UNIVERSITAT DE BARCELONA

Estudis correlacionals i predictius aplicats a l'educació.

Ormazábal, F. J. (Coord.) Vila, R.; Mateo, M.; Torrado, M.; Martínez, F.; Berlanga, V.; del Barrio, J.; Ruiz, A.

Departament de Mètodes d'Investigació i Diagnòstic en Educació (MIDE)
Facultat de Pedagogia
Universitat de Barcelona

Para citar la obra:
Ormazábal, F.J. (coord.) Vila, R.; Mateo, M.; Torrado, M.; Martínez, F.; Berlanga, V.; Del Barrio, J.; Ruiz, A. (2012) *Estudis correlacionals i predictius aplicats a l'educació.* Barcelona: Universitat de Barcelona. Dipòsit Digital http://hdl.handle.net/2445/21388

INDEX

1.1 Conceptes, elements i procés

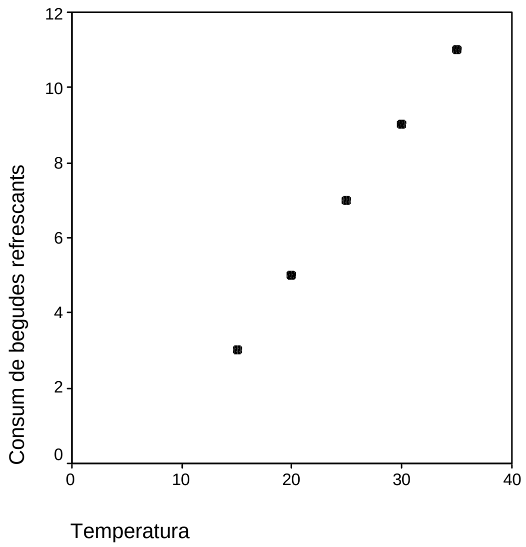
La correlació ens indica si hi ha relació entre dos o més variables, és a dir la variació conjunta (covariació) entre aquestes.

La relació o variació conjunta pot ser:

- ***Positiva:*** Quan augmenten els valors d'una de les variables augmenten també els de l'altra (o altres variables) i a l'inrevés.

  Exemple: La Temperatura (X) i el consum de begudes refrescants (Y).

X	Y
15	3
20	5
25	7
30	9
35	11



Gràfica de dispersió

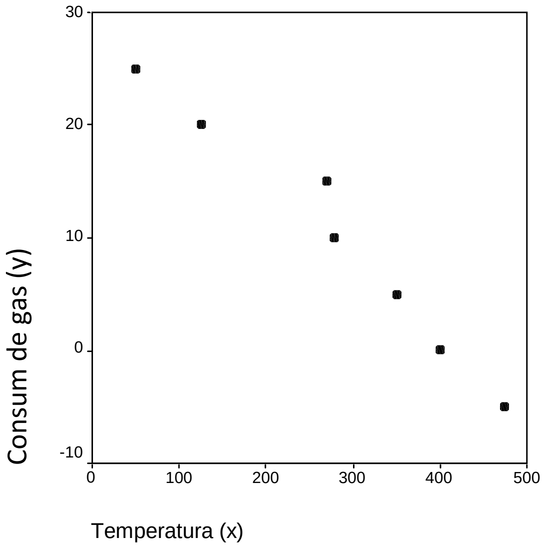
Quan augmenta la temperatura augmenta també el consum de begudes refrescants[1].

[1] Opcions del menú de l'SPSS per a obtenir la gràfica de dispersió: Gràficos > Cuadros de diálogo antiguos>Dispersión/Puntos... > Simple > Definir.

- ***Negativa:*** Quan augmenten els valors d'una de les variables disminueixen els de l'altra (o altres variables) i a l'inrevés.

  Per exemple: la temperatura (X) i el consum de gas natural (Y).

X	Y
-5	474
0	400
5	350
10	278
15	270
20	125
25	50



Quan augmenta la temperatura disminueix el consum de gas natural.

