21

Modelos para más de una variable

Hasta ahora se han visto diferentes modelos estadísticos para el caso de una sola magnitud biológica medida. Pero en los experimentos es frecuente tratar el caso donde hay más de una variable involucrada. En este capítulo se tratará el caso de dos variables con los modelos de *Regresión y Correlación estadística*. El caso de más de dos variables excede los límites del presente trabajo, destinado a los alumnos de las carreras de Farmacia y Bioquímica, porque requiere del manejo de un nivel matemático (espacios vectoriales n-dimensionales) que estos no poseen. Se comienza en este capítulo con el caso más sencillo de relación lineal simple entre dos magnitudes biológicas cualesquiera. Es decir, una relación del tipo Y = X = f (X). A continuación se trata el modelo más general con relaciones del tipo Y = a + b X, la ecuación de una recta o polinomio de primer grado. Modelo conocido con el nombre de *Análisis de Regresión lineal*. Se presentan los métodos cortos de cálculo y el planteo de ensayos de hipótesis respectivos. A continuación de generaliza la regresión para el caso de polinomios de más de un grado y se muestran las transformaciones de variables, convenientes para linealizar los cálculos.

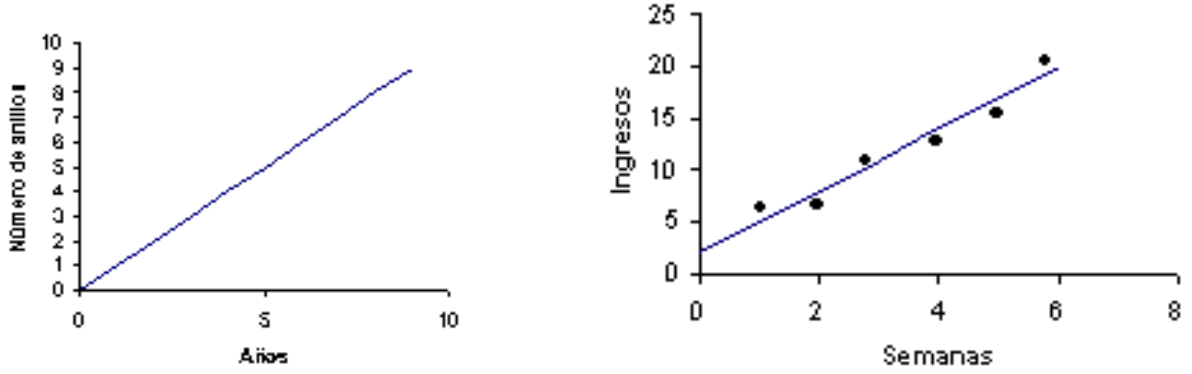
21.1 Introducción

Antes de comenzar, es conveniente aclarar una confusión frecuente en la bibliografía. Hay textos que usan los cálculos de regresión y correlación para los mismos casos por lo similares que son. A veces el lector se confunde y piensa que puede emplear ambos modelos en un mismo problema. Esto no es así, hay una diferencia fundamental entre ellos. El *Análisis de Regresión* se usa cuando el investigador sabe que existe una relación entre las variables porque hay una teoría o investigaciones previas que la han descubierto. Por ejemplo, la relación entre espacio y tiempo ya se sabe que es la velocidad, o como la relación entre voltaje e intensidad de corriente eléctrica. En estos casos, el investigador suele estar interesado en verificar experimentalmente tal relación y el objeto de la regresión es encontrar la curva que mejor ajuste a sus datos experimentales. Cuando no conoce la relación exacta, sino que trata de encontrar una curva de tipo práctica en los llamados modelos empíricos, la idea es tener una manera rápida de relacionar dos magnitudes; como por ejemplo, en la industria cuando se relaciona la velocidad de un motor con el rendimiento del generador eléctrico que mueve. Por su parte, el *Análisis de Correlación* se emplea cuando el investigador sospecha que ambas magnitudes están relacionadas, pero no tiene idea de una ecuación que las combine. Por ejemplo el caso de peso y talla, donde todo lo que se sospecha es que a mayor talla, mayor peso, pero nadie ha descubierto una fórmula que las relacione. Con esto en mente se comienza con:

21.2 Análisis de regresión

La forma más común de concebir las relaciones entre pares de magnitudes es del tipo *causa-efecto*. Lo que trata el análisis estadístico es establecer la forma y la significación de las relaciones funcionales entre las dos variables. La demostración de la relación causa-efecto es tema del procedimiento del método científico, y queda a cargo del investigador. Estadística trata de verificar la función matemática que permite predecir que valores de una variable *Y* corresponden a valores dados de una variable *X*. Se suele escribir como Y = f (X), donde X es la variable independiente.

Figura 21.1: Rectas de regresión.



(a) Número de anillos de un árbol **(b) Ingresos semanales en una Farmacia (miles $)**

El caso más simple de una recta de regresión es del tipo Y = X donde la recta pasa por el origen de coordenadas y su inclinación es de 45º. Este es el caso de la relación entre el número (Y) de anillos de un árbol y su edad en años (X). Ver caso (a) de la figura anterior. El caso más general es cuando la recta no pasa por el origen y su inclinación es cualquiera. La relación matemática es del tipo Y = **a** + **b** X donde **a** es el punto por donde la recta corta al eje Y cuando X = 0 y **b** es la tangente del ángulo de inclinación. Ver caso (b) de la figura anterior donde se muestra la relación de los ingresos de una Farmacia medidos en miles de pesos con el tiempo expresado en semanas, con una ecuación expresada con: Y = 2 + 3 X. Aquí, se supone que se han medido los ingresos reales en una Farmacia y se encontró la recta que mejor ajusta a esa serie de puntos con el Análisis de Regresión.

En todo ejemplo real, las observaciones no coinciden exactamente con la recta de regresión debido a los errores casuales que afectan las mediciones. En Biología se suponen causas de tipo genético y ambientales para explicar la aleatoriedad, además de los errores de medición. Esto significa que para un dado valor de X, el valor de Y que le corresponde no será exactamente: a + b X, sino que esta ecuación usando el valor de X arroja el *valor esperado* de Y denominado Y*. Entonces, para cada valor medido de X se tendrán dos valores: el valor medido en el experimento Y y su valor esperado calculado por la recta de regresión Y*. La diferencia entre ambos (Y-Y*) debe ser lo más chica posible, para tener una buena aproximación. La idea básica del Análisis de Regresión es minimizar matemáticamente el cuadrado de estas diferencias con el método de los *mínimos cuadrados*.

