Análisis de Regresión

http://www.academia.utp.ac.pa/humberto-alvarez/diseno-de-experimentos-y-regresion

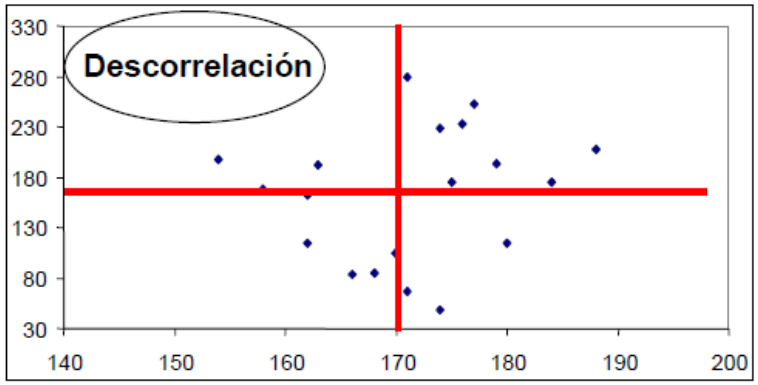
Introducción

- Tiene como objetivo modelar en forma matemática el comportamiento de una variable de respuesta en función de una o más variables independientes (factores).
- Para estimar los parámetros de un *modelo de regresión* son necesarios los datos.
- Estos pueden obtenerse de experimentos planeados, de observaciones de fenómenos no controlados o de registros históricos.

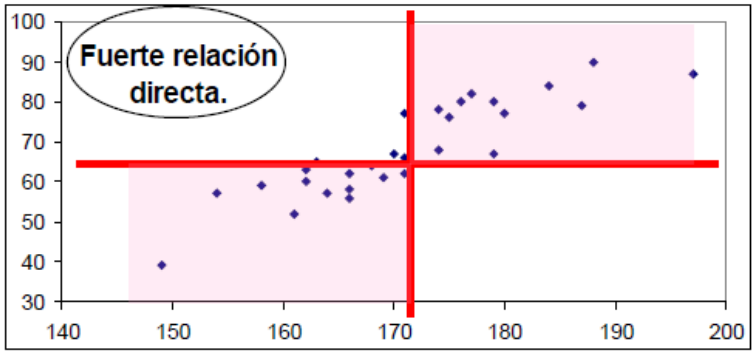
Relaciones entre variables aleatorias y regresión lineal.

- El término regresión fue introducido por Galton en su libro "*Natural inheritance*" (1889) refiriéndose a la "ley de la regresión universal":
  - "Cada peculiaridad en un hombre es compartida por sus descendientes, pero en media, en un grado menor."
    - Regresión a la media
  - Su trabajo se centraba en la descripción de los rasgos físicos de los descendientes (una variable) a partir de los de sus padres (otra variable).
  - Pearson (un amigo suyo) realizó un estudio con más de 1000 registros de grupos familiares observando una relación del tipo:
    - Altura del hijo = 85cm + **0,5** • altura del padre (aprox.)
    - Conclusión: los padres muy altos tienen tendencia a tener hijos que heredan parte de esta altura, aunque tienen tendencia a acercarse (*regresar*) a la media. Lo mismo puede decirse de los padres muy bajos.
- Hoy en día el sentido de regresión es el de predicción de una medida basándonos en el conocimiento de otra.

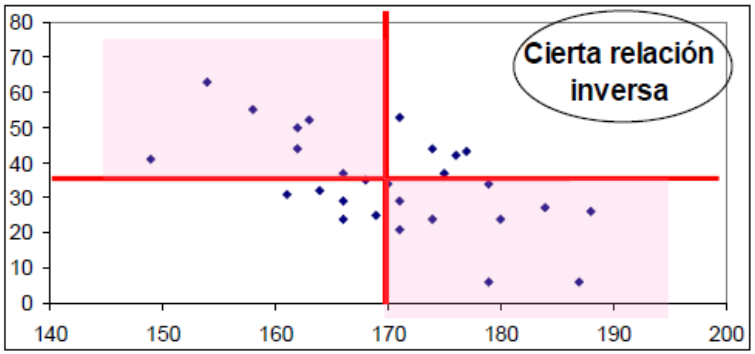
Reconociendo relaciones



Para valores de X por encima de la media tenemos valores de Y por encima y por debajo en proporciones similares: Descorrelación.

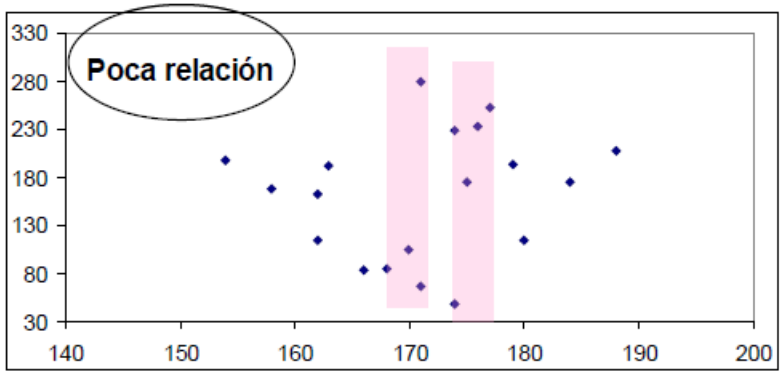


- Para los valores de X mayores que la media le corresponden valores de Y mayores también.
- Para los valores de X menores que la media le corresponden valores de Y menores también.
- Esto se llama relación directa o creciente entre X e Y.