Gràfica de dispersió

- ***Nul·la:*** Les variables no guarden cap relació, o varien independentment una de l'altra.

El ***coeficient de correlació*** indica el grau de relació entre les variables i el sentit d'aquesta relació (positiva o negativa):

- Si el coeficient de correlació és 0 (correlació nul·la) indica que les variables són independents les unes de les altres.

- Si el coeficient és 1 significa que la correlació és perfecta positiva.
- Si el coeficient és -1 significa que la correlació és perfecta negativa.

La ***regressió*** indica la forma com es produeix la variació conjunta i això permet predir els valors d'una variables en funció dels de l'altra. Només té sentit quan existeix correlació o dependència entre les variables.

1.2 Correlació i regressió lineal simple

La correlació i regressió simple fan referència a l'estudi únicament entre dos variables.

Existeixen diferents coeficients de correlació en funció del tipus de variable amb què ens trobem (vegeu taula següent).

Coeficient de correlació	Variable X	Variable Y	SPSS
Pearson (r_{xy})	Quantitativa contínua (o discreta)	Quantitativa contínua (o discreta)	Analizar > Correlaciones > Bivariadas... > Pearson
Spearman (r_s ó ρ_{xy})	Mesurada amb una escala ordinal	Mesurada amb una escala ordinal	Analizar > Correlaciones > Bivariadas... > Spearman
Chi-quadrat (χ^2)	Qualitativa	Qualitativa	Analizar > Estadísticos descriptivos > Tablas de contingencia... > Estadísticos... > Chi-cuadrado
Biserial (r_b)	Quantitativa contínua (o discreta)	Dicotomitzada (qualitativa artificial)	--
Biserial puntual (r_{bp})	Quantitativa contínua (o discreta)	Qualitativa dicotòmica	--
Tetracòrica (r_t)	Dicotomitzada (qualitativa artificial)	Dicotomitzada (qualitativa artificial)	--

1.2.1 Correlació de Pearson

1. Verificar les condicions d'aplicació
2. Gràfica de dispersió i càlcul del coeficient de correlació mitjançant l'SPSS
3. Interpretació del coeficient de correlació
   Als estudis correlacionals les hipòtesis estadístiques són:

   - H_0 (hipòtesi nul·la): independència o correlació no significativa (variació conjunta suficientment petita com per ser explicada per la influència de l'atzar)

   - H_1 (hipòtesi alternativa): correlació significativa (variació conjunta suficientment gran com per no ser explicada només per la influència de l'atzar)

   a. No significativa → fi del procés
   b. Sí significativa → estudi de la regressió

4. Càlcul de la recta de regressió: predicció del valor més probable
   - $X' = \overline{X} + b_{xy} \cdot (Y - \overline{Y})$ o bé $X' = a_{xy} + b_{xy} \cdot Y$
   - $Y' = \overline{Y} + b_{yx} \cdot (X - \overline{X})$ o bé $Y' = a_{yx} + b_{yx} \cdot X$

5. Càlcul de l'interval de regressió (amb un determinat marge d'error):
   - $X \in X' \pm Z_{\alpha/2} \cdot S_{xy}$ on $S_{xy} = S_x \sqrt{1 - r_{xy}^2}$
   - $Y \in Y' \pm Z_{\alpha/2} \cdot S_{yx}$ on $S_{yx} = S_y \sqrt{1 - r_{xy}^2}$

A continuació desenvoluparem els passos anteriors en un exercici.

