problemas da es el conjunto de outliers que corresponden a puntos que quedan demasiado lejos de la línea de regresión. Esos outliers hacen que las colas de la distribución del ruido o azar queden más gordas de lo que debieran. La sugerencia es quitarlos, primero los más alejados e ir cogiendo los más cercanos hasta que ya se pueda aceptar la normalidad. Los outliers más peligrosos son los que están al mismo tiempo alejados de la línea y en un extremo de ella pues un dato así tira el resto de la distribución hacia su lado. Este proceso de maquillaje se justifica alegando, por ejemplo, que hay variables no controldas que inciden demasiado sobre la variable respuesta que se mide. Por supuesto, el objetivo siguiente debería ser tratar de dilucidar dichas variables.

Una anotación con respecto a la aleatoriedad de los residuos. Asegurar que los residuos deban ser aleatorios implica que no exista absolutamente ninguna autocorrelación o correlación entre ellos. Existe un test muy sencillo para decidir que no exista una autocrrelación de primer orden (correlación entre cada dato y su vecino de abajo). Si suponemos que los residuos están numerados de 1 a n, medimos el coeficiente de autocorelación con retrazo 1 de James Durbin y Geoffrey Watson:

$$Durbin - Watson = \sum \frac{(residuo_i - residuo_{i-1})^2}{residuo_i^2}$$

Cuando los residuos son aleatorios, este coeficiente toma el valor 2. Pero como las muestras de v.a. pueden parecer no aleatorias, no se pone problema si este coeficiente está entre 1.5 y 2.5. Si este test dice que los residuos no son aleatorios, se puede abandonar la idea del modelo lineal. Pero si el test no detecta autocorrelación de primer orden queda la duda que existan correlaciones de segundo o tercer o cualquier otro orden, por ejemplo, que sean periódicas. Todas estas sofisticaciones usualmente ni se mencionan, debido a que se cree que el estudio de regresión es robusto y resiste una buena dosis de machete, de violación de los supuestos. Y con correlación se es aún más tolerante pero uno debe estar preparado para dar respuesta a preguntas sobre los residuos cuando se trata de hacer un test, una extrapolación a la población, para luego tomar decisiones.

8. Regresión doble

Hemos estudiado un modelo de causa-efecto en el cual la causa puede modelarse por una variable unidimensional. La generalización inmediata es pasar de líneas de regresión a planos de regresión. Una vez hecha esta generalización, el paso a cualquier número de dimensiones es más de lo mismo.

Veámos como se ejecutan los cálculos. El método es muy apropiado para Excel, Gnumeric y OpenOffice.

Nuestro modelo es:

$y = output = a + b_1 x_1 + b_2 x_2 + \epsilon$

donde los x_i son las dos variables cuantitativas que inciden sobre el resultado. El ruido se describe por ϵ, que se aume como v.a. normalmente distribuida con media cero y desviación σ, la misma en cualquier parte del input.

Tenemos 3 incógnitas a, b_1, b_2, por lo que necestamos 3 ecuaciones independientes. Supongamos que los datos vienen en forma de tabla con n renglones.

Se comienza con una tabla con tres columas de la forma x_1, x_2, y. Dicha tabla se extiende con las siguientes columnas: $x_1 y, x_2 y, x_1^2, x_2^2,\ y^2$ y se calculan las sumas sobre cada columna. Veámos como usamos esta tabla para calcular el modelo

$y = a + b_1 x_1 + b_2 x_2$

Este modelo tiene 3 incógnitas, así que necesitamos 3 ecuaciones independientes, las cuales se formulan así: sumando el modelo sobre todos los valores (x_1, x_2, y) obtenemos

$\sum y = \sum a + \sum b_1 x_1 + \sum b_2 x_2$

esta es nuestra primera ecuación:

Si multiplicamos la ecuación del modelo por x_1 a cada lado y después sumamos, obtenemos

$\sum x_1 y = \sum a x_1 + \sum b_1 x_1 x_1 + \sum b_2 x_1 x_2$

Si hacemos lo mismo con x_2:

$\sum x_2 y = \sum a x_2 + \sum b_1 x_1 x_2 + \sum b_2 x_2 x_2$

Ahora tenemos 3 ecuaciones de las cuales uno puede despejar las 3 incógnitas:

$\sum y = na + b_1 \sum x_1 + b_2 \sum x_2$
$\sum x_1 y = a \sum x_1 + b_1 \sum x_1^2 + b_2 \sum x_1 x_2$
$\sum x_2 y = a \sum x_2 + b_1 \sum x_1 x_2 + b_2 \sum x_2^2$

Notemos que cada coeficiente puede leerse directamente de las sumas de la tabla extendida. De estas ecuaciones, uno resuelve las incógnitas e inmediatamente uno puede hacer predicciones:

Valor predicho de y para valores de x_1 *y* $x_2 = \widehat{y} = a + b_1 x1 + b_2 x_2$

Nuestras predicciones tienen un error debido a σ. El estadígrafo que nos ayuda a cuantificar dicha incertidumbre es s_y, el error estándard de la estimación o predicción, en éste caso con 2 variables independientes siendo n es el número de puntos de la forma (x_1, x_2, y):

$s_y = \sqrt{\frac{\sum (y_i - \widehat{y_i})^2}{n-2-1}}$

Esto nos sirve para calcular el intervalo de confianza de la predicción para un par de x_1 y x_2:

$((a + b_1 x_1 + b_2 x_2) - t_{\alpha/2} s_y, (a + b_1 x_1 + b_2 x_2) + t_{\alpha/2} s_y)$, con $n - k - 1$ g.l.