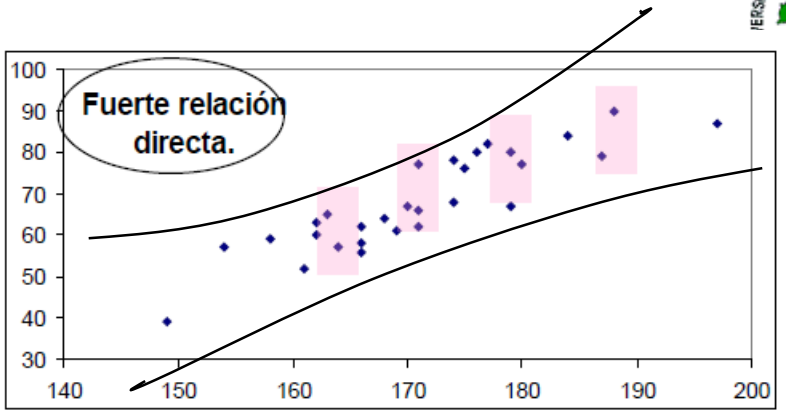


Para los valores de X mayores que la media le corresponden valores de Y menores. Esto es relación inversa o decreciente.

Buena o mala relación

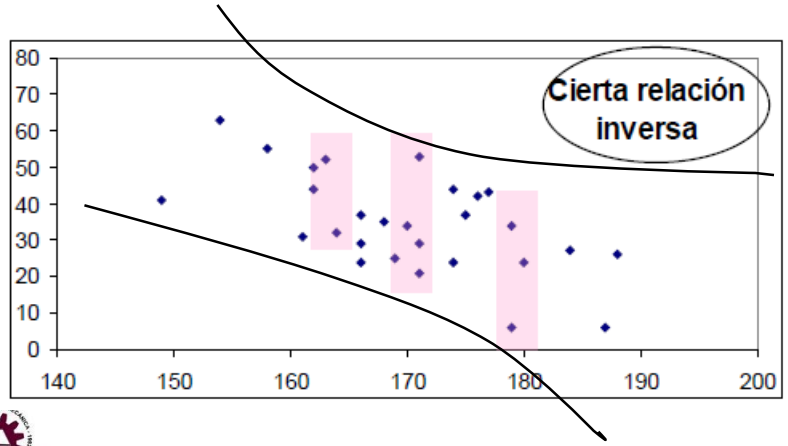


Dado un valor de X no podemos decir gran cosa sobre Y. Mala relación. Independencia.



- Conocido X sabemos que Y se mueve por una horquilla estrecha. Buena relación.

- Lo de "horquilla estrecha" hay que entenderlo con respecto a la dispersión que tiene la variable Y por si sola, cuando no se considera X.



Covarianza

- La covarianza entre dos variables, $\mathbf{S_{xy}}$, indica si la posible relación entre dos variables es directa o inversa:
- Directa: $S_{xy} > 0$
- Inversa: $S_{xy} < 0$
- Descorreladas: $S_{xy} = 0$
- El signo de la covarianza da una idea de la nube de datos, pero no dice nada del grado de relación.

$$S_{xy} = \frac{1}{n}\sum_i (x_i - \bar{x})(y_i - \bar{y})$$

Coeficiente de correlación de Pearson (r)

$$r = \frac{S_{xy}}{S_x S_y}$$

- Indica si los datos tienen una tendencia a disponerse alineadamente (excluyendo rectas horizontales y verticales).
- Tiene el mismo signo que S_{xy} . Por tanto de su signo obtenemos el que la posible relación sea directa o inversa.
- Es útil para determinar si hay relación lineal entre dos variables, pero no servirá para otro tipo de relaciones (cuadrática, logarítmica,...)
- Sólo toma valores en [-1,1].
- Cuanto más cerca esté r de +1 o -1 mejor será el grado de relación lineal.

Regresión lineal simple

- Sean dos variables *X* y *Y*, suponga que se quiere explicar el comportamiento de *Y.*
- *Y* se le llama la *variable dependiente* o la *variable de respuesta* y a *X* se le conoce como *variable independiente* o *variable regresora*.
- La variable *X* no necesariamente es aleatoria, ya que en muchas ocasiones el investigador fija sus valores. *Y* siempre es una variable aleatoria.
- Una manera de estudiar el comportamiento de *Y* con respecto a *X* es mediante un modelo de regresión que consiste en ajustar un modelo matemático de la forma: $Y = f(X)$ a la pareja de puntos, con base en los valores que toma *X*.

El modelo de regresión:

- Suponga que las variables X y Y están relacionadas linealmente y que para cada valor de X, la variable dependiente, Y, es una variable aleatoria.
- Es decir, que cada observación de Y puede ser descrita por el modelo:

$$y_i = \beta_0 + \beta_1 x_i + \varepsilon_i$$

- donde ε es el error aleatorio normalmente distribuido con media cero y varianza σ^2. También suponga que los errores aleatorios no están correlacionados.
- El modelo se conoce como regresión lineal simple.