21.3 Diseños experimentales

Los diseños experimentales en regresión son dos: el Modelo I y el II. Ambos se basan en cuatro hipótesis básicas.

1. *La variable independiente se mide sin error*. Esto significa que está bajo el control del investigador y se consideran "fijos" a los valores de X que eligió. Por ejemplo, al manipular las dosis de un cierto medicamento hipo-tensor, se fijan estos valores de X y por lo tanto no se la puede considerar como una variable aleatoria. En cambio, el valor de la presión sanguínea del paciente no puede ser fijada por el investigador; entonces Y varía en forma libre. En forma análoga a la vista en los modelos de Anova, se considera *Modelo I* de regresión, al caso donde los valores de X pueden ser manipulados a voluntad. Cuando esto no es así, entonces se tiene el *Modelo II* de regresión donde ambas variables se consideran aleatorias porque no están bajo el control de investigador. Por ejemplo, se toma una muestra aleatoria de una población y a cada individuo seleccionado se le mide su presión sanguínea y su nivel hormonal. Ambas variables no quedan bajo el control del investigador y deben ser consideradas aleatorias.

2. *El valor esperado de la variable Y* para un dado valor de X, se determina con la relación*:

$$E(Y) = \mu_y = Y^* = \alpha + \beta X$$

Se usan las letras griegas para describir una relación Paramétrica entre las variables.

3. *Para cualquier valor de X, los valores de Y se distribuyen independiente y normalmente.*

$$Y^* = \alpha + \beta X + \varepsilon$$

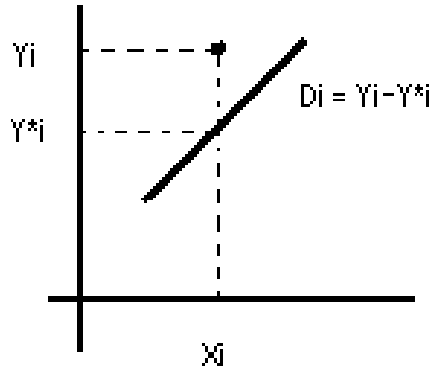
Donde ε es un término de error con una distribución normal de media igual a cero y desvío σ. Esto supone que cada valor X tiene un gran número de valores posibles de Y a hacer la medición, con una distribución normal, cuyo eje de simetría es una vertical (eje z: dentro de un espacio tridimensional imaginario) que pasa por el punto Y* de la recta de regresión, orientada sobre la línea que une el punto X con él Y* correspondiente. Esto para los casos donde hay más de un valor de Y para cada valor de X.

4. *Todas las muestras a lo largo de la línea de regresión son homocedásticas*. Se supone que todas las distribuciones normales mencionadas en el punto anterior tienen la misma varianza.

21.4 Cálculos básicos en regresión

La manera más sencilla de ilustrar estos cálculos es partiendo del concepto básico de la recta de regresión: minimizar el cuadrado de las diferencias entre el valor medido Yi y el valor correspondiente esperado por la recta Y*i. El caso más sencillo es cuando hay un solo valor medido de Yi, para cada valor X_i. En la Figura 21.2 siguiente se muestra el planteo:

Figura 21.2: Diferencias a minimizar.



Cada punto medido tiene un par de coordenadas (X_i ; Yi) Con la recta se estima el valor Y*i. La diferencia entre el valor medido y el estimado es Di = Yi – Y*i. Cuanto mejor sea la estimación, menor será esta diferencia. El método consiste en *minimizar* la suma de sus cuadrados: derivando respecto de las dos incógnitas a y b, igualando a cero y despejando. Queda un sistema de dos ecuaciones con dos incógnitas, que al resolverlo permiten hallar las denominadas *ecuaciones paramétricas de regresión.*

Para minimizar se usan las relaciones (ver Apéndice 1):

$$\frac{\partial}{\partial a}\sum D_i^2 = 0 \qquad\qquad \frac{\partial}{\partial b}\sum D_i^2 = 0$$

Resolviendo estas relaciones se obtienen:

$$\left.\begin{aligned} \sum Yi &= a.N + b\sum X_i \\ \sum X_i Yi &= a.\sum x_i + b\sum X_i^2 \end{aligned}\right\}$$

Ecuaciones normales o paramétricas de regresión.

Ejemplo) Se ha medido la altura de 15 padres y de sus hijos primogénitos en metros. Hallar la recta de regresión. Los datos son:

	X (padres)	Y (hijos)	X^2	X Y	Y^2
1	1,71	1,75	2,9241	2,9925	3,0625
2	1,67	1,76	2,7889	2,9392	3,0976
3	1,62	1,64	2,6244	2,6568	2,6896
4	1,75	1,76	3,0625	3,08	3,0976
5	1,59	1,64	2,5281	2,6076	2,6896
6	1,81	1,77	3,2761	3,2037	3,1329
7	1,69	1,72	2,8561	2,9068	2,9584
8	1,68	1,73	2,8224	2,9064	2,9929
9	1,76	1,75	3,0976	3,08	3,0625
10	1,72	1,72	2,9584	2,9584	2,9584
11	1,79	1,81	3,2041	3,2399	3,2761
12	1,68	1,69	2,8224	2,8392	2,8561
13	1,64	1,66	2,6896	2,7224	2,7556
14	1,68	1,71	2,8224	2,8728	2,9241
15	1,58	1,62	2,4964	2,5596	2,6244
SUMA	25,37	25,73	42,9735	43,5653	44,1783

Reemplazando en las ecuaciones paramétricas resulta

$$\left.\begin{array}{l} 25{,}73 = a\,(15) + b\,(25{,}37) \\ 43{,}5653 = a\,(25{,}37) + b\,(42{,}9735) \end{array}\right\} \quad \text{De donde se calculan a y b.}$$

Sin embargo hay una forma más corta. La recta de regresión debe pasar necesariamente por el centro de gravedad de los datos, es decir por el valor medio de X e Y. Esto es, si se divide la primer ecuación por N en ambos miembros, queda igual a:

$(1/N)\sum Yi = a + (b/N)\sum X_i$ O sea: $\overline{Y} = a + b\,\overline{X}$ O bien: $a = \overline{Y} - b\,\overline{X}$

Se puede calcular los valores medios de la tabla anterior resultando:

$\overline{X} = 25{,}37 / 15 = 1{,}6913$ m $\overline{Y} = 25{,}73 / 15 = 1{,}7153$ m

Reemplazando por los valores hallados se despeja: $a = (1{,}7153 - b\ 1{,}6913)$

Reemplazando el valor de a en la segunda ecuación, queda una sola ecuación con una incógnita:

$43{,}565 = (1{,}72 - b\ .\ 1{,}69)\ 25{,}37 + b\ 42{,}97 = 1{,}72\ (25{,}37) + b\ [(42{,}97) - (1{,}69)(25{,}37)]$

Despejando se calcula: $b = -0{,}7$. Con este valor, se puede calcular a, reemplazando en la fórmula de valores medios:

$\overline{Y} = a + b\,\overline{X}$

$1{,}72 = a + (-0{,}7)\ (1{,}6913)$

Resultando: $a = 2{,}9$

Y la ecuación de la recta es: **$Y = 2{,}9 - 0{,}7\ X$**

Si en lugar de calcular la regresión de Y sobre X (X es la variable independiente), se desea calcular la regresión de X sobre Y, porque Y es la variable independiente, entonces las ecuaciones quedan muy similares, pero con otras incógnitas:

$$\left.\begin{array}{l} \sum X_i = c.N + d\sum Y_i \\ \sum X_i Yi = c.\sum Y_i + d\sum Y_i^2 \end{array}\right\} \quad \textit{Ecuaciones normales para regresión de X sobre Y.}$$

Aplicando estas nuevas relaciones al ejemplo anterior resulta:

$$\left.\begin{array}{l} 25{,}37 = c\,(15) + d\,(25{,}73) \\ 43{,}5653 = c\,(25{,}73) + d\,(44{,}1783) \end{array}\right\} \quad \text{De donde se calculan c y d.}$$

Para simplificar se usa:

$$\overline{X} = c + d\ \overline{Y} \qquad \text{O bien:} \qquad c = \overline{X} - d\ \overline{Y}$$

Reemplazando por los valores hallados se despeja: c = (1, 6913 – d 1,7153)

Reemplazando el valor de a en la segunda ecuación, queda una sola ecuación con una incógnita:

43,5653 = (1,6913 – d 1,7153)(25,73) + d 44,1783

0,04815 = d (0,04363) y despejando resulta: d = 1,1 y así se obtiene c = - 0,2

Y la ecuación de esta otra recta es: **X = - 0,2 + 1,1 Y**

Como se puede apreciar, la regresión de Y sobre X difiere de la regresión de X sobre Y. Solo cuando la coincidencia entre los puntos reales y la recta de regresión sea perfecta, entonces ambas rectas de regresión serán iguales.

Existen otras maneras de simplificar los cálculos, usando un cambio de coordenadas. Esto se hace, colocando el origen en el centro de gravedad de los datos. Sencillamente, a cada valor de X e Y se le resta su media respectiva. Las ecuaciones quedan más cortas. El lector interesado podrá encontrar estas técnicas en la Bibliografía específica.

21.5 Cálculos cortos en regresión

La manera más rápida de efectuar los cálculos, incluso los necesarios para los tests de hipótesis, se ilustra en el ejemplo siguiente:

Ejemplo) Nelson (1964) midió la pérdida en peso (en mg) de 9 tandas de veinticinco coleópteros *Tribolium,* luego de seis días de no ingerir comida a nueve humedades diferentes. Los datos son:

	Hum.	Peso	x =	y =	x^2	y^2	x.y	Y*	D =	D^2	y*=	y^{*2}
	X	Y	X-50,39	Y-6,02					Y-Y*		Y*-6,02	
	0	8,98	-50,389	2,9578	2539,04	8,748449	-149,04	8,704	0,276	0,07616	2,68181	7,1921
	12	8,14	-38,389	2,1178	1473,707	4,484983	-81,299	8,0654	0,0746	0,00557	2,04314	4,1744
	29,5	6,67	-20,889	0,6478	436,3457	0,419616	-13,531	7,134	-0,464	0,21527	1,11175	1,236
	43	6,08	-7,3889	0,0578	54,59568	0,003338	-0,4269	6,4155	-0,3355	0,11254	0,39325	0,1546
	53	5,9	2,6111	-0,122	6,817901	0,014938	-0,3191	5,8833	0,0167	0,00028	-0,13897	0,0193
	62,5	5,83	12,111	-0,192	146,679	0,036949	-2,328	5,3776	0,4524	0,20463	-0,64458	0,4155
	75,5	4,68	25,111	-1,342	630,5679	1,80156	-33,705	4,6858	-0,0058	3,3E-05	-1,33647	1,7861
	85	4,2	34,611	-1,822	1197,929	3,320494	-63,069	4,1801	0,0199	0,00039	-1,84208	3,3933
	93	3,72	42,611	-2,302	1815,707	5,300227	-98,1	3,7544	-0,0344	0,00118	-2,26786	5,1432
Σ	**453,5**	**54,2**	**0**	**0**	**8301,389**	**24,13056**	**-441,82**	**54,2**	**0**	**0,61606**	**0**	**23,514**
	50,3889	**6,022**	**:Medias**					**6,0222**				

b	**-0,053222**
a	**8,704027**

Paso 1) En las primeras dos columnas se vuelcan los datos, se suman y se saca el promedio.

Paso 2) En la tercer y cuarta columna, se colocan los valores de las dos primeras menos sus respectivos promedios y se determinan: $y = f(x)$ dos variables que pasan por el centro de gravedad de los datos, de acuerdo a la transformación efectuada:

$x = X - \overline{X} = X - 50{,}3889$ e $y = Y - \overline{Y} = Y - 6{,}022$

Paso 3) En la quinta y sexta columnas se calculan los cuadrados de los nuevos valores x e y. Mientras que en la séptima se calcula su producto.