Decimos que un modelo lineal se justifica cuando podemos alegar que el modelo explica una proporción adecuada de la variabilidad de la variable respuesta. Equivalentemente, podemos estudiar la hipótesis nula:

$H_o : b_1 = b_2 = ... = b_n = 0$
H_a : al menos uno de los $b_i \neq 0$.

Para estudiar estas hipótesis, el análisis de varianza nos ilumina: la variación total se divide en dos partes: una explicada por el modelo lineal y la otra, la variabilidad residual, identificada con el ruido. El quebrado o razón entre estas dos variaciones, apropiadamente normalizadas por sus grados de libertad siguen una distribución F.

La variación total es

$TSS = \sum(y_i - \bar{y})^2 = \sum(y_i - \widehat{y}_i)^2 + \sum(\widehat{y}_i - \bar{y})^2$

Definiendo $RSS = \sum(\widehat{y}_i - \bar{y})^2 =$ (suma de cuadrados explicada por el modelo de regresión (sum of squares explained by the regression model) y

$NSS = \sum(y_i - \widehat{y}_i)^2$ suma de cuadrados atribuible al ruido (sum of squares defined as noise), obtenemos:

$TSS = RSS + NSS.$

Ahora nosotros podemos calcular el estadígrafo de Fisher:

$F = \frac{RSS/2}{NSS/(n-2-1)}$

el cual sigue una distribución F con 2 g.l. en el numerador y $n - 2 - 1$ g.l. en el denominador.

Como hay tanta aritmética por hacer, uno prefiere usar paquetes y por eso es buena idea ver cómo se manejan. Nuestra elección es R, un paquete serio, gigantesco, extendible a las necesidades particulares y gratis.

9. Trabajando con el paquete R

Excel, OpenOffice y Gnumeric todos son muy apropiados para programar hojas y macros que nos den los análisis estadísticos requeridos. Veamos cómo se procede con el paquete R.

Recordemos que la **significancia de un evento**, **p-value** o **valor-p** da la probabilidad de que un evento más extremo que el observado suceda por azar. Lo que hay que tener presente (ver pág 290 del libro adjunto) es que para uno rechazar la Ho se debe tener un p-value inferior a la significancia de trabajo, por defecto 0.05. Si el p-value es mayor, se acepta la Ho. Un p-value grande (mayor que 0.05) dice que hay una multitud de eventos más extremos que el observado, que la diferencia entre lo que se ve y lo que se cree puede explicarse por azar.

En general los paquetes no reportan el valor crítico para una significancia dada, sino `la significancia`, `p-value`, o `valor-p` del estadígrafo de contraste que en regresión básica puede ser una t o una F. A menos que el paquete lo explicite, el p-value es de cola superior. Cuando usa un valor absoluto, se trata de dos colas. Se usa la notación científica, por ejemplo $2{,}46e-5$ significa 0,0000246.

A continuación vienen ejemplos y ejercicios. La pregunta más básica de todas es si hay o no regresión lineal con pendiente diferente de cero. Muy comedidamente solicitamos en cada caso hacer la gráfica o diagrama de dispersión o scatterplot de los datos.

35 ***Ejemplo*** *Primero veamos qué produce el paquete R para los datos del ejercicio 4a y 5a:*

Para el punto 4a) con datos (1,2), (2,1), (1,3), (3,3), (4,3), (4,2) corrimos el programa siguiente:

```
#Programa en R
#Limpia la memoria
rm(list = ls())
#Datos por pares (x,y)
x <- c(1,2,1,3,4,4)
y <- c(2,1,3,3,3,2)
#El símbolo ~ indica que se estudia una relación entre
#la variable dependiente a la izquierda y las independientes
#a la derecha.
#La regresión se estudia por medio de
#los modelos lineales (linear models).
#mod es el nombre del modelo a estudiar.
mod<-lm(y ~ x)
summary(mod)
```

El output del programa fue:

```
Call:
lm(formula = y ~ x)

Residuals:
      1       2       3       4       5       6
-0.1754 -1.2807  0.8246  0.6140  0.5088 -0.4912

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)   2.0702     0.8157   2.538   0.0641 .
x             0.1053     0.2915   0.361   0.7362
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.8983 on 4 degrees of freedom
Multiple R-squared:  0.03158,Adjusted R-squared:  -0.2105
F-statistic: 0.1304 on 1 and 4 DF,  p-value: 0.7362
```

Vemos que en la tabla de `Coefficients` el valor de a y de b. Éstos aparecen en la primera columna numérica y su respuesta coincide con la nuestra: el intersecto es 2.07 y la pendiente es 0.10. En el último renglón vemos el valor de $F_{obs} = 0{,}13$. No hay F crítico sino la significancia o `p-value` de la F_{obs} que es 0.7362. Éste valor nos dice que hay multitudes de valores más extremos que el observado y que por lo tanto lo que observamos se puede explicar por azar: la pendiente resultó diferente de cero por puro azar, se acepta la H_o que dice que la pendiente es cero, que la variable de entrada no influye sobre la salida.