Cálculo de los parámetros:

- Mediante mínimos cuadrados:

$$S = \sum_{i=1}^{n} (\varepsilon_i)^2 = \sum_{i=1}^{n} \left(y_i - [\beta_0 + \beta_1 x_i]\right)^2$$

- Derivando con respecto β_0 y β_1

$$\frac{\partial S}{\partial \beta_0} = -\sum_{i=1}^{n} 2\left(y_i - [\beta_0 + \beta_1 x_i]\right)$$

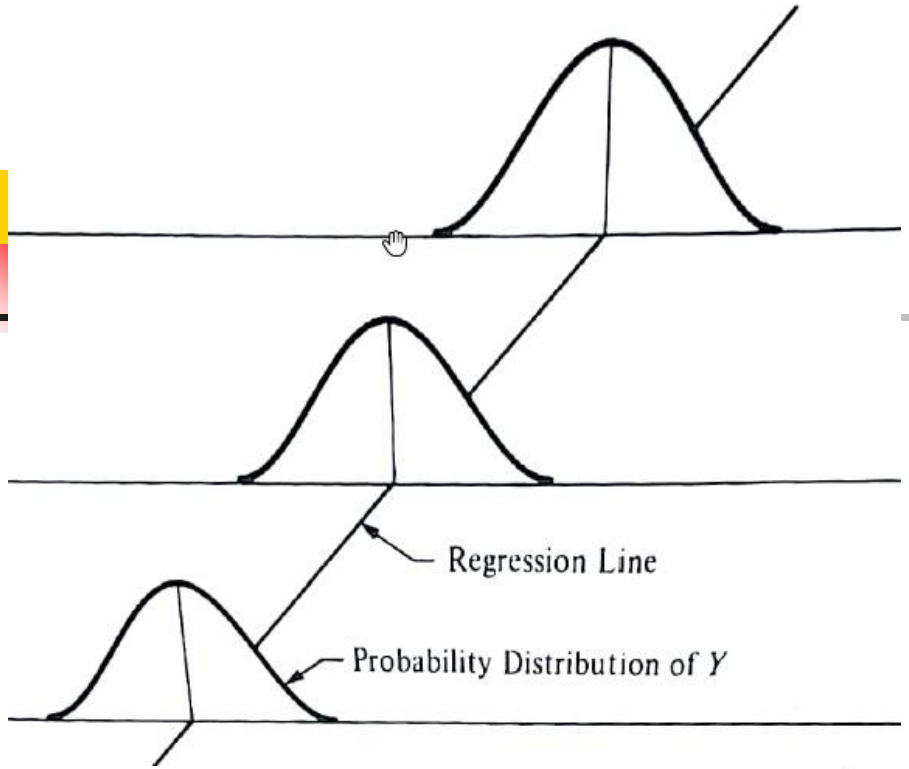
$$\frac{\partial S}{\partial \beta_1} = -\sum_{i=1}^{n} 2x_i\left(y_i - [\beta_0 + \beta_1 x_i]\right)$$

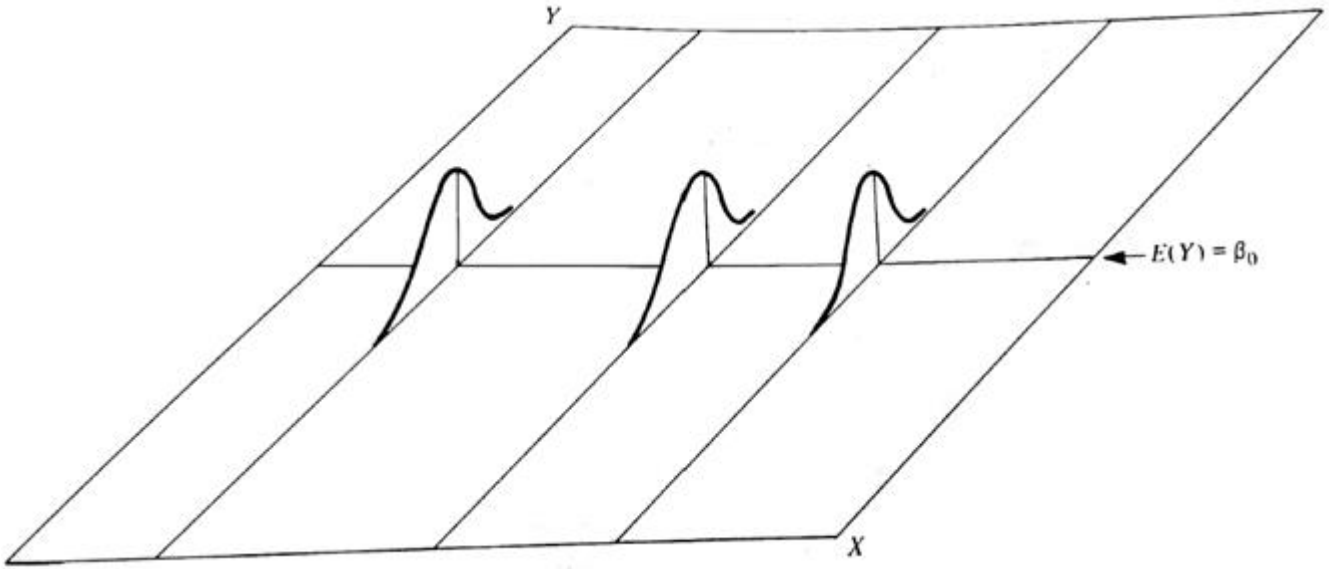
- Igualando a cero y resolviendo para β_0 y β_1

$$\hat{\beta}_1 = \frac{S_{xy}}{S_{xx}} \qquad S_{xy} = \sum_{i=1}^{n} (x_i - \bar{x})(y_i - \bar{y}) = \sum_{i=1}^{n} x_i y_i - \frac{(\Sigma_{i=1}^{n} x_i)(\Sigma_{i=1}^{n} y_i)}{n}$$

$$\hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x} \qquad S_{xx} = \sum_{i=1}^{n} (x_i - \bar{x})^2 = \sum_{i=1}^{n} x_i^2 - \frac{(\Sigma_{i=1}^{n} x_i)^2}{n}$$