Paso 4) Ahora se pueden calcular los coeficientes de la recta de regresión a y b con la ecuación de la recta de regresión de x sobre y, escrita en el nuevo sistema de coordenadas:

$y = b\,x = \left(\frac{\sum xy}{\sum x^2}\right).x$ Por lo tanto: $b = (-441{,}82) / (8301{,}389) = -0{,}0532$

Es decir se calcula la pendiente de la recta b como el cociente entre los totales de la quinta y séptima columna. Conocido este valor, se puede calcular el otro considerando que la recta pasa por el centro de gravedad de los puntos, o sea por sus valores promedio:

$\overline{Y} = a + b\,\overline{X}$ O sea: $a = \overline{Y} - b\,\overline{X} = 6{,}022 - (-0.0532)\,50{,}389 = 8{,}704$

$\overline{Y} = 8{,}7 - 0{,}05\,\overline{X}$

21.6 Ensayos de hipótesis en regresión

La manera de ensayar la hipótesis de que la regresión existe es con un Cuadro de Regresión, similar al visto en el caso del Anova. Para ello se divide a la suma de cuadrados que representa la variabilidad total en dos términos, lo mismo que con los grados de libertad, y se encuentra un estadígrafo F con distribución de Fisher. Para ello se empieza con:

Variación Total: De una variable cualquiera Y, se calcula como la suma de cuadrados de las diferencias entre cada valor medido y su promedio (es la suma de cuadrados en Y). Esto es:

$$VT = \sum \left(Y - \overline{Y}\right)^2$$

Si al término entre paréntesis se le suma y le resta una misma cantidad Y* este no se altera. Realizando el reemplazo y las cuentas, resulta:

$$VT = \sum \left(Y - \overline{Y}\right)^2 = \sum \left(Y - Y^* + Y^* - \overline{Y}\right)^2 = \sum \left(Y - Y^*\right)^2 + \sum \left(Y^* - \overline{Y}\right)^2 = VNE + VE$$

Donde VNE es la *Variación No Explicada* porque los valores de Y se comportan en forma aleatoria o no previsible. Mientras que el segundo término, cada valor de la diferencia tiene un patrón

definido. Y* es la recta de regresión de la cual se conoce su fórmula y los valores de X respectivos son fijados por el investigador; además su promedio es una constante. Esto significa que se tiene de alguna manera una "explicación" y por ello VE es la *Variación Explicada*. En resumen, a la variabilidad total se la ha descompuesto en dos términos, en forma análoga a como se procedió con la suma de cuadrados en Anova.

Como el cociente entre la variación explicada y la total, es el cuadrado del *Coeficiente de Correlación* muchas veces se confunden ambos temas: Regresión y Correlación, como si fuese la misma cosa. La única analogía es que los cálculos para ambos modelos son casi los mismos, como se verá más adelante. Por ahora, se puede continuar el ejemplo del punto anterior con:

Paso 5) En la octava columna se calculan los valores de Y* (estimados a partir de la recta de regresión) esto es: $Y^* = a + b\,X$

Paso 6) En la novena columna se calculan las diferencias $Di = Yi - Yi^*$

Paso 7) En la décima columna se calculan los cuadrados de estas diferencias y su suma permite obtener la Variación No Explicada:

$$VNE = \sum (Y - Y^*)^2 = 0{,}61606$$

Paso 8) En la undécima columna se calculan las diferencias y* entre cada valor Yi y su promedio. Esto es, a la columna 2 se le resta una constante: 6,0222.

Paso 9) En la última columna se elevan al cuadrado los valores anteriores y su suma permite obtener la Variación Explicada:

$$VE = \sum \left(Y^* - \overline{Y}\right)^2 = \sum (y^*)^2 = 23{,}514$$

Paso 10) Se calculan los grados de libertad respectivos, sabiendo que si hay n mediciones, la VT tendrá n términos cuadráticos libres, a los cuales hay que restar un grado de libertad por la ecuación de vínculo entre ellos, dada con el cálculo de su media. Por lo tanto: $\upsilon_T = n - 1 = 8$. Por su parte la Variación Explicada tendrá un único grado de libertad $\upsilon_E = 1$. Entonces, los grados de libertad de la Variación No Explicada son: $\upsilon_{NE} = n - 2 = 7$.

Paso 11) Se puede armar ahora la Tabla de Regresión, en forma análoga al Cuadro de Anova:

TABLA DE REGRESIÓN

Fuente de Variación	Suma de cuadrados	Grados de Libertad	Cuadrados Medios	F
Explicada (debida a la regresión lineal)	23,5145	1	23,51	267,2***
No Explicada (error)	0,6161	7	0,088	
Total	24,1306	8		

El cuadrado medio explicado se debe a la regresión lineal y mide la cantidad de variación de Y, tenida en cuenta por la variación de X. Si a este valor se lo resta de la variación total, la variación residual o remanente es la no explicada y se la usa como el cuadrado medio error. El valor del estadígrafo F se lo calcula como el cociente entre estos cuadrados medios. En este caso su valor tan grande hace altamente significativo el rechazo de la hipótesis nula, que suponía que no había regresión. O sea, se tiene evidencia científica validada por el modelo, de que gran parte de la varianza encontrada puede ser explicada por la regresión de Y sobre X.

Se pueden hacer otro ensayo de hipótesis acerca del "*coeficiente de regresión"*: b como se muestra a continuación:

(H_0) $\beta = 0$ No hay regresión en la población de donde se tomaron las muestras.
(H_1) $\beta \neq 0$ Hay regresión.

El ensayo usa el modelo de Student con $t = (b - \beta) / DS_b$ versus $t_{\alpha;\nu} = t_{\alpha;n-2}$
Donde el error típico de estimación de b está dado por:

$$DS_b = \sqrt{\frac{MSerror}{\sum x^2}} = \sqrt{\frac{0,088}{8301,39}} = 0,003256$$

Luego $t = (- 0,05322 - 0)/0,003256 = -16,35^{***}$ $(t_{0,001\ ;\ 7} = 5,408)$
Y se rechaza la hipótesis nula con valores altamente significativos.
También se pueden hallar los Límites de confianza para el coeficiente de regresión con:

$$\beta \in (b \pm t_{\alpha;\nu}\ DS_b)$$

Donde para un 95% de confianza resulta $t_{\alpha;\nu} = t_{0,05,\ 7} = 2,356$. O sea:

$$\beta \in (- 0,053 \pm 0,008) \rightarrow 95\%\ CI\ (-0,061\ ;\ -0,045)$$

Se concluye que el valor verdadero del coeficiente de regresión β está dentro de dicho intervalo, por lo tanto es diferente de cero: hay regresión (con un 95% de probabilidades a favor).

21.7 Regresión por el origen: Recta de Calibración

En muchas oportunidades, la teoría empleada para la regresión exige que la recta pase a través del origen de coordenadas. Entonces, ya se tiene un punto para el cual no se encontrará variaciones en el muestreo. Tal punto debe tratarse en una forma diferente a otro cualquiera observado. Un ejemplo de esto es el caso visto del número de anillos de un árbol y su edad. Otro, más frecuente en Bioquímica y Farmacia, es el caso de la *Recta de Calibración* de un instrumento de laboratorio cualquiera, como una balanza, un espectrofotómetro, etc. Generalizando, para todo instrumento que requiera hacer el "*ajuste del cero*" antes de comenzar a usarlo. En una balanza, esto es la primer pesada en vacío, cuando se ajusta la escala al cero luego de ser nivelada.