Lo dicho implica que el ruido es enorme. La manera oficial de decirlo es que la variación explicada por la línea hallada debe ser mínima. Eso se lee en el coeficiente de determinación que es el cuadrado de la correlación. Obtuvimos que $r = 0{,}177$ por lo que su cuadrado es 0.0313 que coincide con el valor dado por R: `Multiple R-squared: 0.03158`. Por tanto, la variación explicada por el modelo o línea de regresión es del 3 %. Es lo mismo que decir que el ruido causa el 97 % de la variación total.

Ocurre también que uno puede estudiar la H_o respecto a la pendiente por medio de una t (caso 3 del resumen de fórmulas). Dicho estudio está reportado en la tabla, renglón 2: la t_{obs} vale 0.361 cuyo $p - value$ de dos colas es 0.7362, altísimo (el mismo que el de F): el ruido domina.

El reporte de residuos nos dice cuáles datos son los que se tiran el modelo. En orden decreciente son: el 2, el 3, el 4 y el 5. El dato 2 es el más peligroso. Podría quitarse para estudiar de nuevo la H_o.

También se ve reportado el estudio del intersecto bajo la $H_o : a = 0$. El $p - value$ casi raya con el 5 % pero no alcanza, se acepta la H_o.

¿Qué dicen los errores estándard? Nos dan la desviación de la campana del ruido sacada del TLC. Si dicha desviación es grande o chica hay que leerlo en el $p-value$ de la t, cosa que ya hicimos. Existe un test apropiado para determinar si dicha campana es en realidad una normal, lo cual es un supuesto importante para poder creer el análisis que hemos hecho.

36 ***Ejercicio*** *Analizar el output del paquete R para los datos del ejercicio 4c y 5c: (1,2), (1,3), (2, 5), (3,5), (4,9).*

El programa fue:

```
#Programa en R
#Limpia la memoria
rm(list = ls())
#Datos por pares (x,y)
x <-c(1,1,2, 3,4)
y <-c(2,3,5,5, 9)
#El símbolo ~ indica que se estudia una relación entre
#la variable dependiente a la izquierda y las independientes
#a la derecha.
#La regresión se estudia por medio de
#los modelos lineales (linear models).
#mod es el nombre del modelo a estudiar.
mod<-lm(y ~ x)
#Liste los coeficientes de regresión
mod$coefficients
summary(mod)
```

Resultados:

```
Call:
lm(formula = y ~ x)

Residuals:
      1       2       3       4       5
-0.4706  0.5294  0.5882 -1.3529  0.7059

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)   0.5294     0.9825   0.539   0.6274
x             1.9412     0.3946   4.919   0.0161 *
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 1.029 on 3 degrees of freedom
Multiple R-squared:  0.8897,Adjusted R-squared:  0.8529
F-statistic:  24.2 on 1 and 3 DF,  p-value: 0.01609
```

Respuesta:

Vemos que en la tabla el valor de a y de b aparecen en la primera columna numérica y que su respuesta coincide con la nuestra: el intersecto es 0.5 y la pendiente es 1.9. En el último renglón vemos el valor de $F_{obs} = 24{,}2$. No hay el F crítico sino la significancia o `p-value` de la F_{obs} que es 0.016. Éste valor nos dice que hay menos del 2 % de valores más extremos que el observado y que por lo tanto lo que observamos no se puede explicar por azar: la pendiente resultó diferente de cero no por azar sino porque la variable de entrada x sí afecta a la de salida y. Se rechaza la H_o que dice que la pendiente es cero, y aconsejamos creer que la pendiente es 1.9.

Lo dicho implica que el ruido es pequeño. La manera oficial de decirlo es que la variación explicada por la línea hallada debe ser casi total. Eso se lee en el coeficiente de determinación que es el cuadrado

de la correlación. Obtuvimos que $r = 0{,}94$ por lo que su cuadrado es 0.8836 que coincide con el valor dado por R: `Multiple R-squared: 0.8897`. Por tanto, la variación explicada por el modelo o línea de regresión es del 88 %. Es lo mismo que decir que el ruido causa el 12 % de la variación total. Vemos que los datos presentan un buen ajuste al modelo, o sea que éste tiene valor predictivo.

Ocurre también que uno puede estudiar la H_o respecto a la pendiente por medio de una t (caso 3 del resumen de fórmulas). Dicho estudio está reportado en la tabla, renglón 2: la t_{obs} vale 4.9 cuyo $p-value$ de dos colas es 0.01. Hay una estrellita al frente que nos advierte que tenemos un valor significativo en el nivel del 5 % (el mismo que el de F): el ruido no domina. El modelo es robusto. Como el $p-value$ de dos colas es 0,01 el $p-value$ de una cola es la mitad, pues el área se divide entre 2: nos da 0.005 que dice que con una cola los valores más extremos que el nuestro son el 5 por mil. Maravilloso: tenemos un modelo de regresión altamente predictivo.

El reporte de residuos nos dice cuáles datos son los que hicieron que el $p-value$ no fuese menor: el cuarto dato y después el 5, el 3 y el 2.

