Correlación

Sean X y Y las mediciones de dos caracteres *x* y *y* las desviaciones de tales mediciones respecto de su media. La correlación entre los valores X y Y medidos en los mismos individuos se estima mediante el coeficiente de correlación (*r*).

$$r = \frac{Covarianza_{x\,y}}{Media\ geométrica\ de\ las\ varianzas_{x,\,y}}$$

$$r = \frac{\frac{1}{n-1}\sum(X-\overline{X})(Y-\overline{Y})}{\sqrt{\frac{1}{(n-1)}(\sum(X-\overline{X})^2(Y-\overline{Y})^2}};\ definición\ del\ coeficiente\ de\ correlación$$

$$r = \frac{(\sum(X-\overline{X})(Y-\overline{Y})}{\sqrt{(\sum(X-\overline{X})^2(Y-\overline{Y})^2}} = \frac{SP_{xy}}{\sqrt{SS_x SS_y}} = \frac{(x\cdot y)}{\sqrt{(x\cdot x)(y\cdot y)}};\ formula\ común$$

Los cálculos de $SS_x\ y\ SS_y$ los vimos anteriormente, y el cálculo de los productos X, Y, es:

$$SS_x SS_y = \sum XY - \frac{\sum X \bullet \sum Y}{n}$$

Ejemplo: Suponga que medimos dos caracteres (X, Y) diferentes en la misma planta en 7 individuos. También, puede suponer que se midieron en progenitores (X) y progenie (Y); o bien, que es el *mismo* carácter medido en dos experimentos (o ambientes) distiontos.

X	4	4	5	5	6	7	8	$\sum X = 39$	$\sum(X-\overline{X})^2 = 13.71$
Y	8	10	11	12	14	16	15	$\sum Y = 86$	$\sum(Y-\overline{Y})^2 = 49.43$
$X\cdot Y$	32	40	55	60	84	112	12	$\sum XY = 503$	

$$\sum XY = 503;\quad \frac{\sum X \bullet \sum Y}{n} = \frac{39\cdot 86}{7} = 479.14$$

$$SS_x SS_y = 503 - 479.14 = 23.86$$

$$\sqrt{\sum\left(X-\overline{X}\right)^2\left(Y-\overline{Y}\right)^2} = \sqrt{13.71 \cdot 49.43} = 26.03$$

Finalmente,
$r = \frac{23.86}{26.03} = 0.917$.

Este resultado muestra que X y Y están altamente correlacionados. En el caso extremos, es decir, cuando X y Y muestran una variación idéntica (i.e., $x_i = y_i$), r = 1. En tal situación:

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

Si la variación en ambos caracteres es estrictamente proporcional, entonces también r =1. Por tanto, es importante notar que r no distingue entre una variación idéntica o proporcional. También puede ocurrir que las desviaciones de X o de Y sean iguales o proporcionales, pero en *direcciones opuestas*; por lo tanto, r = -1. En general, $-1 \geq r \leq +1$.

Supongan que X y Y varían de forma totalmente independiente. Entonces, esperamos que la suma de productos cruzados, $\sum\left(X-\overline{X}\right)\left(Y-\overline{Y}\right)$, sea cercana a cero cuando la muestra sea muy *grande*. Sin embargo, si la muestra es pequeña, existe la posibilidad aleatoria de que r no sea = 0.

El coeficiente de correlación estimado se compara con el valor esperado con n-2 grados de libertad, donde n es el número de pares de los caracteres X, Y. Nuestra hipótesis nula supone que la correlación verdadera entre X y Y es cero. Por tanto, juzgamos la correlación muestral con un cierto grado de significancia.

De la tabla se aprecia que la r es mayor a la r de tablas con un α de 0.01. La hipótesis se contrasta con *dos colas*. Es decir, la hipótesis establece que la r es mayor que el valor absoluto de r esperada con una significancia específica.

Regresión

La relación entre dos variables puede ser expresado por el coeficiente de regresión (b_{yx}), a diferencia de r que no diferencia igualdad y proporcionalidad en la variación de dos caracteres.

Tomando como ejemplo la estatura de progenitores -progenie podemos expresar la regresión de Y en X (Figura) mediante la ecuación:

$$Y-\overline{Y} = b_{yx}\left(X-\overline{X}\right)+e$$

Donde $Y-\overline{Y}$ es la desviación de una progenie respecto de la media de todas las progenies; $X-\overline{X}$ es la desviación de sus progenitores respecto la media de su generación; b_{yx} es el *coeficiente de regresión* y *e* (error) es una desviación aleatoria respecto en valor esperado.