En un espectrofotómetro, es la primera lectura colocando agua destilada en la cubeta. Para ilustrar este caso se presentan los siguientes datos, con dos enunciados diferentes:

Ejemplo 1) Para determinar si un espectrofotómetro está calibrado se han medido 13 valores de referencia o patrones (**Y**) (soluciones calibradas) y los valores observados de transmitancia (**X**) se muestran en la siguiente tabla – ver caso práctico en Apéndice 2:

N	X	Y
1	13,6	52
2	13,9	48
3	21,1	72
4	25,6	89
5	26,4	80
6	39,8	130
7	40,1	139
8	43,9	173
9	51,9	208
10	53,2	225
11	65,2	259
12	66,4	199
13	67,7	255
Total	528,8	1929

$\sum X^2 = 26.062,1$ $\sum Y^2 = 356.259$

$\sum XY = 95.755,7$ $b = \frac{\sum XY}{\sum X^2} = 3,67$

Teniendo en cuenta que la recta pasa por el origen, la ecuación de la misma es:

$$Y^* = 3,67\ X$$

Para efectuar la validación estadística se usan las relaciones siguientes:

$$\sum x^2 = \sum (X - \overline{X})^2 = \sum X^2 - \left(\sum X\right)^2 / N = 26.062,1 - [(528,8)^2 / 13] = 4.552,14$$

$$\sum y^2 = \sum (Y - \overline{Y})^2 = \sum Y^2 - \left(\sum Y\right)^2 / N = 356.259 - [(1.929)^2 / 13] = 70.025,08$$

$$\sum xy = \sum XY - \left(\sum X\right)\left(\sum Y\right) / N = 95.755,7 - [(528,8)\ (1.929)\ (1/13)] = 17.289,92$$

Entonces ahora se pueden calcular las variabilidades siguientes:

$VT = \sum (Y - \overline{Y})^2 = \sum y^2 = 70.025,08$ con 12 grados de libertad

$$VE = \sum \left(Y^* - \overline{Y}\right)^2 = \frac{\left(\sum xy\right)^2}{\sum x^2} = \frac{(17.289,92)^2}{4.552,14} = 65.670,51$$ con 1 grado de libertad

$VNE = \sum \left(Y - Y^*\right)^2 = VT - VE = 70.025,08 - 65.670,51 = 4.354,57$ con 11 grados de libertad

Con estos datos se puede armar la tabla de regresión siguiente

Fuente de Variación	Suma de cuadrados	Grados de Libertad	Cuadrados Medios	F
Explicada (debida a la regresión lineal)	65.670,51	1	65.670,51	165,9***
No Explicada (error)	4.354,57	11	395,87	
Total	70.025,08	12		

Se tiene prueba altamente significativa de que existe la regresión de Y sobre X, aunque se ha tomado la ecuación general sin tener en cuenta que pasa por el origen.

Cuando se tengan dudas que la recta de regresión pasa por el origen, o sea a = 0, se puede hacer otra validación estadística con:

Ho : a = 0 La recta pasa por el origen.
H1 : a ≠ 0 La recta no pasa por el origen.

El valor muestral se calcula considerando que en el centro de gravedad de los datos debe ser:

$$y = Y - \overline{Y} = b\,x = \left(\frac{\sum xy}{\sum x^2}\right).\ (X - \overline{X})$$

$Y - 148,4 = 3,67\ (x - 40,68) = 3,67\ x - 149,29$

$Y = -0,89 + 3,67\ x$

Entonces, con esta recta se tendrá la contribución atribuible a la media con:

La variación debida a la media es: $(\Sigma\ Y)^2 / N = (1929)^2 / 13 = 286.233,9$

La variación debida a b es: $VE = \frac{\left(\sum xy\right)^2}{\sum x^2} = 65.670,5$

Entonces la total es 286.233,9 + 65.670,5 = 351.904,4 con 11 grados de libertad

Como la suma de cuadrados debida a la regresión es: $(\sum XY)^2 / \sum X^2 = 351.819$

La regresión adicional debida al ajuste de la media es: SS m = (351.904 – 351.819) = 85

Como tiene un grado de libertad es: MS m = SS m / 1 = 85 entonces,

F = MS m / MS error = 85 / 395,87 = 0,215 No significativo

Se concluye que no hay evidencia que muestre que la recta no pasa por el origen.

Sin embargo el problema principal aquí, no es determinar si hay regresión (porque ya se sabe que así será), esto es no hace falta probar que $b \neq 0$; sino determinar que el valor encontrado no difiera significativamente del *factor* de dilución correcto, que es lo que usa el bioquímico para sus mediciones. Imaginando que el valor esperado del factor de dilución es $\beta = 3,5$:

Ho : $b = \beta = 3,5$ El sistema está bien calibrado para hacer las mediciones.
H1 : $b \neq \beta$ El sistema no está bien calibrado.

Se usa el modelo de Student con:

$t = (b - \beta) / DS_b$ versus $t_{\alpha;\nu} = t_{\alpha;n-2}$

Donde el error típico de estimación de b está dado por:

$$DS_b = \sqrt{\frac{MSerror}{\sum x^2}} = \sqrt{\frac{395,87}{4.552,14}} = 0,295$$

Luego: $t = (3,67 - 3,5) / 0,295 = 0,58$ (no significativo) ($t_{0,95;\,11} = 2,201$)

Por lo que no se puede rechazar la hipótesis nula. No hay prueba estadística como para creer que el sistema de medición no está calibrado. Se puede usar otro ejemplo, si se tratase de una balanza, la idea es que cada valor medido Y, de los patrones utilizados X, sea $Y = X$. Esto es, una recta que pasa por el origen a 45º, o bien $\beta = 1$.

Ejemplo 2) En un estudio de Bacteriología, se midieron las reversiones inducidas a la independencia por 10^7 células sobrevivientes (Y), por dosis (ergs / bacterias) 10^{-5} (X) de *Escherichia coli* estreptomiceno-dependiente, sometidas a radiación ultravioleta monocromática de 2,967 Å de longitud de onda. (Datos de Zelle, M.R.; Univ. Cornell, del libro de Steel y Torrie).

														Total
X	13,6	13,9	21,1	25,6	26,4	39,8	40,1	43,9	51,9	53,2	65,2	66,4	67,7	**528,8**
Y	52	48	72	89	80	130	139	173	208	225	259	199	255	**1929**

Para simplificar los cálculos se han tomado los mismos datos del enunciado anterior. Por lo tanto la recta de calibración está dada por la ecuación:

$$Y^* = 3,67\ X$$

Habrá 3,67 retornos inducidos por dosis. Entonces la recta de regresión de retornos inducidos Y, por la dosis administrada X se expresa de la manera anterior.