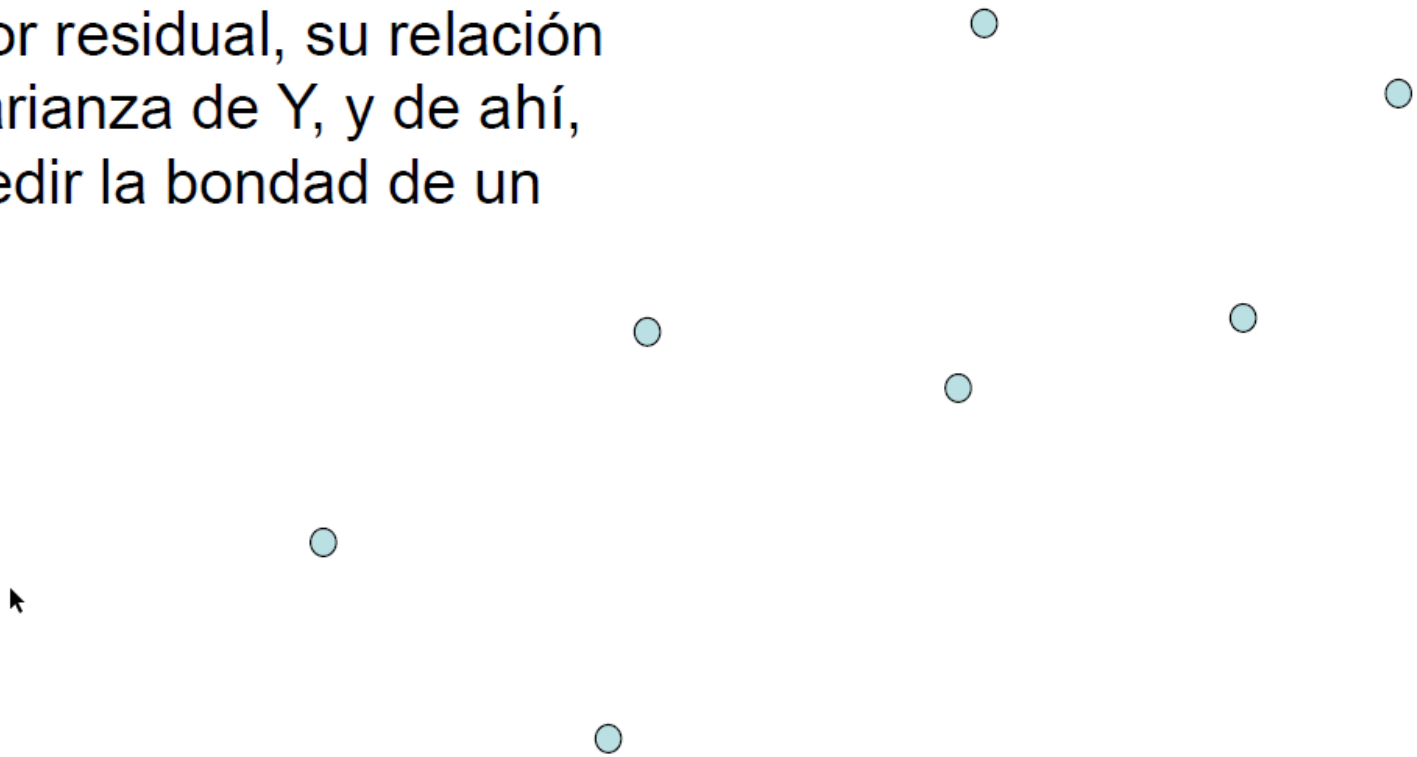
Comportamiento de la línea de regresión





¿Qué tan buena es la regresión?

Imaginemos un diagrama de dispersión, y vamos a tratar de comprender en primer lugar qué es el error residual, su relación con la varianza de Y, y de ahí, cómo medir la bondad de un ajuste.