También se ve reportado el estudio del intersecto bajo la $H_o : a = 0$. El $p-value$ es tremendamente grande, hay multitud de valores más grandes que el observado. Se acepta la H_o, es permitido creer que el intersecto poblacional es cero y si el muestral no lo fue, es una simple consecuencia del azar.

¿Qué dicen los errores estándard? Nos dan la desviación de la campana del ruido sacada del TLC. Una campana para a y otra para b. Si dicha desviación es grande o chica hay que leerlo en el $p-value$ de la t, cosa que ya hicimos. Existe un test apropiado para determinar si dicha campana es en realidad una normal, lo cual es un supuesto importante para poder creer el análisis que hemos hecho.

37 ***Ejercicio*** *Explique de qué manera un caso fallido se volvió una maravilla. Los datos se escribieron directamente en el programa y vienen de las ventas en millares de un nuevo celular a través del tiempo en meses:*

Explique todo lo del caso fallido siguiente:

```
#Programa en R
> #Limpia la memoria
+ rm(list = ls())
#Datos por pares (x,y)
+ x <- c(1,2,3,4,5,6,7)
+ y <- c(2,5,8,9,8,5,2)
+ modeloLineal <- lm(y ~ x)
+ summary(modeloLineal)

Call:
lm(formula = y ~ x)

Residuals:
      1       2       3       4       5       6       7
-3.5714 -0.5714  2.4286  3.4286  2.4286 -0.5714 -3.5714

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) 5.571e+00  2.665e+00   2.091   0.0908 .
x           4.511e-16  5.959e-01   0.000   1.0000
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 3.153 on 5 degrees of freedom
Multiple R-squared:  1.111e-31,Adjusted R-squared:   -0.2
F-statistic: 5.554e-31 on 1 and 5 DF,  p-value: 1
```

Como los datos no se ajustaron a una línea, ¿se ajustarán a una parábola de la forma $y = a+bc+cx^2$? Demuestre que eso es un éxito y explique por qué.

```
#Programa en R
> #Limpia la memoria
+ rm(list = ls())
#Datos por pares (x,y)
+ x <- c(1,2,3,4,5,6,7)
+ y <- c(2,5,8,9,8,5,2)
+ modeloparabólico <- lm(y ~ poly(x, 2, raw = TRUE))
+ summary(modeloparabólico)

Call:
lm(formula = y ~ poly(x, 2, raw = TRUE))

Residuals:
      1       2       3       4       5       6       7
 0.2381 -0.5714  0.1429  0.3810  0.1429 -0.5714  0.2381

Coefficients:
                      Estimate Std. Error t value Pr(>|t|)
(Intercept)           -3.57143    0.76042  -4.697 0.009331 **
poly(x, 2, raw = TRUE)1  6.09524    0.43579  13.987 0.000152 ***
poly(x, 2, raw = TRUE)2 -0.76190    0.05324 -14.311 0.000139 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.488 on 4 degrees of freedom
Multiple R-squared:  0.9808,Adjusted R-squared:  0.9713
F-statistic: 102.4 on 2 and 4 DF,  p-value: 0.000367
```

Respuesta: los datos no se ajustan a un modelo lineal. En cambio, el modelo parabólico es como agua para el chocolate: todos los coeficientes son altamente significativos y aconsejamos creer que el modelo se ajusta a

$$y = -3{,}57 + 6{,}09x - 0{,}76x^2.$$

Podemos, además, garantizar un alto poder predictivo. Primero las ventas suben, se satura el mercado y viene una declinación. De acá podemos sacar confiadamente una estimación del personal necesario para el proyecto de ventas y del tiempo del contrato.

10. Coeficiente de correlación

Veamos cómo se leen los resultados del paquete R sobre el coeficiente de correlación.

38 ***Ejemplo*** *Primero veamos qué produce el paquete R para los datos del ejercicio 4a y 5a: (1,2), (1,3), (2, 5), (3,5), (4,9).*

El programa completo en R es el siguiente:

```
#Programa en R
#Limpia la memoria
rm(list = ls())
#Datos por pares (x,y)
x <-c(1,1,2,3,4)
y <-c(2,3,5,5, 9)
#El símbolo ~ indica que se estudia una relación entre
#la variable dependiente a la izquierda y las independientes
```

```
#a la derecha.
#La regresión se estudia por medio de
#los modelos lineales (linear models).
#mod es el nombre del modelo a estudiar.
mod<-lm(y ~ x)
#Liste los coeficientes de regresión
mod$coefficients
summary(mod)
#Halle el coeficiente de correlación
cor(x,y)
#Decida la Ho: no hay correlación lineal
#contra la alterna: hay correlación lineal
#causada por un efecto sistemático.
cor.test(x, y)
#Haga el dibujo de dispersión,
#de las parejas de puntos (x,y)
plot(x, y)
#Añada la línea de regresión de mínimos cuadrados.
abline(mod)
```