Lo primero es probar que la recta de regresión existe, esto es probar que $b \neq 0$ como se hizo en el caso anterior. Luego si llegase a haber dudas respecto a si la recta pasa o no, por el origen de coordenadas, se puede hacer un test como se vio más arriba. El recurso utilizado aquí es transformar la variable medida para que el resultado se aproxime a una línea recta.

21.8 Más de un valor de Y

Muchas veces en la práctica se encuentran casos donde se obtienen más de un valor de Y por cada valor de X controlado por el investigador. El ordenamiento de los datos se parece mucho al orden en que se presentan las observaciones en un análisis de varianza de un factor. Para ilustrar este punto, se ha desarrollado un ejemplo con tamaños muestrales diferentes que es el caso más completo.

Ejemplo 1) Un investigador mide los porcentajes *p* de supervivencia del coleóptero *Tribolium castaneum* a cuatro densidades de siembra. Los datos de los porcentajes fueron transformados con Y = arc sen $\sqrt{p}$ porque cumplen mejor los supuestos de distribución normal y homoscedasticidad. El número de huevos por gramo de arena es la variable X controlada por el investigador. La supervivencia se calcula por el número de insectos que llegan a la edad adulta. Se preparan 4 densidades de siembra diferente y los resultados, transformados se muestran a continuación:

	Densidad de siembra X			
	5 / g	**20 / g**	**50 / g**	**100 / g**
Supervivencia Y	61,68	68,21	58,69	53,13
	58,37	66,72	58,37	49,89
	69,30	63,44	58,37	49,82
	61,68	60,84		
	69,30			
Total	320,33	259,21	175,43	152,84
Ni	5	4	3	3
Medias	64,07	64,80	58,81	50,95

Ref: Ejemplo de Sokal-Rohlf (pág. 480)

Se desea saber si existen diferencias en la supervivencia de los cuatro grupos y además, si se puede establecer una línea de regresión de supervivencia sobre la densidad.

Para resolver este caso se procederá en dos etapas. En la primera se busca mediante modelos de ANOVA decidir si hay diferencia entre los grupos de siembra. En la segunda se procede con el análisis de regresión. Por regla general, si no hay significación en ANOVA es bastante improbable que exista una línea de regresión.

Etapa 1) ANOVA: Se procede con los pasos habituales para este modelo:

Paso 1) Se calcula la suma total de las observaciones T = 907,81 = $\sum\sum Y$

Paso 2) Se calcula la suma de los cuadrados de los datos: Tx^2 = 55.503,6547

Paso 3) Se calcula la suma de los totales grupales al cuadrado, divididos por su tamaño muestral respectivo:

Tx = $[(320,33)^2 / 5] + [(259,21)^2 / 4] + [(175,43)^2 / 3] + [(152,84)^2 / 3]$ = 55.364,968

Paso 4) Se calcula el término de corrección T^2 / N = 54.941,2664

Paso 5) Se calculan las sumas de cuadrados con:

SS_T = *Paso 2 – Paso 4* = 55.503,6547 – 54.941,2664 = 562,3883

SS_E = *Paso 3 – Paso 4* = 55.364,9680 – 54.941,2664 = 423,7016

SST = SST – SSE = 562,3883 – 423,7016 = 138,6887

Paso 6) Los grados de libertad son $\nu_T = N - 1 = 14$; $\nu_E = a - 1 = 3$ y $\nu_D = N - a = 11$

Paso 7) Se formula la Tabla de ANOVA como sigue:

Variación	SS	ν	MS	F
Entre grupos	423,7016	3	141,2339	11,2**
Dentro de grupos	138,6777	11	12,6079	
Total	562,389	14		

Los grupos difieren muy significativamente entre sí.

Etapa 2) Regresión: Ahora se debe comprobar si las diferencias entre los valores de supervivencia pueden ser explicados por una regresión lineal sobre la densidad de siembra.

Paso 8) Se calcula la sumatoria de los valores de X multiplicados por su tamaño muestral con:

$$\sum \text{Ni X} = 5\,(5) + 4\,(20) + 3\,(50) + 3\,(100) = 555$$

Paso 9) Se calcula la sumatoria de los valores de X^2 multiplicados por su tamaño muestral con:

$$\sum \text{Ni X}^2 = 5\,(5)^2 + 4\,(20)^2 + 3\,(50)^2 + 3\,(100)^2 = 39.225$$

Paso 10) Se calcula la sumatoria de los productos de X e $\overline{Y}$ por su respectivo tamaño muestral con:

$$\sum \text{Ni X } \overline{Y} = \sum X \sum Y = 5\,(320,33) + 20\,(259,21) + 50\,(175,43) + 100\,(152,84) = 30.841,35$$

Paso 11) Se calcula el término de corrección para X con:

$$\text{TC x} = \frac{\left(\sum \text{Ni X}\right)^2}{\sum \text{Ni}} = (\textit{Paso 8})^2 \;/\; \text{N} = (555)^2 \;/\; 15 = 20.535$$

Paso 12) Se calcula la suma de cuadrados de *x* con:

$\sum x^2 = \sum Ni\, X^2$ - TC x = *Paso 9 – Paso 11* = 39.225 – 20.535 = 18.690

Paso 13) Se calcula la sumatoria de los productos:

$$\sum x.y = \sum X\left(\sum Y\right) - \frac{\left(\sum NX\right)\left(\sum\sum Y\right)}{\sum Ni} = Paso\ 10 - \frac{Paso\ 8\,.\,Paso\ 1}{N} = -\ 2.747{,}62$$

Paso 14) Se calcula la suma de los cuadrados explicada con:

$$SS\ ex = \frac{\left(\sum XY\right)^2}{\sum X^2} = (Paso\ 13)^2 / Paso\ 12 = (-2.747{,}62)^2 / (18.690) = 403{,}9281$$

Paso 15) Se calcula la suma de cuadrados no explicada:

SS no ex. = SS_E – SS ex. = *Paso 5 – Paso 14* = 423,7016 - 403,9281 = 19,7735

Paso 16) Se arma la Tabla de Anova completada con el análisis de regresión:

TABLA DE ANOVA + REGRESION

Fuente de Variación	Suma de cuadrados	Grados de Libertad	Cuadrados Medios	F
Entre Grupos**	423,7016	3	141,2339	11,20**
Regresión lineal	403,9281	1	403,9281	40,86*
Desvíos respecto a la Regresión	19,7735	2	9,8868	< 1(ns)
Dentro grupos (error)	138,6867	11	12,6079	
Total	562,3883	14		

Para comprobar si las desviaciones respecto de la regresión lineal son significativas se hace el ensayo: F = $MS_{Y;X}$ / MS_D < 1 por lo tanto se acepta la Ho, que las desviaciones respecto a la regresión lineal son nulas. Esto significa que no hay variación residual, o dispersión, alrededor de la línea de regresión. Por lo tanto se acepta a la recta como una buena explicación.