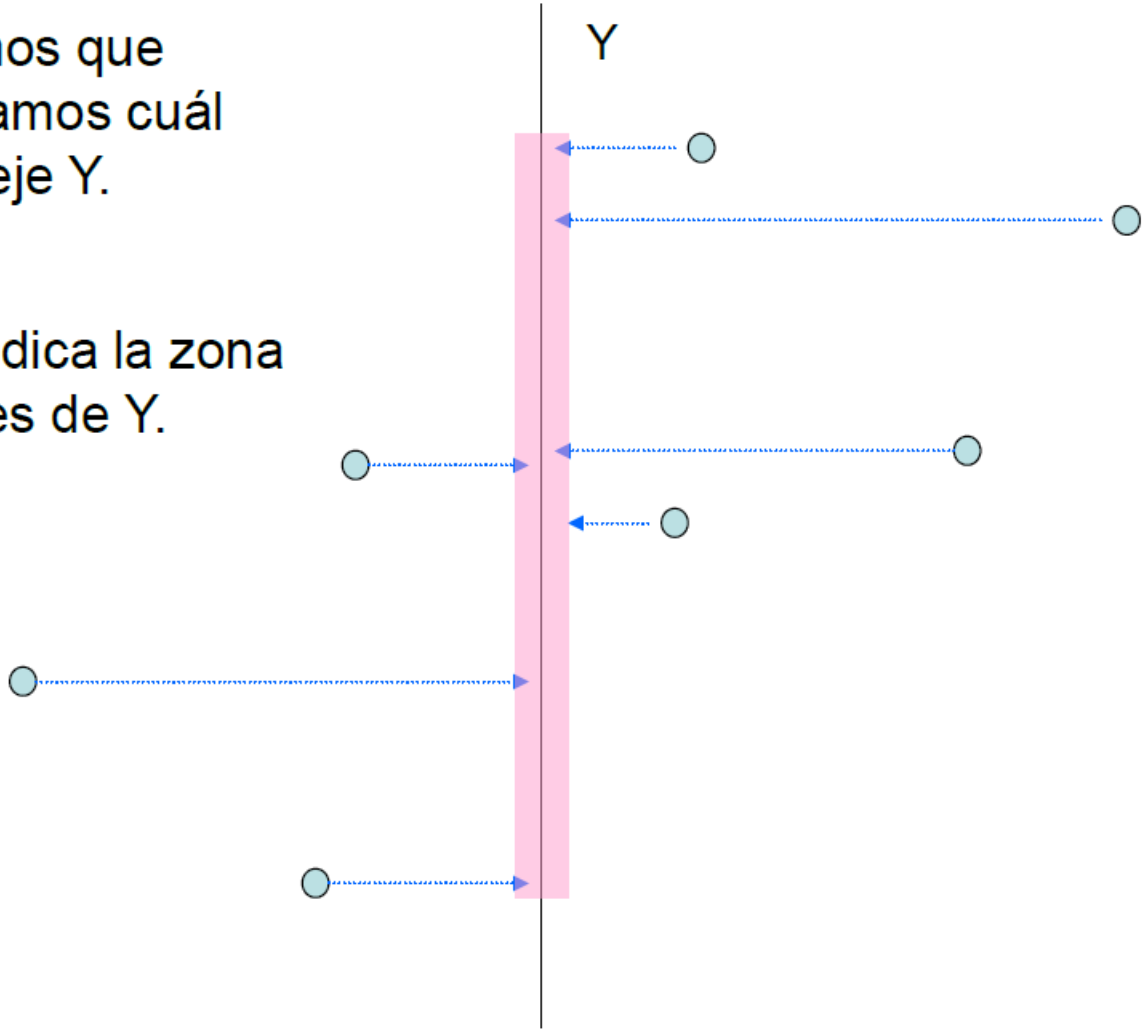


Interpretación de la variabilidad de Y

En primer lugar olvidemos que existe la variable X. Veamos cuál es la variabilidad en el eje Y.

La franja sombreada indica la zona donde varían los valores de Y.

Proyección sobre el eje Y = olvidar X.

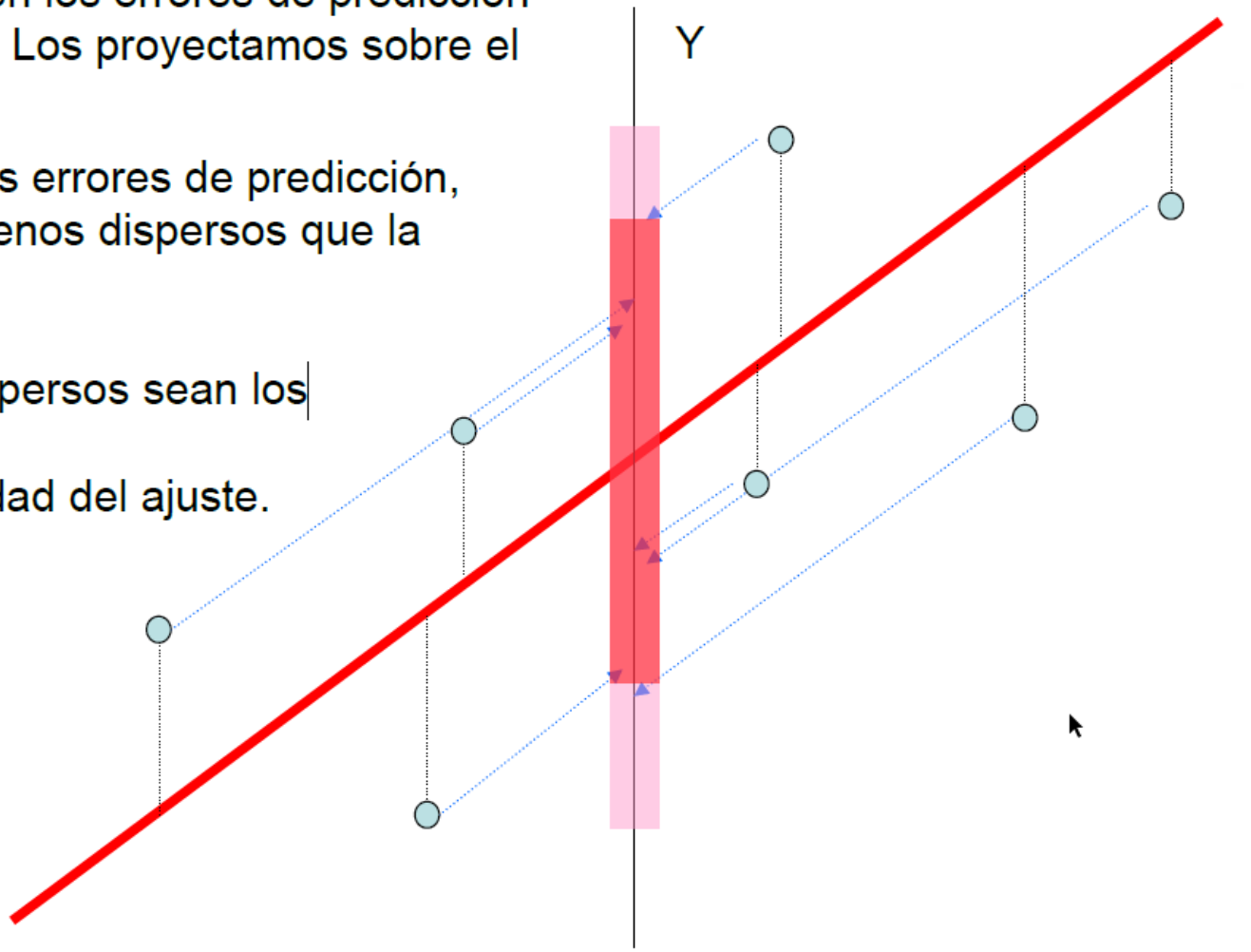


El residuo

Fijémonos ahora en los errores de predicción (líneas verticales). Los proyectamos sobre el eje Y.