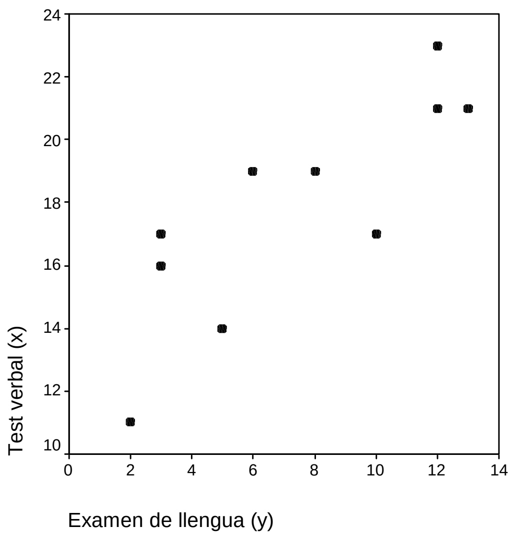
Es vol fer un estudi de correlació i regressió entre els resultats obtinguts a un test verbal (V) i a un examen de llengua (L). Per fer això disposem dels resultats d'un grup d'alumnes de batxillerat escollits a l'atzar:

1) Condicions d'aplicació

Subjecte	V(x)	L(y)
1	21	13
2	23	12
3	21	12
4	17	10
5	19	8
6	19	6
7	14	5
8	17	3
9	16	3
10	11	2

- Escales de mesura: interval o de raó
- Tipus de variables: quantitatives continues (també discretes)
- Tipus de distribució: Normal (es verifica mitjançant la prova de Kolmogorov)
- Funció lineal (es pot intuir en la gràfica)
- Homocedasticidad o homogeneïtat de variances (s'accepta que es compleix quan les dues variables segueixen la llei normal)

2) Gràfica de dispersió i càlcul del coeficient de correlació mitjançant l'SPSS



Correlaciones

		Test verbal (x)	Examen de llengua (y)
Test verbal (x)	Correlación de Pearson	1	,824**
	Sig. (bilateral)	,	,003
	N	10	10
Examen de llengua (y)	Correlación de Pearson	,824**	1
	Sig. (bilateral)	,003	,
	N	10	10

**. La correlación es significativa al nivel 0,01 (bilateral).

3) Interpretació del coeficient de correlació

Com la significació (,003) és menor que 0,05 aleshores rebutgem la H_0 amb un risc α de 0,05 i acceptem la H_1 i podem concloure que la correlació (0,824) és significativa (no pot ser explicada només per la influència de l'atzar).

4) Càlcul de la recta de regressió

A la correlació lineal simple, teòricament, hi ha dues rectes de regressió:

1. Recta de regressió d'Y en funció d'X ($R_{y/x}$). Expressa la variació de la variable Y en funció de la variable X. Permet predir el valor més probable (y') de la variable Y, en funció d'un determinat valor d'X.
2. Recta de regressió d'X en funció d'Y ($R_{x/y}$). Expressa la variació de la variable X en funció de la variable Y. Permet predir el valor més probable (x') de la variable X, en funció d'un determinat valor d'Y.

En la pràctica utilitzem solsament la primera (Y en funció d'X) per que a la variable dependent l'anomenem X i a la variable independent l'anomenem Y.

Suposem que per al nostre estudi ens interessa poder fer una estimació de la nota que es podrà treure en l'examen de llengua sabent la puntuació que té una persona en el test verbal, per tant necessitem calcular la recta de regressió d'Y en funció d'X.

El càlcul en l'SPSS es demana a través del menú Analizar > Regressión > Lineal...> on hem de posar com a "Dependiente" la variable de l'examen de llengua (Y) i com a "Independiente" la variable del test verbal (X).

Resumen del modelo

Modelo	R	R cuadrado	R cuadrado corregida	Error típ. de la estimación
1	,824[a]	,679	,639	2,505

a. Variables predictoras: (Constante), Test verbal (x)

Coeficientes[a]

Modelo		Coeficientes no estandarizados		Coeficientes estandarizados	t	Sig.
		B	Error típ.	Beta		
1	(Constante)	-9,661	4,222		-2,288	,051
	Test verbal (x)	,958	,233	,824	4,114	,003

a. Variable dependiente: Examen de llengua (y)

Els coeficients no estandaritzats ens donen els valors a_{yx} (-9,661) i b_{yx} (0,958) de la recta de regressió que necessitem. Per tant tenim que $Y' = -9{,}661 + 0{,}958 \cdot X$.