El siguiente ensayo es para determinar si existe la regresión lineal, es decir si b difiere significativamente de cero. Para eso se hace el ensayo:

F = MS_b / $MS_{Y;X}$ = (403,9281) / (9,8868) = 40,86* >> $F_{0,95\ ;\ 1\ ;\ 2}$ = 18,5

Luego se tiene evidencia significativa, como para afirmar que existe una recta de regresión que explica la regresión lineal de la supervivencia, respecto a la densidad de siembra. Resta entonces encontrar dicha recta:

Paso 17) Se calcula el coeficiente de regresión b con:

$b = \frac{(\sum XY)}{\sum X^2}$ = *Paso 13 / Paso 12* = (- 2.747,62) / (18.690) = - 0,14701

Paso 18) Se calcula la ordenada al origen a con:

$a = \overline{Y} - b\ \overline{X} = (T / N) - b [(\sum Ni\ X) / N] = (907,81 / 15) – (-0,14701 [555 / 15] = 65,96$

Por lo tanto la recta de regresión se expresa con:

$$Y^* = 65,96 – 0,14701 . X$$

Se ha probado que a medida que la densidad de siembra aumenta, la supervivencia disminuye. Y que esta relación se puede expresar con la ecuación de arriba.

21.9 Curvas de regresión

El caso anterior era el más simple, cuando la curva de regresión es una recta. Pero para el caso más general la *curva de regresión* toma una forma polinomial con:

$$Y^* = a + b\ X + c\ X^2 + d\ X^3 + ...$$

La idea es que cualquier curva puede ser aproximada con un desarrollo en serie polinomial. Ahora se tiene un conjunto de potencias crecientes de la variable independiente X, cada una con un coeficiente de regresión diferente: a, b, c, d, etc. Por ejemplo, en el caso de una parábola habrá tres términos polinómicos. A medida que se utilicen potencias más altas, el ajuste de la curva de regresión a los datos reales, será cada vez mejor. Sin embargo, con cada potencia añadida se perderá un grado de libertad y se necesitarán más mediciones. Si n = 5 datos, para el cuadrado medio residual o error, los grados de libertad son n-2 = 3, entonces el polinomio mayor que se podrá usar es el de tercer grado. Por otra parte, es muy raro encontrar polinomios de más de tres grados en las investigaciones biológicas. Los más comunes son:

$Y^* = a + b\ X$ (recta de regresión)

$Y^* = a' + b'\ X + c'\ X^2$ (parábola de regresión)

$Y^* = a'' + b''\ X + c''\ X^2 + d''\ X^3$ (parábola cúbica de regresión)

Los coeficientes de cada una de las tres curvas anteriores son diferentes, por lo que deben ser calculados cada vez. Luego de obtenida la recta, se puede aumentar una potencia de X y buscar la parábola de regresión. Pero entonces, hay que recomenzar los cálculos de nuevo y por regla general, los nuevos coeficientes hallados (a', b') son diferentes de los anteriores (a, b). Como estas regresiones polinomiales son ajustes empíricos, si al comprobar la significación esta resulta significativa, significa que ahora se tiene un mejor ajuste que el lineal y conviene intentar la parábola cúbica.

Se recomienza todo de nuevo y los nuevos coeficientes serán diferentes a los anteriores con lo que la significación deberá ser testeada otra vez. Es claro, que si antes de comenzar se hubiese tenido información acerca del tipo de polinomio buscado, se hubiera comenzado por allí y no con la recta. Los cálculos y ensayos relacionados con este tema se pueden encontrar en el libro en Steel y Torrie (1960) mencionados en la bibliografía.

21.10 Problemas propuestos

1) Marcar la respuesta correcta a cada una de las afirmaciones siguientes, o completar la frase:

1) Para un mismo problema se puede usar Regresión o Correlación.	**V**	**F**
2) El análisis de regresión se usa cuando se conoce la relación teórica Y = f (X).	**V**	**F**
3) Para el análisis de peso y talla se puede usar la regresión.	**V**	**F**
4) El método de los mínimos cuadrados es el usado en regresión.	**V**	**F**
5) Explicar los diseños experimentales posibles en regresión:.....	.....	
6) En un Modelo I la variable dependiente se mide sin error.	**V**	**F**
7) Explicar las 4 hipótesis básicas de estos modelos:.....	.....	
8) Todas las muestras a lo largo de la curva de regresión son homocedásticas.	**V**	**F**
9) Las ecuaciones paramétricas de regresión se deducen minimizando la suma de cuadrados.	**V**	**F**
10) Expresar el par de ecuaciones normales de regresión:.....	.....	
11) La ecuación de regresión de Y sobre X, es igual a la de X sobre Y.	**V**	**F**
12) Explicar los pasos básicos para los cálculos cortos de regresión:.....	.....	
12) La variación Total es la suma del cuadrado de las diferencias entre cada Y y su media.	**V**	**F**
13) La Variación No Explicada es la suma del cuadrado de las diferencias de Y con X.	**V**	**F**
14) La Variación explicada es la diferencia de VT – VNE.	**V**	**F**
15) El ensayo de hipótesis en regresión se hace con una Tabla de ANOVA.	**V**	**F**
16) Para testear si hay regresión se puede usar el Modelo de Student.	**V**	**F**
17) Para testear si la recta pasa por el origen hay que:.....	.....	
18) Una curva de calibración instrumental debe pasar por el origen.	**V**	**F**
19) Es lo mismo que haya uno o más valores de Y por cada X en los cálculos.	**V**	**F**
20) Cuando hay más de un valor de Y primero se testea que haya regresión.	**V**	**F**
21) Las curvas de regresión se resuelven con la técnica de polinomios ortonormales.	**V**	**F**
22) Las transformaciones más comunes para regresión son:.....	**.....**	
23) Para las curvas de dosis-mortalidad se usa la transformación probit.	**V**	**F**