Se observa que los errores de predicción, residuos, están menos dispersos que la variable Y original.

Cuanto menos dispersos sean los residuos,
mejor será la bondad del ajuste.



Bondad de ajuste

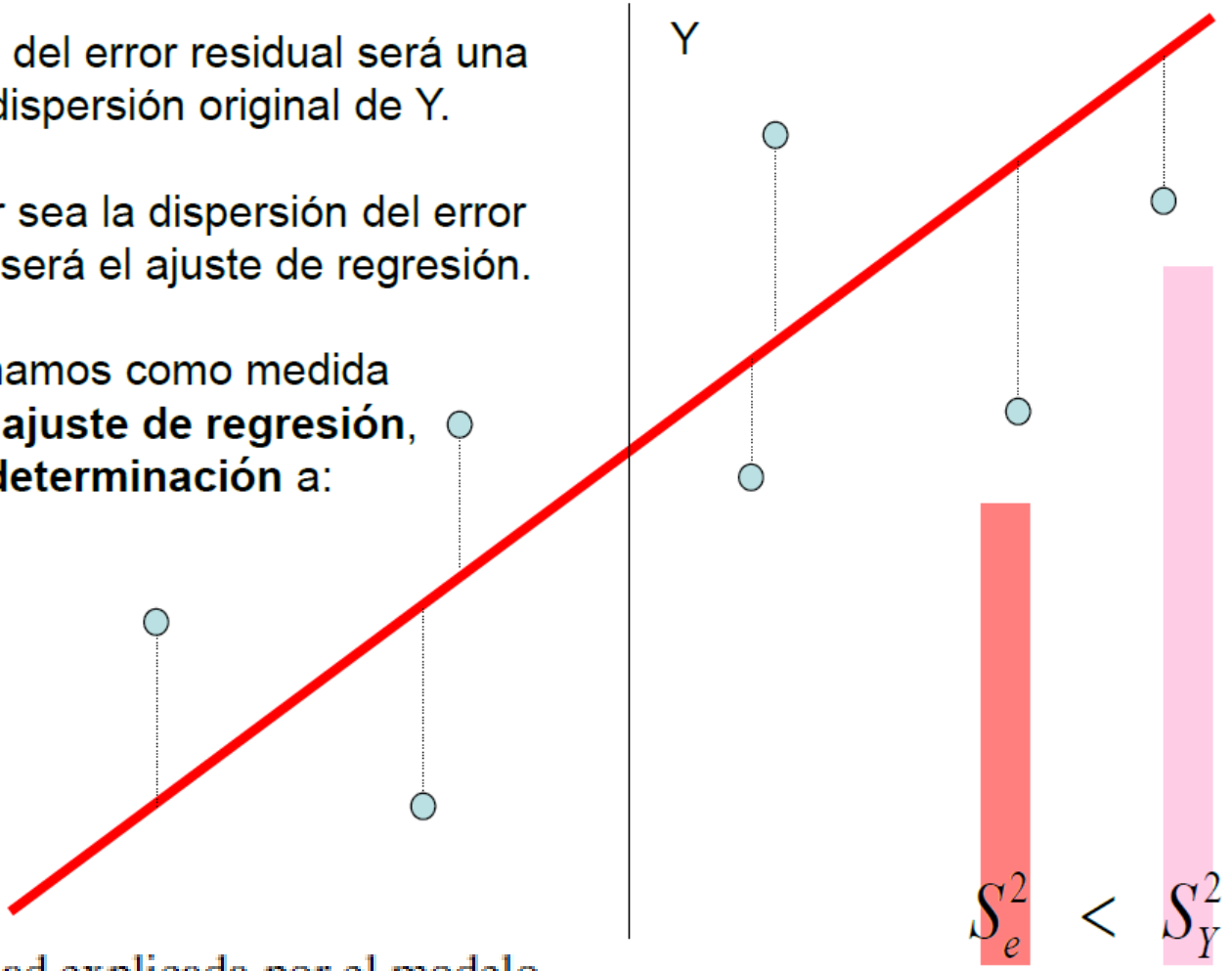
Resumiendo:

- La dispersión del error residual será una fracción de la dispersión original de Y.

- Cuanto menor sea la dispersión del error residual mejor será el ajuste de regresión.

Eso hace que definamos como medida de **bondad de un ajuste de regresión**, o **coeficiente de determinación** a:

$$R^2 = 1 - \frac{S_e^2}{S_y^2}$$



$$R^2 = \frac{\text{Variabilidad explicada por el modelo}}{\text{Variabilidad total}}$$

Prueba de hipótesis del modelo de regresión

- Una forma de probar que tan bien se explica la relación entre *X* y *Y.*
- Una forma de hacer esto es probar una serie de hipótesis sobre el modelo.
- Para ello se supone una distribución normal para el término de error, ε_i, con media cero y varianza σ^2.
- La hipótesis de mayor interés plantea que la pendiente es significativamente
- diferente de cero.
- Esto se logra al probar la siguiente hipótesis:

$$H_0 \quad : \quad \beta_1 = 0$$

$$H_A \quad : \quad \beta_1 \neq 0$$

Análisis de regresión para el modelo

Parámetro	Estimación	Error estándar	Estadístico	Valor-*p*
Intercepción	$\hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x}$	$\sqrt{CM_E\left[\frac{1}{n} + \frac{\bar{x}^2}{S_{xx}}\right]}$	$\frac{\hat{\beta}_0}{\sqrt{CM_E\left[\frac{1}{n} + \frac{\bar{x}^2}{S_{xx}}\right]}}$	$\Pr\left(T > \lvert t_0 \rvert\right)$
Pendiente	$\hat{\beta}_1 = \frac{S_{xy}}{S_{xx}}$	$\sqrt{CM_E / S_{xx}}$	$\frac{\hat{\beta}_1}{\sqrt{CM_E / S_{xx}}}$	$\Pr\left(T > \lvert t_0 \rvert\right)$