Suposant que una persona té una puntuació de 20 en el test verbal, el valor més probable en l'examen de llengua és Y' = -9,66 + 0,958 · 20 = 9,51.

5) Càlcul de l'interval de regressió (amb un determinat marge d'error)

$S_{yx} = S_y\sqrt{1-r_{xy}^2} = 3'9547\sqrt{1-0'824^2} = 2'24$ (vegeu nota [2])

$$Y \in Y \pm Z_{\alpha/2} \cdot S_{yx} = 9'51 \pm 1'96 \cdot 2'24 = 9'51 \pm 4'39 \prec \begin{matrix} 13'90 \\ 5'12 \end{matrix}$$

Aquest resultat ens indica que per a un individu que ha obtingut 20 punts a X (test verbal) el més probable (Y') és que obtingui 9,51 a Y, però en el 95% dels casos (amb un nivell de confiança del 95% o un risc d'error del 5%) la seva puntuació a Y oscil·larà entre un màxim de 13'90 i un mínim de 5'12.

De la mateixa manera es podria estudiar la regressió d'X en funció d'Y, utilitzant l'equació de la recta X=f(y) per a calcular el valor més probable d'X (X') per a un valor determinat d'Y, i després calcular l'interval de regressió al voltant d'X'.

1.3 Altres coeficients de correlació simple

1.3.1 Spearman

Si volem estudiar la relació entre variables mesurades a nivell ordinal es pot utilizar la correlació d'Spearman. La interpretació d'aquest coeficient és igual que en el cas del coeficient de Pearson i el seu càlcul en l'SPSS també es fa a través del menú principal Analizar > Correlaciones > Bivariadas... > i seleccionant el coeficient de Spearman.

1.3.2 Chi-quadrat

Es pot aplicar amb qualsevol tipus de variables (qualitatives). No té cap mena de restriccions (no cal una n mínima, no cal que les distribucions s'ajustin a la corba normal...).

[2] El valor d'S_{yx} apareix en l'output de l'SPSS (Resumen del modelo) com a "Error típ. de la estimación" (en aquest cas dóna 2'505), tot i que al calcular-ho manualment dóna 2'24 ja que s'utilitza la quasidesviació típica.

Cal tenir en compte però, a diferència que els coeficients de Pearson i d'Spearman, que la chi-quadrat indica si hi ha o no correlació però no el sentit ni la intensitat de la correlació.

Coeficients que indiquen la intensitat de la correlació a partir de chi-quadrat:

$$2\times 2 \longrightarrow \Phi = \sqrt{\frac{\chi^2}{n}}$$

$$k\times l \longrightarrow C = \sqrt{\frac{\chi^2}{\chi^2 + n}}$$

El càlcul de chi-quadrat té una condició d'aplicació que consisteix en que com a màxim un 20% de les freqüències teòriques (o esperades) pot ser inferior a 5. Si aquesta condició no es compleix aleshores s'ha de reagrupar alguna categoria i tornar a fer el càlcul de la chi-quadrat des del principi[3].

Exemples

Exemple a) 120 alumnes de Sociologia es distribueixen de la següent forma: 39 d'origen basc y la resta, d'altres províncies. Es pretén estudiar si l'origen de l'alumnat està relacionat amb la concepció sobre la Sociologia, sabent que 64 alumnes tenen una concepció concreta i 56 universal. Amb una $\alpha = 0'01$.

Tabla de contingencia Concepte sociologia * Origen

			Origen		
			Basc	No Basc	Total
Concepte sociologia	Concret	Recuento	29	35	64
		Frecuencia esperada	20,8	43,2	64,0
	Universal	Recuento	10	46	56
		Frecuencia esperada	18,2	37,8	56,0
Total		Recuento	39	81	120
		Frecuencia esperada	39,0	81,0	120,0

[3] Si la taula de contingència és de 2x2 i hi ha alguna freqüència esperada inferior a 5, aleshores com que ja no es pot reagrupar cap categoria, en comptes de consultar la "Chi-cuadrado de Pearson" que apareix a l'output s'ha de consultar la "Corrección por continuidad".