Aparte de la gráfica, la cual da una línea que se ajusta casi perfectamente a los datos, el paquete produce el siguiente output:

```
> #Programa en R
+ #Limpia la memoria
+ rm(list = ls())
+ #Datos por pares (x,y)
+ x <-c(1,1,2, 3,4)
+ y <-c(2,3,5,5, 9)
+ #El símbolo ~ indica que se estudia una relación entre
+ #la variable dependiente a la izquierda y las independientes
+ #a la derecha.
+ #La regresión se estudia por medio de
+ #los modelos lineales (linear models).
+ #mod es el nombre del modelo a estudiar.
+ mod<-lm(y ~ x)
+ #Liste los coeficientes de regresión
+ mod$coefficients
(Intercept)           x
  0.5294118   1.9411765
+ summary(mod)

Call:
lm(formula = y ~ x)

Residuals:
      1       2       3       4       5
-0.4706  0.5294  0.5882 -1.3529  0.7059

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)   0.5294     0.9825   0.539   0.6274
x             1.9412     0.3946   4.919   0.0161 *
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

```
Residual standard error: 1.029 on 3 degrees of freedom
Multiple R-squared:  0.8897,Adjusted R-squared:  0.8529
F-statistic:  24.2 on 1 and 3 DF,  p-value: 0.01609

+ #Halle el coeficiente de correlación
+ cor(x,y)
[1] 0.9432422
+ #Decida la Ho: no hay correlación lineal
+ #contra la alterna: hay correlación lineal
+ #causada por un efecto sistemático.
+ cor.test(x, y)

Pearson's product-moment correlation

data:  x and y
t = 4.9193, df = 3, p-value = 0.01609
alternative hypothesis: true correlation is not equal to 0
95 percent confidence interval:
 0.3633641 0.9963528
sample estimates:
      cor
0.9432422

+ #Haga el dibujo de dispersión,
+ #de las parejas de puntos (x,y)
+ plot(x, y)
+ #Añada la línea de regresión de mínimos cuadrados.
+ abline(mod)
```

Vemos que el coeficiente de correlación es positivo y es casi 1. El problema es que tenemos pocos datos y por eso debemos descartar que dicho valor se haya obtenido por el azar del muestreo. Por ello hacemos un test para decidir entre la Ho y la Ha:

Ho: el coeficiente de correlación (poblacional) es cero aunque sus contrapartes muestrales no lo sean. Ha: hay un efecto sistemático que crea una correlación no nula.

Para decidir la disyuntiva tenemos dos opciones. Primera: miramos el intervalo de confianza del 95 %, el cual es $(0.3633641, 0.9963528)$. Como 0 no pertenece a dicho intervalo, rechazamos la Ho: aconsejamos creer que la correlación no es cero. Segunda: miramos el valor-p: es 0.01609, el cual nos dice que el azar es capaz de repetir tan alto valor de correlación con una probabilidad de menos de 2 entre 100. Como esto es menos que 5 entre 100, aceptamos con una significancia del 5 % la hipótesis de que la correlación no es cero. Ambas opciones dan el mismo veredicto.

39 ***Ejemplo*** *Veamos qué produce el paquete R para los datos del ejercicio cuyos datos se ajustan a una parábola y no a una línea: Las ventas por mes de un celular nuevo: (1,2), (2,5), (3,8), (4,9), (5,8), (6,5), (7,2).*

El programa completo en R es el siguiente:

```
#Programa en R
#Limpia la memoria
rm(list = ls())
#Datos por pares (x,y)
x <- c(1,2,3,4,5,6,7)
y <- c(2,5,8,9,8,5,2)
#El símbolo ~ indica que se estudia una relación entre
#la variable dependiente a la izquierda y las independientes
#a la derecha.
```

```
#La regresión se estudia por medio de
#los modelos lineales (linear models).
#mod es el nombre del modelo a estudiar.
mod<-lm(y ~ x)
#Liste los coeficientes de regresión
mod$coefficients
summary(mod)
#Halle el coeficiente de correlación
cor(x,y)
#Decida la Ho: no hay correlación lineal
#contra la alterna: hay correlación lineal
#causada por un efecto sistemático.
cor.test(x, y)
#Haga el dibujo de dispersión,
#de las parejas de puntos (x,y)
plot(x, y)
#Añada la línea de regresión de mínimos cuadrados.
abline(mod)
```

El resultado es:

```
> #Programa en R
+ #Limpia la memoria
+ rm(list = ls())
+ #Datos por pares (x,y)
+ x <- c(1,2,3,4,5,6,7)
+ y <- c(2,5,8,9,8,5,2)
+ #El símbolo ~ indica que se estudia una relación entre
+ #la variable dependiente a la izquierda y las independientes
+ #a la derecha.
+ #La regresión se estudia por medio de
+ #los modelos lineales (linear models).
+ #mod es el nombre del modelo a estudiar.
+ mod<-lm(y ~ x)
+ #Liste los coeficientes de regresión
+ mod$coefficients
 (Intercept)           x
5.571429e+00 4.510967e-16
+ summary(mod)

Call:
lm(formula = y ~ x)

Residuals:
      1       2       3       4       5       6       7
-3.5714 -0.5714  2.4286  3.4286  2.4286 -0.5714 -3.5714

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) 5.571e+00  2.665e+00   2.091   0.0908 .
x           4.511e-16  5.959e-01   0.000   1.0000
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 3.153 on 5 degrees of freedom
Multiple R-squared:  1.111e-31,Adjusted R-squared:   -0.2
```