Si fuera el caso, improbable, que los valores no se desviaran aleatoriamente la ecuación se restringe a

$$Y-\overline{Y}=b_{yx}\left(X-\overline{X}\right),$$

$$Y=b_{yx}\left(X-\overline{X}\right)+\overline{Y}$$

df \ α	0.2	0.1	0.05	0.02	0.01	0.001
1	0.951057	0.987688	0.996917	0.999507	0.999877	0.999999
2	0.800000	0.900000	0.950000	0.980000	0.990000	0.999000
3	0.687049	0.805384	0.878339	0.934333	0.958735	0.991139
4	0.608400	0.729299	0.811401	0.882194	0.917200	0.974068
5	0.550863	0.669439	0.754492	0.832874	0.874526	0.950883
6	0.506727	0.621489	0.706734	0.788720	0.834342	0.924904
7	0.471589	0.582206	0.666384	0.749776	0.797681	0.898260
8	0.442796	0.549357	0.631897	0.715459	0.764592	0.872115
9	0.418662	0.521404	0.602069	0.685095	0.734786	0.847047
10	0.398062	0.497265	0.575983	0.658070	0.707888	0.823305
11	0.380216	0.476156	0.552943	0.633863	0.683528	0.800962
12	0.364562	0.457500	0.532413	0.612047	0.661376	0.779998
13	0.350688	0.440861	0.513977	0.592270	0.641145	0.760351
14	0.338282	0.425902	0.497309	0.574245	0.622591	0.741934
15	0.327101	0.412360	0.482146	0.557737	0.605506	0.724657
16	0.316958	0.400027	0.468277	0.542548	0.589714	0.708429
17	0.307702	0.388733	0.455531	0.528517	0.575067	0.693163
18	0.299210	0.378341	0.443763	0.515505	0.561435	0.678781
19	0.291384	0.368737	0.432858	0.503397	0.548711	0.665208
20	0.284140	0.359827	0.422714	0.492094	0.536800	0.652378
21	0.277411	0.351531	0.413247	0.481512	0.525620	0.640230
22	0.271137	0.343783	0.404386	0.471579	0.515101	0.628710
23	0.265270	0.336524	0.396070	0.462231	0.505182	0.617768
24	0.259768	0.329705	0.388244	0.453413	0.495808	0.607360
25	0.254594	0.323283	0.380863	0.445078	0.486932	0.597446
26	0.249717	0.317223	0.373886	0.437184	0.478511	0.587988
27	0.245110	0.311490	0.367278	0.429693	0.470509	0.578956
28	0.240749	0.306057	0.361007	0.422572	0.462892	0.570317
29	0.236612	0.300898	0.355046	0.415792	0.455631	0.562047
30	0.232681	0.295991	0.349370	0.409327	0.448699	0.554119

Esto implica que existe una proporcionalidad de las desviaciones de las progenies y sus progenitores con sus respectivas medias. O, de forma más general, una proporcionalidad de las desviaciones de X y Y de sus medias.

Por tanto, aun en presencia de desviaciones, la ecuación nos da el valor más probable de Y para un valor de X, ya que el promedio de las desviaciones aleatorias de un valor esperado es cero ($\overline{e} = 0$).

Los valores esperados, designados con "y gorro" son:

$$\hat{Y} - \overline{Y} = b_{yx}(X - \overline{X})$$

Por tanto,

$$b_{yx} = \frac{\hat{Y} - \overline{Y}}{X - \overline{X}}$$

Todos los valores $\hat{Y}$ están en la recta de regresión (figura). El coeficiente de regresión es la tangente de α, donde α es el ángulo que forma la recta de regresión y el eje de la abscisa ($\frac{b}{a} = \frac{SS_{yx}}{SS_x}$), siempre que $\sum e^2$ sea mínimo. Esta condición se cumple cuando

$$b_{yx} = \frac{(\sum(X-\overline{X})(Y-\overline{Y})}{\sum(X-\overline{X})^2} = \frac{SP_{xy}}{SS_x}$$

Con esta expresión se calcula el coeficiente de regresión. La recta cruza el punto central del sistema con coordenadas $\overline{X}, \overline{Y}$ (*Figura*).

Ejemplo. Los datos que se anexan en la tabla contigua representan 8 pares (P_i) de valores de X, Y de un carácter expresado, por ejemplo, en progenitores e hijos. Con estos valores, podemos obtener el coeficiente de regresión.