Pruebas de chi-cuadrado

	Valor	gl	Sig. asintótica (bilateral)	Sig. exacta (bilateral)	Sig. exacta (unilateral)
Chi-cuadrado de Pearson	10,263[b]	1	,001		
Corrección por continuidad[a]	9,049	1	,003		
Razón de verosimilitud	10,627	1	,001		
Estadístico exacto de Fisher				,002	,001
Asociación lineal por lineal	10,177	1	,001		
N de casos válidos	120				

a. Calculado sólo para una tabla de 2x2.

b. 0 casillas (,0%) tienen una frecuencia esperada inferior a 5. La frecuencia mínima esperada es 18,20.

Medidas simétricas

		Valor	Sig. aproximada
Nominal por nominal	Phi	,292	,001
	V de Cramer	,292	,001
N de casos válidos		120	

a. Asumiendo la hipótesis alternativa.

b. Empleando el error típico asintótico basado en la hipótesis nula.

En primer lloc es comprova que el percentatge de freqüències esperades (o teòriques) que hi ha amb valors inferiors a 5: en aquest exemple hi ha un 0%. Per tant es compleix la condició d'aplicació.

El valor de Chi-quadrat és 10,263 i la seva significació de 0,001; com que aquesta significació és menor que l'alpha 0,01 podem concloure que l'origen de l'alumnat té relació significativa amb la concepció de la sociologia. Ara bé, si volem saber quan intensa és aquesta relació, s'ha de calcular el coeficient Phi, que en l'exemple dóna 0,292.

Exemple b) Un grup de 80 alumnes que ha fet dos exàmens parcials de l'assignatura d'estadística ha obtingut les següents qualificacions. Hi ha correlació entre els dos parcials ($\alpha = 0,05$)?

Tabla de contingencia PARCIAL1 * PARCIAL2

			PARCIAL2		Total
			aprovat	suspens	
PARCIAL1	aprovat	Recuento	40	20	60
		Frecuencia esperada	37,5	22,5	60,0
	suspens	Recuento	10	10	20
		Frecuencia esperada	12,5	7,5	20,0
Total		Recuento	50	30	80
		Frecuencia esperada	50,0	30,0	80,0

Pruebas de chi-cuadrado

	Valor	gl	Sig. asintótica (bilateral)	Sig. exacta (bilateral)	Sig. exacta (unilateral)
Chi-cuadrado de Pearson	1,778[b]	1	,182		
Corrección por continuidad[a]	1,138	1	,286		
Razón de verosimilitud	1,743	1	,187		
Estadístico exacto de Fisher				,195	,143
Asociación lineal por lineal	1,756	1	,185		
N de casos válidos	80				

a. Calculado sólo para una tabla de 2x2.

b. 0 casillas (,0%) tienen una frecuencia esperada inferior a 5. La frecuencia mínima esperada es 7,50.

Medidas simétricas

		Valor	Sig. aproximada
Nominal por nominal	Phi	,149	,182
	V de Cramer	,149	,182
N de casos válidos		80	

a. Asumiendo la hipótesis alternativa.

b. Empleando el error típico asintótico basado en la hipótesis nula.

Es compleix la condició d'aplicació pel que fa a les freqüències esperades.

La chi-quadrat dóna 1,778 i la seva significació és 0,182. Donat que la significació és més gran que l'alpha (0,05) es pot concloure que no hi ha relació entre els dos parcials.

La intensitat de la relació trobada (Phi) és de 0,149 (no significativa).

Exemple c) Es vol estudiar la relació entre el nivell instructiu-educatiu i els ingressos de 400 pares ($\alpha = 0{,}01$).