```
F-statistic: 5.554e-31 on 1 and 5 DF,  p-value: 1

+ #Halle el coeficiente de correlación
+ cor(x,y)
[1] 0
+ #Decida la Ho: no hay correlación lineal
+ #contra la alterna: hay correlación lineal
+ #causada por un efecto sistemático.
+ cor.test(x, y)

Pearson's product-moment correlation

data:  x and y
t = 0, df = 5, p-value = 1
alternative hypothesis: true correlation is not equal to 0
95 percent confidence interval:
 -0.7530581  0.7530581
sample estimates:
cor
  0

+ #Haga el dibujo de dispersión,
+ #de las parejas de puntos (x,y)
+ plot(x, y)
+ #Añada la línea de regresión de mínimos cuadrados.
> abline(mod)
```

El paquete R dibuja la línea de ajuste de mínimos cuadrados de forma horizontal, con pendiente cero. La correlación da cero con p-value 1. Lo cual quiere decir que, con raras excepciones, toda muestra al azar tendrá una correlación no cero. Por otro lado, el IC del 95 % es

(-0.7530581 , 0.7530581)

el cual contiene al cero. Ambos métodos aconsejan lo mismo: es normal creer que la correlación es cero.

40 ***Ejercicio*** *Analizar el output del programa siguiente sobre unos datos que relacionan el peso (en decenas de kilogramos) de señoritas tomadas al azar y su número de admiradores declarados.*

```
> #Programa en R
+ #Limpia la memoria
+ rm(list = ls())
+ #Datos por pares (x,y)
+ x <- c(1,2,3,3,6,5,7)
+ y <- c(9,7,8,5,6,4,5)
+ #El símbolo ~ indica que se estudia una relación entre
+ #la variable dependiente a la izquierda y las independientes
+ #a la derecha.
+ #La regresión se estudia por medio de
+ #los modelos lineales (linear models).
+ #mod es el nombre del modelo a estudiar.
+ mod<-lm(y ~ x)
+ #Liste los coeficientes de regresión
+ mod$coefficients
(Intercept)           x
  8.5198020  -0.5792079
+ summary(mod)
```

```
Call:
lm(formula = y ~ x)

Residuals:
      1       2       3       4       5       6       7
 1.0594 -0.3614  1.2178 -1.7822  0.9554 -1.6238  0.5347

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)   8.5198     1.1330   7.520 0.000658 ***
x            -0.5792     0.2599  -2.228 0.076315 .
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 1.396 on 5 degrees of freedom
Multiple R-squared:  0.4983,Adjusted R-squared:  0.3979
F-statistic: 4.966 on 1 and 5 DF,  p-value: 0.07632

+ #Halle el coeficiente de correlación
+ cor(x,y)
[1] -0.705896
+ #Decida la Ho: no hay correlación lineal
+ #contra la alterna: hay correlación lineal
+ #causada por un efecto sistemático.
+ cor.test(x, y)

Pearson’s product-moment correlation

data:  x and y
t = -2.2284, df = 5, p-value = 0.07632
alternative hypothesis: true correlation is not equal to 0
95 percent confidence interval:
 -0.9525806  0.1006835
sample estimates:
      cor
-0.705896

+ #Haga el dibujo de dispersión,
+ #de las parejas de puntos (x,y)
+ plot(x, y)
+ #Añada la línea de regresión de mínimos cuadrados.
> abline(mod)
>
```

Respuesta: si los datos vienen de un experimento controlado, asumir un modelo lineal predice que el intersecto definitivamente no es cero y que es algo abusivo alegar que la pendiente no es cero. Si los datos son observaciones al azar, faltan datos para alegar que el coficiente de correlación es diferente de cero. Vale la pena registrar más datos pues el IC por poco queda sin el cero. Por tanto, no podemos alegar aún que los chicos las prefieren flacas. Sin embargo, las niñas en general son muy sensibles a estos temas y para ellas la evidencia sí es suficiente. Eso se debe a que ellas trabajan con una significancia del alrededor de 0.10.

41 ***Ejercicio*** *Dos niñas muy bien preparadas en Ciencia Política decidieron lanzar su espacio de 10 minutos en la TV (más 5 de propaganda) y cada una piensa que ella es la mejor pero que ambas son muy buenas. Por eso nos pidieron analizar sus datos que relacionan el tiempo que cada niña aparece*

en pantalla con el rating medido por los twiters recibidos. Los datos se componen de tripletas donde la primera coordenada representa el tiempo de la primera niña, la segunda el tiempo de la segunda niña y la tercera el rating en escala arbitraria. Los datos son:

(8,2,4), (2,8,5), (4,6,3), (3,7,6), (2,8,4), (5,5,3), (7,3,5).

Solución: este problema es muy difícil porque hay competencia sobre un marco de colaboración. Con todo, podemos hacer un primer rastreo con los modelos lineales: si más tiempo de una de las niñas produce proporcionalmente mayor rating, dicha niña es la mejor. El siguiente programa en R pretende formular el modelo de regresión de 2 variables de entrada y una de salida:

```
#Programa en R
#Limpia la memoria
rm(list = ls())
#Datos por pares (x,y)
n1 <- c(8,2,4,3,2,5,7)
n2 <- c(2,8,6,7,8,5,3)
r  <- c(4,5,3,6,4,3,5)
#El símbolo ~ indica que se estudia una relación entre
#la variable dependiente a la izquierda y las independientes
#a la derecha.
#La regresión se estudia por medio de
#los modelos lineales (linear models).
#mod es el nombre del modelo a estudiar.
mod<-lm(r ~ n1 + n2)
#Liste los coeficientes de regresión
mod$coefficients
summary(mod)
```

Este programa produjo el siguiente output:

```
> #Programa en R
+ #Limpia la memoria
+ rm(list = ls())
+ #Datos por pares (x,y)
+ n1 <- c(8,2,4,3,2,5,7)
+ n2 <- c(2,8,6,7,8,5,3)
+ r  <- c(4,5,3,6,4,3,5)
+ #El símbolo ~ indica que se estudia una relación entre
+ #la variable dependiente a la izquierda y las independientes
+ #a la derecha.
+ #La regresión se estudia por medio de
+ #los modelos lineales (linear models).
+ #mod es el nombre del modelo a estudiar.
+ #Se estudia la dependencia lineal de la variable r
+ #del conjunto de variables n1 y n2.
+ mod<-lm(r ~ n1 + n2)
+ #Liste los coeficientes de regresión
+ mod$coefficients
(Intercept)          n1          n2
 4.66101695 -0.08474576          NA
> summary(mod)

Call:
lm(formula = r ~ n1 + n2)

Residuals:
       1        2        3        4        5        6        7
```

```
 0.01695  0.50847 -1.32203  1.59322 -0.49153 -1.23729  0.93220

Coefficients: (1 not defined because of singularities)
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  4.66102    1.02050   4.567  0.00602 **
n1          -0.08475    0.20647  -0.410  0.69848
n2                NA         NA      NA       NA
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 1.199 on 5 degrees of freedom
Multiple R-squared:  0.03259,Adjusted R-squared:  -0.1609
F-statistic: 0.1685 on 1 and 5 DF,  p-value: 0.6985
```

El paquete acepta estudiar el efecto de la primera niña pero al enfrentar la segunda no puede y responde NA que significa `Not Available` (no existente). ¿Qué ocurre? Lo que pasa es que tenemos un problema técnico: los datos no son independientes pues el tiempo que no ocupa una de las niñas lo ocupa la otra. Podemos verificar que la causa del problema es la dependencia entre las dos series de datos si modificamos los datos un poquito para que se pierda la dependencia:

```
#Programa en R
#Limpia la memoria
rm(list = ls())
#Datos por pares (x,y)
n1 <- c(8,2,4,3,7,5,7)
n2 <- c(2,4,6,7,7,5,3)
r  <- c(4,5,3,6,4,3,5)
#El símbolo ~ indica que se estudia una relación entre
#la variable dependiente a la izquierda y las independientes
#a la derecha.
#La regresión se estudia por medio de
#los modelos lineales (linear models).
#mod es el nombre del modelo a estudiar.
mod<-lm(r ~ n1 + n2)
#Liste los coeficientes de regresión
mod$coefficients
summary(mod)
#End
```

Como resultado vemos lo siguiente, en donde los datos modificados de la segunda niña también son aceptados por el paquete:

```
> #Programa en R
+ #Limpia la memoria
+ rm(list = ls())
+ #Datos por pares (x,y)
+ n1 <- c(8,2,4,3,7,5,7)
+ n2 <- c(2,4,6,7,7,5,3)
+ r  <- c(4,5,3,6,4,3,5)
+ #El símbolo ~ indica que se estudia una relación entre
+ #la variable dependiente a la izquierda y las independientes
+ #a la derecha.
+ #La regresión se estudia por medio de
+ #los modelos lineales (linear models).
+ #mod es el nombre del modelo a estudiar.
+ mod<-lm(r ~ n1 + n2)
+ #Liste los coeficientes de regresión
```

```
+ mod$coefficients
(Intercept)          n1          n2
 5.42801556 -0.16147860 -0.06420233
+ summary(mod)

Call:
lm(formula = r ~ n1 + n2)

Residuals:
        1         2         3         4         5         6         7
-0.007782  0.151751 -1.396887  1.505837  0.151751 -1.299611  0.894942

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)   5.4280     2.3730   2.287   0.0841 .
n1           -0.1615     0.2563  -0.630   0.5629
n2           -0.0642     0.2978  -0.216   0.8399
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 1.3 on 4 degrees of freedom
Multiple R-squared:  0.09069,Adjusted R-squared:  -0.364
F-statistic: 0.1995 on 2 and 4 DF,  p-value: 0.8268
> #End
```

De lo visto podemos mejorar nuestra programación:

```
#Programa en R
#Limpia la memoria
rm(list = ls())
#Datos por pares (x,y)
n1 <- c(8,2,4,3,2,5,7)
n2 <- c(2,8,6,7,8,5,3)
r  <- c(4,5,3,6,4,3,5)
#El símbolo ~ indica que se estudia una relación entre
#la variable dependiente a la izquierda y las independientes
#a la derecha.
#La regresión se estudia por medio de
#los modelos lineales (linear models).
#mod es el nombre del modelo a estudiar.
mod<-lm(r ~ n1)
#Liste los coeficientes de regresión
mod$coefficients
summary(mod)
#Haga el dibujo de dispersión,
#de las parejas de puntos (n1,r)
plot(n1,r)
#Añada la línea de regresión de mínimos cuadrados.
abline(mod)
#
#Halle el coeficiente de correlación entre n1 y n2
cor(n1,n2)
#Decida la Ho: no hay correlación lineal
#contra la alterna: hay correlación lineal
#causada por un efecto sistemático.
cor.test(n1, n2)
```