X	Y	X Y
4	3	12
5	4	20
6	6	36
7	5	35
9	7	63
10	8	80
11	9	99
12	14	168

$\sum X = 64$ $\sum Y = 56$ $\sum XY = 513$

$\overline{X} = 8$ $\overline{Y} = 7$

$$\sum(X - \overline{X})^2 = 60$$

$$SP_{xy} = \sum XY - \frac{\sum X \bullet \sum Y}{n} = 513 - \frac{56 \cdot 64}{8} = 65$$

$$b_{yx} = \frac{(\sum(X-\overline{X})(Y-\overline{Y})}{\sum(X-\overline{X})^2} = \frac{SP_{xy}}{SS_x} = \frac{65}{60} = 1.083$$

Tabla. Valores de $\hat{Y}$ y *e*.

X	Y	X Y	$\hat{Y}$	$e = Y - \hat{Y}$
4	3	12	2.668	0.332
5	4	20	3.751	0.249
6	6	36	4.834	1.166
7	5	35	5.917	-0.917
9	7	63	8.083	-1.083
10	8	80	9.166	-1.166
11	9	99	10.249	-1.249
12	14	168	11.332	2.668

$\sum \hat{Y} = 56$ $\sum e = 0$

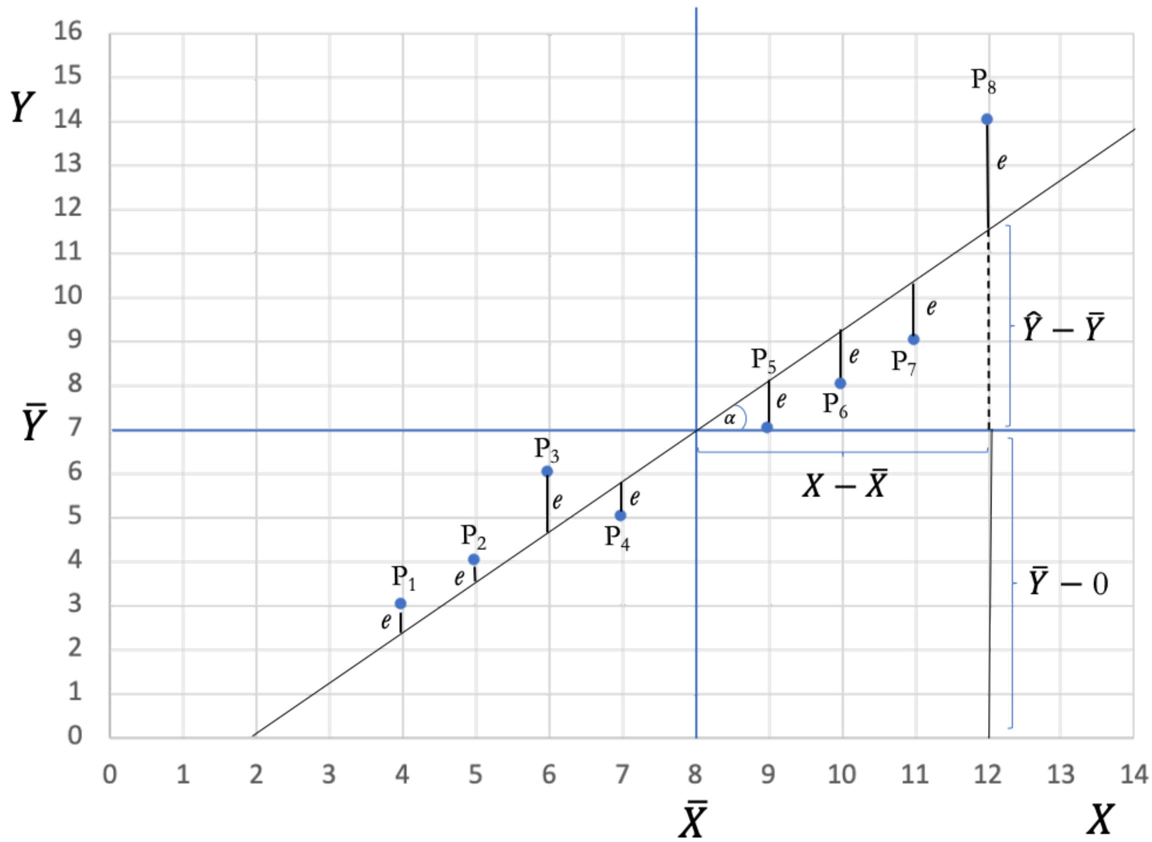

Figura. Regresión lineal de P_i pares de caracteres Y en X (datos en tabla). Se ilustra en valor de Y = 14 (P_8): $Y - \bar{Y} = e + (\hat{Y} - \bar{Y})$.