Tabla de contingencia Nivell d'ingressos * Nivell instrucctiu

			Nivell instrucctiu			Total
			baix	mig	alt	
Nivell d'ingressos	baix	Recuento	63	78	29	170
		Frecuencia esperada	51,0	76,5	42,5	170,0
	mig	Recuento	48	82	45	175
		Frecuencia esperada	52,5	78,8	43,8	175,0
	alt	Recuento	9	20	26	55
		Frecuencia esperada	16,5	24,8	13,8	55,0
Total		Recuento	120	180	100	400
		Frecuencia esperada	120,0	180,0	100,0	400,0

Pruebas de chi-cuadrado

	Valor	gl	Sig. asintótica (bilateral)
Chi-cuadrado de Pearson	22,931[a]	4	,000
Razón de verosimilitud	21,744	4	,000
Asociación lineal por lineal	19,436	1	,000
N de casos válidos	400		

a. 0 casillas (,0%) tienen una frecuencia esperada inferior a 5. La frecuencia mínima esperada es 13,75.

Medidas simétricas

		Valor	Sig. aproximada
Nominal por nominal	Coeficiente de contingencia	,233	,000
N de casos válidos		400	

a. Asumiendo la hipótesis alternativa.

b. Empleando el error típico asintótico basado en la hipótesis nula.

Es comprova que no hi ha més d'un 20% de freqüències esperades inferiors a 5. El valor de chi-quadrat és 22'931 amb una significació de 0'000, la qual en ser menor que 0,01 permet concloure que hi ha una relació significativa entre el nivell instructiu i el nivell econòmic dels subjectes estudiats. La intensitat d'aquesta relació és de 0'233.

1.4 Correlació i regressió parcial i múltiple

En la recerca educativa, donada la complexitat dels fenòmens que s'estudien, és molt probable que es considerin més de dues variables al mateix temps per tal d'intentar explicar els problemes.

Suposem, per exemple, que en una mostra de 200 alumnes hem mesurat les variables intel·ligència general (G), alçada (A) i edat (E). Hem calculat els estadístics de les tres variables i els coeficients de correlació lineal simple entre aquestes.

Aquestes dades estan recollides en la taula següent:

	G (x_1)	A (x_2)	E (x_3)	$\overline{X}$	S_x
G (x_1)	(r_{11}) 1	(r_{12}) .6	(r_{13}) .8	100	20
A (x_2)	(r_{21}) .6	(r_{22}) 1	(r_{23}) .7	150	10
E (x_3)	(r_{31}) .8	(r_{32}) .7	(r_{33}) 1	12	2

Les correlacions simples trobades en realitat no són exactament així, perquè en qualsevol cas cadascuna de les variables que formen una parella està relacionada també amb l'altra. Les tres estan correlacionades.

La ***correlació parcial*** permet estudiar la relació entre dues variables, eliminant la influència de les restants. En general, en un conjunt d' *m* variables, la relació entre dues d'elles, eliminant la influència de les restants. També es podria aconseguir el mateix amb un disseny en què es controlés el possible efecte de les variables "restants", és a dir, les que no es plantegen ni com a independents ni com a dependents.

La ***correlació múltiple*** permet estudiar la relació entre una de les variables i les altres. En general, en un conjunt d' *m* variables, la relació d'una d'elles amb totes les altres.

1.4.1 Correlació parcial

$$r_{12.3} = \frac{r_{12} - r_{13} \cdot r_{23}}{\sqrt{(1 - r_{13}).(1 - r_{23})}}$$

Estudi de la relació entre les variables 1 y 2, eliminant la influència de la variable 3:

Estudi de la relació entre les variables 2 y 3, eliminant la influència de la variable 1:

$$r_{23.1} = \frac{r_{23} - r_{21} \cdot r_{31}}{\sqrt{(1 - r_{21}).(1 - r_{31})}}$$

Estudi de la relació entre les variables 3 y 1, eliminant la influència de la variable 2:

$$r_{31.2} = \frac{r_{31} - r_{32} \cdot r_{12}}{\sqrt{(1 - r_{32}).(1 - r_{12})}}$$