2) Se desea encontrar la recta de regresión para los siguientes datos, y luego validar los resultados obtenidos tanto para a como para b:

X	Y
0	5
10	20
20	40
30	60
40	78
50	98
60	118
70	135
80	152

3) Se desea saber si la balanza investigada está calibrada. Los datos recogidos son:

X	1	2	5	10	20	50	100	150	200	500	1000	1500	2000	3000
Y	1,2	1,9	4,8	10,5	21	49	98	152	202	496	1010	1520	1986	3020

4) Encontrar la recta de regresión para cuando hay 5 valores de Y por cada X, realizando las validaciones correspondientes:

	X			
	15	**20**	**30**	**50**
Y	65	60	53	43
	68	58	52	44
	67	59	51	42
	69	57	54	41
	70	58	50	40

5) Encontrar la recta de regresión para cuando hay 3 valores de Y por cada X, realizando las validaciones correspondientes:

	X			
	15	**20**	**30**	**50**
Y	65	60	53	43
	68	58	52	44
	67	59	51	42

6) Para los datos de la tabla siguiente decidir si existe una recta de regresión entre las dos variables presentadas.

X	95	115	135	155	175	195	221	235	255	275
Y	90	100	120	140	160	170	180	190	200	215

7) Para los datos de la tabla siguiente decidir si existe una recta de regresión entre las dos variables presentadas.

X	**Y**
40	16
60	28
80	43
100	55
120	79
140	96
160	110
180	125
200	150

Apéndice 1:

La diferencia entre un valor medido (Y_i) y su valor estimado (Y_i*) es $D_i = Y_i - Y_i$*
Para minimizar la suma de los cuadrados de estas diferencias se usan las relaciones:

$$\frac{\partial}{\partial a}\sum D_i^2 = 0$$
$$\frac{\partial}{\partial a}\sum \left[Y_i - (a + bX_i)\right]^2 = 0$$

Si la sumatoria es convergente, la derivada de la suma es igual a la suma de las derivadas

$$0 = \sum 2.\left[Y_i - (a + bX_i)\right].\left[0 - 1 - 0\right] = (-2)\sum \left[Y_i - a - b.X_i\right] = \sum Y_i - \sum a - \sum b.X_i$$

Operando con esta relación se obtiene la primera ecuación paramétrica de la recta que es:

$$\sum Yi = a.N + b\sum X_i$$

Para obtener la segunda ecuación paramétrica de la recta de regresión se usa la relación:

$$\frac{\partial}{\partial b}\sum D_i^2 = 0$$

$\frac{\partial}{\partial b}\sum \left[Y_i - (a + bX_i)\right]^2 = 0$ Análogamente al anterior si la sumatoria converge es:

$$0 = \sum 2.\left[Y_i - (a + bX_i)\right].\left[0 - 0 - X_i\right] = (-2)\sum \left[Y_i.X_i - a.X_i - b.X_i^2\right]$$

Resolviendo esta relación se obtiene la segunda ecuación paramétrica:

$$\sum X_i Yi = a.\sum x_i + b\sum X_i^2$$

Para mostrar que es un mínimo, basta derivar una vez más y ver que el resultado sea positivo

$$\frac{\partial^2}{\partial^2 a}\sum D_i^2 = (-2)\ (-N) = 2\ N$$

$$\frac{\partial^2}{\partial^2 b}\sum D_i^2 = (-2)\ (-\sum X_i^2) = 2\ \sum X_i^2$$ (que siempre será un número positivo)

Con esto se demuestra que la recta de regresión minimiza la suma de los cuadrados de las diferencias entre cada valor medido y su correspondiente valor estimado con la ecuación de la recta. Por eso también se la llama: *recta de mínimos cuadrados*.

Apéndice 2:

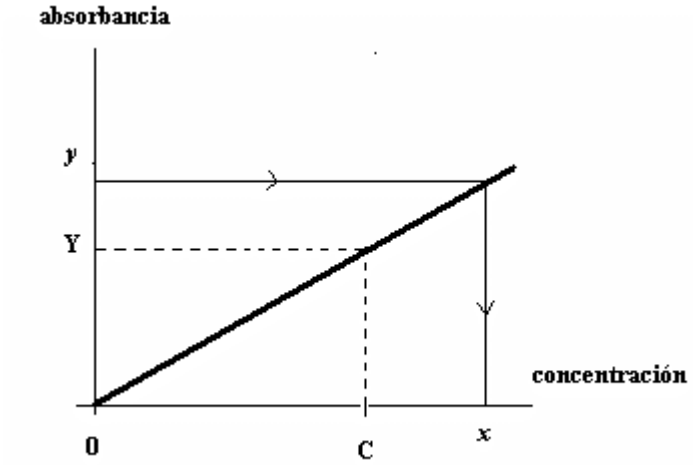
En forma diaria el bioquímico prepara su espectro para realizar las mediciones. Lo primero es el ajuste de cero. Para esto carga agua destilada en la cubeta del espectro (previamente nivelado) y el resultado tiene que dar una absorbancia nula (o 100% de transmitancia). Ajusta el aparato para tener esos valores, es decir para una concentración nula (X = 0) debe tener una absorbancia nula (Y = 0). Este es el primer punto de la recta de calibración denominado ajuste de cero.

Para sacar el segundo punto usa una sustancia calibrada, control o estándar (que viene provisto con cada kit de mediciones). Suponiendo que el estándar (de glucosa) tiene una concentración de 0,9 (expresada en las unidades habituales de trabajo), entonces usa el kit para proceder a la medición y obtiene una ligera coloración de la sustancia final que coloca en la cubeta del espectro. Esta coloración hace que no toda la luz enviada atraviese la cubeta, sino que un cierto porcentaje será absorbido (o desviado) y mide con el espectro la absorbancia (Y = A) que corresponde a una sustancia que tiene una concentración (X = C) de glucosa.

La costumbre es calcular el *factor* (k) con esos dos valores como si fuera una regla de tres simple: si el paciente tiene una absorbancia *y* entonces tendrá una concentración *x* calculada con:

$x = (C / A)\, y = k\ y$

Lo que se está haciendo en realidad es usar una recta de calibración:



Donde la tangente del ángulo de la recta es la inversa del factor. A cada paciente se le mide la absorbancia *y* para obtener la concentración buscada *x*.

Notar que este procedimiento diario implica que solo dos puntos son suficientes para obtener una recta de calibración del espectro. Cuando en realidad habría que usar más puntos y con ellos el procedimiento de regresión, para obtener una buena recta de calibración.