La desviación de cualquier valor de Y respecto a su media tiene dos componentes, el segundo es la desviación aleatoria ($e = Y - \hat{Y}$), el primero es la desviación de valor estimado de regresión en X.

$$Y - \bar{Y} = (\hat{Y} - \bar{Y}) + (Y - \hat{Y})$$

Si hay independencia entre ambos componentes, podemos establecer una igualdad entre las sumas de cuadrados, total SS_y, de la regresión y residual, con grados de libertad n-1, 1, y n-2, respectivamente:

$$\sum(Y - \bar{Y})^2 = \sum(\hat{Y} - \bar{Y})^2 + \sum(Y - \hat{Y})^2$$

Calculamos la suma de cuadrados para los componentes:

$(Y - \bar{Y})^2$	$(Y - \bar{Y})^2$	$(Y - \hat{Y})^2$
16	18.766224	0.110224

9	10.556001	0.062001
1	4.691556	1.359556
4	1.172889	0.840889
0	1.172889	1.172889
1	4.691556	1.359556
4	10.556001	1.560001
49	18.766224	7.118224

$\sum(Y-\overline{Y})^2 =$	$\sum(\hat{Y}-\overline{Y})^2 =$	$\sum(Y-\hat{Y})^2 =$
84	70.373	13.583

A partir de estos valores podemos construir un análisis de varianza:

Tabla. Análisis de la varianza del modelo de regresión

Efecto	*SS*	*d.f.*	*CM*	*F*	*P*
Modelo	70.373	1	70.373	31.09	
Error	13.583	6	2.263		
Total	84	7	12		

R^2 =0.837

R^2_{ajustada} = 1- =0.8113

Relación entre *b* y *r*

Dado que las ecuaciones para *b* y *r* son similares excepto por el denominador, si ambas variables tienen la misma varianza (i.e., $SS_x = SS_y$), entonces $b = r$. Si como en el ejemplo hipotético de que X y Y sean el mismo carácter en progenitores y su progenie, quiere decir que ambas generaciones poseen la misma varianza.

Ya que $b_{yx} = \dfrac{(\sum(X-\overline{X})(Y-\overline{Y})}{\sum(X-\overline{X})^2}$, y $r = \dfrac{(\sum(X-\overline{X})(Y-\overline{Y})}{\sqrt{(\sum(X-\overline{X})^2(Y-\overline{Y})^2}}$

$$b/r = \frac{\dfrac{(\sum(X-\overline{X})(Y-\overline{Y})}{\sum(X-\overline{X})^2}}{\dfrac{(\sum(X-\overline{X})(Y-\overline{Y})}{\sqrt{(\sum(X-\overline{X})^2(Y-\overline{Y})^2}}}$$

$$b = r\cdot \frac{\sqrt{(\sum (X-\overline{X})^2 (Y-\overline{Y})^2}}{\sum (X-\overline{X})^2} = r\cdot \sqrt{\frac{\sum (Y-\overline{Y})^2}{\sum (X-\overline{X})^2}}$$

Si los términos en la raíz los dividimos entre los grados de libertad (*n*-1),

Cuando no es así, $b = r\cdot \frac{s_y}{s_x}$. (*r* por la razón de las desviaciones estándar de X y Y).

Si $b^2 = r^2 \bullet \left(\sqrt{\frac{\sum (Y-\overline{Y})^2}{\sum (X-\overline{X})^2}} \right)^2$ se cumple la igualdad

$$b^2 \bullet \sum \left(X - \overline{X}\right)^2 = r^2 \bullet \sum \left(Y - \overline{Y}\right)^2$$

Recordemos que $\hat{Y} - \overline{Y} = b_{yx}(X - \overline{X})$

$$(\hat{Y} - \overline{Y})^2 = b_{yx}{}^2(X - \overline{X})^2 = r^2 \bullet \sum \left(Y - \overline{Y}\right)^2$$

$$\sum (Y - \overline{Y})^2 = r^2 \bullet \sum \left(Y - \overline{Y}\right)^2 + (1 - r^2)\cdot \sum \left(Y - \overline{Y}\right)^2$$

Es decir, la *SS*y se compone de r” debida a la regresión de X en Y y la otra, (por diferencia) 1-*r*, debida a las desviaciones aleatorias de la recta de regresión. Generalmente r^2 se interpreta como la proporción de varianza en Y asociada con la varianza en X, por lo que puede haber una relación de dependencia. En nuestro caso, la significancia se evalúa como *r*. Veamos el valor de $\boldsymbol{r^2}$ en la tabla de Anova.