Amb les dades del cas anterior:

$$r_{12.3} = \frac{0,6 - 0,8\,.0,7}{\sqrt{(1 - 0,8^2).(1 - 0,7^2)}} = 0,09 \qquad r_{23.1} = \frac{0,7 - 0,6\,.0,8}{\sqrt{(1 - 0,6^2).(1 - 0,8^2)}} = 0,46$$

$$r_{31.2} = \frac{0,8 - 0,7\,.0,6}{\sqrt{(1 - 0,7^2).(1 - 0,6^2)}} = 0,67$$

La significació es pot comprovar calculant la Z de cada coeficient, segons la fórmula:

$$Z_{r_{xy.z}} = r_{xy.z}\sqrt{n - m}$$

i comparant cada Z amb $Z_{\alpha/2}$. En aquest cas $Z_{\alpha/2}$ = 1'96 i per tant:

- $Z_{r12.3}$ = 1'26 < 1'96 Res s'oposa a acceptar H_0 (corr. no significativa)
- $Z_{r23.1}$ = 6'46 > 1'96 Rebutgem H_0 amb un risc de 0'05 (corr. significativa)
- $Z_{r31.2}$ = 9'41 > 1'96 Rebutgem H_0 amb un risc de 0'05 (corr. significativa)

Es pot observar com, en eliminar l'efecte de l'Edat (que és la que de veritat correlaciona amb la intel·ligència en períodes evolutius), la correlació entre l'Alçada i la Intel·ligència general queda reduïda a nivells no significatius, mentre que les altres dues (Intel·ligència-Edat i Edat-Alçada) es mantenen.

Limitacions: Només es pot utilitzar quan les variables de les quals s'elimina la influència siguin molt precises i objectives (alçada, pes, edat, pressió sanguínia... en general variables biomètriques o fisiològiques). En educació, com en ciències humanes i socials, la majoria de les variables no compleixen aquesta condició, per la qual cosa la correlació parcial té poca aplicació.

1.4.2 Correlació múltiple

X_0 variable dependent (en el nostre cas Intel·ligència general)

X_1 variable independent (Alçada)

X_2 variable independent (Edat)

$$R^2_{0.12} = \beta_{01} r_{01} + \beta_{02} r_{02}$$

$$\beta_{01} = \frac{r_{01} - r_{02} r_{12}}{1 - r^2_{12}} ; \beta_{02} = \frac{r_{02} - r_{01} r_{12}}{1 - r^2_{12}}$$

$$\beta_{01} = \frac{0'6 - 0'8 \cdot 0'7}{1 - 0'7^2} = \frac{0'04}{0'51} = 0'08 ; \beta_{02} = \frac{0'8 - 0'6 \cdot 0'7}{1 - 0'7^2} = \frac{0'38}{0'51} = 0'75$$

$$R^2_{0.12} = 0'08 \cdot 0'6 + 0'75 \cdot 0'8 = 0'648 ; R_{0.12} = \sqrt{0'648} = 0'805$$

Regressió múltiple

$$X'_0 = \overline{X_0} + b_{01}(X_1 - \overline{X_1}) + b_{02}(X_2 - \overline{X_2})$$

$$b_{01} = \beta_{01} \frac{S_0}{S_1} ; b_{02} = \beta_{02} \frac{S_0}{S_2}$$

$$X'_0 = 100 + 0'08\frac{20}{10}(X_1 - 150) + 0'75\frac{20}{2}(X_2 - 12)$$

si X_1=155 i X_2=15

$$X'_0 = 100 + 0'08\frac{20}{10}(155 - 150) + 0'75\frac{20}{2}(15 - 12) = 100 + 0'8 + 22'5 = 123'3$$

interval:

$$X_0 \in X'_0 \pm Z_{\alpha/2} S_{0.12}$$

$$S_{0.12} = S_0\sqrt{1 - r_{0.12}^2} = 20\sqrt{1 - 0'648} = 11'87$$

$$X_0 \in 123'3 \pm 1'96 \cdot 11'87 = 123'3 \pm 23'27 \prec \begin{matrix} 146'57 \\ 100'03 \end{matrix}$$