Intervalo de confianza para los parámetros de la regresión

$$\hat{\beta}_1 \pm t_{(\alpha/2, n-2)} \sqrt{CM_E / S_{xx}} \qquad \hat{\beta}_0 \pm t_{(\alpha/2, n-2)} \sqrt{CM_E \left[\frac{1}{n} + \frac{\bar{x}^2}{S_{xx}}\right]}$$

ANOVA del modelo de regresión

- Se descompone la variabilidad observada, y a partir de ello se prueba la hipótesis.

Fuente de variación	Suma de cuadrados	Grados de libertad	Cuadrado medio	F_0	Valor-p
Regresión	$SC_R = \hat{\beta}_1 S_{xy}$	1	CM_R	CM_R/CM_E	$\Pr(F > F_0)$
Error o residual	$SC_E = S_{yy} - \hat{\beta}_1 S_{xy}$	$n-2$	CM_E		
Total	S_{yy}	$n-1$			

Ejemplo:

- Se quiere investigar la forma en que se relaciona la cantidad de fibra (madera) en la pulpa con la resistencia del producto (papel).
- Los datos obtenidos en un estudio experimental se muestran en la tabla

Porcentaje de fibra	Resistencia
4	134
6	145
8	142
10	149
12	144
14	160
16	156
18	157
20	168
22	166
24	167
26	171
28	174
30	183

Análisis de regresión lineal

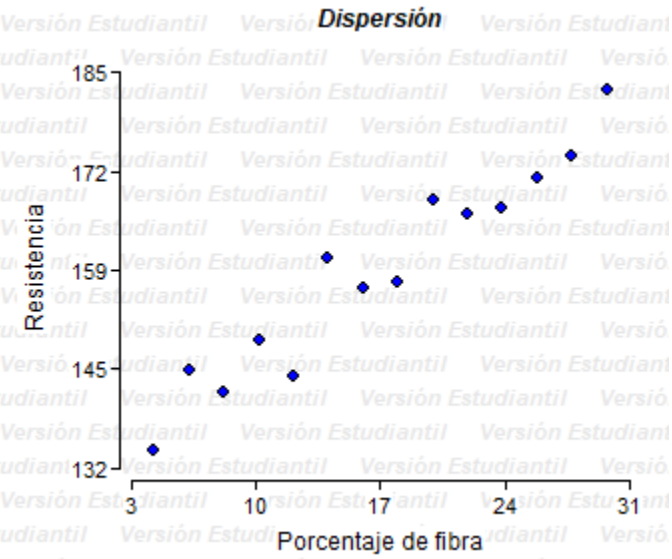
Variable	N	R^2	R^2 Aj	ECMP	AIC	BIC
Resistencia	14	0.93	0.92	20.22	81.51	83.43

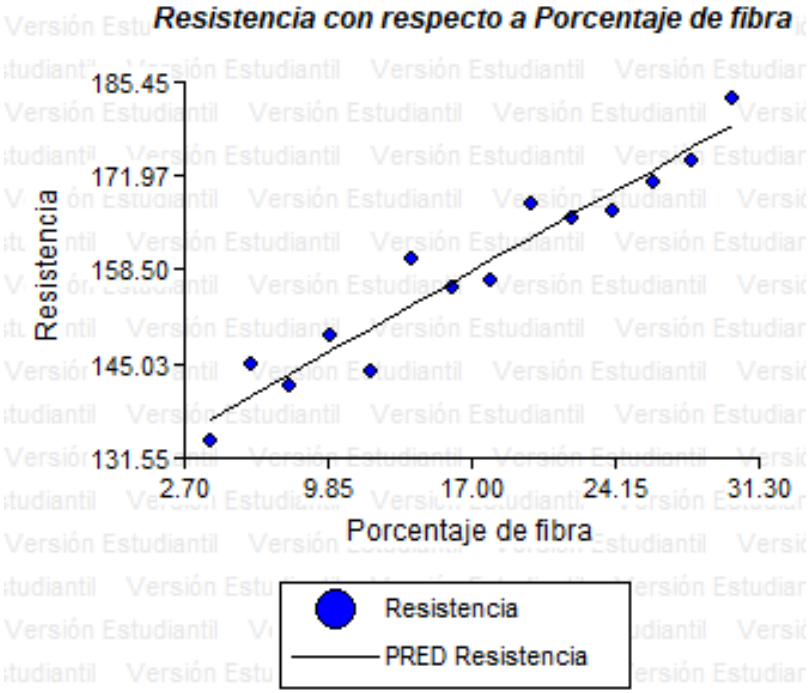
Coeficientes de regresión y estadísticos asociados

Coef	Est.	E.E.	LI(95%)	LS(95%)	T	p-valor	CpMallows	VIF
const	130.67	2.42	125.41	135.94	54.05	<0.0001		
Porcentaje de fibra	1.62	0.13	1.34	1.90	12.64	<0.0001	159.75	1.00

Cuadro de Análisis de la Varianza (SC tipo III)