```
#Haga el dibujo de dispersión,
#de las parejas de puntos (n1,n2)
plot(n1, n2)
mod2 <-lm(n1 ~ n2)
#Añada la línea de regresión de mínimos cuadrados.
abline(mod2)
#End
```

El paquete produjo el siguiente output:

```
> #Programa en R
+ #Limpia la memoria
+ rm(list = ls())
+ #Datos por pares (x,y)
+ n1 <- c(8,2,4,3,2,5,7)
+ n2 <- c(2,8,6,7,8,5,3)
+ r  <- c(4,5,3,6,4,3,5)
+ #El símbolo ~ indica que se estudia una relación entre
+ #la variable dependiente a la izquierda y las independientes
+ #a la derecha.
+ #La regresión se estudia por medio de
+ #los modelos lineales (linear models).
+ #mod es el nombre del modelo a estudiar.
+ mod<-lm(r ~ n1)
+ #Liste los coeficientes de regresión
+ mod$coefficients
(Intercept)          n1
 4.66101695 -0.08474576
+ summary(mod)

Call:
lm(formula = r ~ n1)

Residuals:
       1        2        3        4        5        6        7
 0.01695  0.50847 -1.32203  1.59322 -0.49153 -1.23729  0.93220

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  4.66102    1.02050   4.567  0.00602 **
n1          -0.08475    0.20647  -0.410  0.69848
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 1.199 on 5 degrees of freedom
Multiple R-squared:  0.03259,Adjusted R-squared:  -0.1609
F-statistic: 0.1685 on 1 and 5 DF,  p-value: 0.6985

+ #Haga el dibujo de dispersión,
+ #de las parejas de puntos (n1,r)
+ plot(n1,r)
+ #Añada la línea de regresión de mínimos cuadrados.
+ abline(mod)
+ #
+ #Halle el coeficiente de correlación entre n1 y n2
+ cor(n1,n2)
```

```
[1] -1
+ #Decida la Ho: no hay correlación lineal
+ #contra la alterna: hay correlación lineal
+ #causada por un efecto sistemático.
+ cor.test(n1, n2)

Pearson's product-moment correlation

data:  n1 and n2
t = -Inf, df = 5, p-value < 2.2e-16
alternative hypothesis: true correlation is not equal to 0
95 percent confidence interval:
 -1 -1
sample estimates:
cor
 -1

+ #Haga el dibujo de dispersión,
+ #de las parejas de puntos (n1,n2)
+ plot(n1, n2)
+ mod2 <-lm(n1 ~ n2)
+ #Añada la línea de regresión de mínimos cuadrados.
+ abline(mod2)
> #End
```

El paquete también hizo dos gráficas. En RKWard de Ubuntu Linux las dos gráficas aparecen en el mismo panel y uno puede cambiar la gráfica presentada por medio de las flechas de navegación del mismo panel. Una gráfica indica que la línea de mínimos cuadrados da una efecto negativo de la primera niña pero el análisis estadístico dice que dicha lectura es apresurada: hay una certeza total de que eso es tan sólo un efecto del azar. Si se repitiera el experimento, no habría porque admirarse si los datos dieran al revés. Por otro lado, la gráfica de la correlación entre las dos variables n1 y n2 es perfecta, tal como se espera: dicha dependencia estadística inhabilita un modelo lineal con 2 variables de entrada.

Podemos ahora dar nuestro veredicto definitivo con seguridad y sencillez: los modelos lineales no dan razón alguna para decir que una niña es mejor que la otra en cuanto a rating. Podemos añadir un consejo: debido a que los datos de la segunda niña son redundantes (o si prefiere, los de la primera), el estudio queda desabrido, decolorido. Podría enmendarse el experimento como sigue: además de las dos niñas en acción, se añade un tercer actor, un invitado, el cual se toma un tiempo con mucha espontaneidad de tal manera que el tiempo tomado es una variable aleatoria. Tenemos que las 3 series son dependientes entre sí, pero nó las dos primeras, las de las niñas, que son las que interesan. Con dicho cambio, podemos acumular datos suficientes para que se pueda dilucidar el papel lineal de cada niña: a favor, neutro o comunitario, o en contra.

El intersecto puede interpretarse como un efecto de la franja, quizá como la población cautiva que ve lo que sea. Es una suerte de efecto cooperativo de todo el sistema. Pero en cuanto a las niñas, ellas compiten sanamente lo cual quiere decir que combinan la cooperación con la competencia para maximizar el producto total (rating que genera entradas por propaganda). Una situación tan complicada tal vez no pueda modelarse con paquetes sino debe simularse y tener en cuenta la teoría de juegos y coaliciones. Para tal fin convendría aprender un lenguaje de programación. A falta de mejores ideas, Java puede servir y también el mismo R.