F.V.	SC	gl	CM	F	p-valor
Modelo.	2400.53	1	2400.53	159.75	<0.0001
Porcentaje de fibra	2400.53	1	2400.53	159.75	<0.0001
Error	180.33	12	15.03		
Total	2580.86	13			





SUMMARY OUTPUT

Regression Statistics	
Multiple R	0.964432318
R Square	0.930129695
Adjusted R Square	0.92430717
Standard Error	3.876481166
Observations	14

ANOVA

	df	*SS*	*MS*	*F*	*gnificance F*
Regression	1	2400.53	2400.53	159.747	2.7E-08
Residual	12	180.325	15.0271		
Total	13	2580.86			

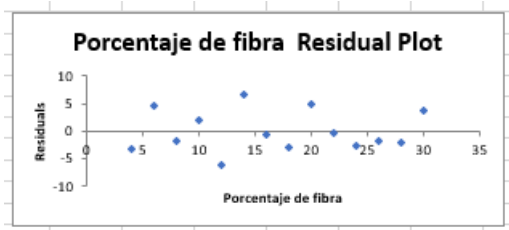
	Coefficients	*ɔndard Err*	*t Stat*	*P-value*	*.ower 95%*	*Jpper 95%*	*ower 95.0%*	*pper 95.0%*
Intercept	130.6747253	2.41779	54.0472	1.1E-15	125.407	135.943	125.407	135.943
Porcentaje de fibra	1.624175824	0.1285	12.6391	2.7E-08	1.34419	1.90416	1.34419	1.90416

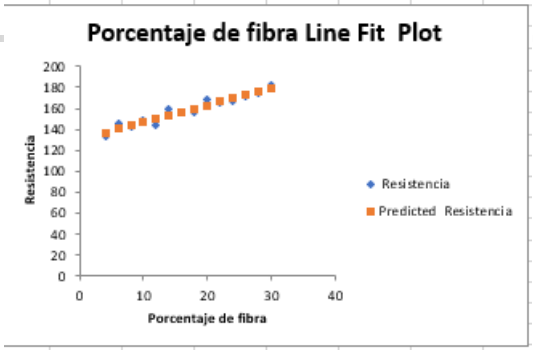
RESIDUAL OUTPUT

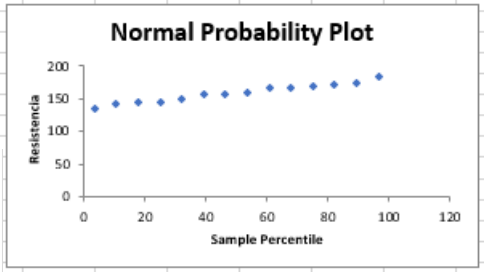
Observation	*edicted Resistenc*	*Residuals*
1	137.1714286	-3.17143
2	140.4197802	4.58022
3	143.6681319	-1.66813
4	146.9164835	2.08352
5	150.1648352	-6.16484
6	153.4131868	6.58681
7	156.6615385	-0.66154
8	159.9098901	-2.90989
9	163.1582418	4.84176
10	166.4065934	-0.40659
11	169.6549451	-2.65495
12	172.9032967	-1.9033
13	176.1516484	-2.15165
14	179.4	3.6

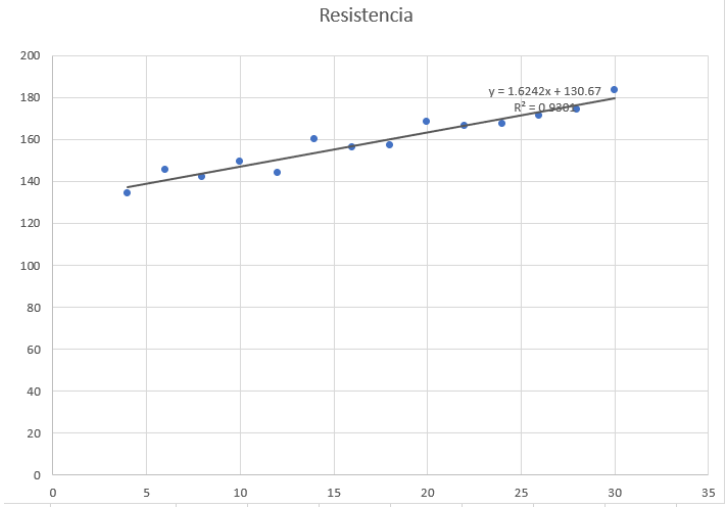
PROBABILITY OUTPUT

Percentile	*Resistencia*
3.57143	134
10.7143	142
17.8571	144
25	145
32.1429	149
39.2857	156
46.4286	157
53.5714	160
60.7143	166
67.8571	167
75	168
82.1429	171
89.2857	174
96.4286	183









Regresión lineal múltiple

- En muchas situaciones prácticas existen varias variables independientes que se cree que influyen o están relacionadas con una variable de respuesta *Y*, y por lo tanto será necesario tomar en cuenta si se quiere predecir o entender mejor el comportamiento de *Y*.
- Sea X_1, X_2, …, X_k variables independientes o regresoras, y sea *Y* una variable de respuesta, entonces el *modelo de regresión lineal múltiple* con *k* variables independientes es el polinomio de primer orden:

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_k X_k + \varepsilon$$

- Donde los β_j son los parámetros del modelo que se conocen como *coeficientes de regresión* y ε es el error aleatorio, con media cero, $E(\varepsilon) = 0$ y $V(\varepsilon) = \sigma^2$.
- Así:
  - β_0 es el término independiente. Es el valor esperado de Y cuando $X_1, \ldots, X_k$ son cero.
  - $\beta_1, \beta_2, \ldots \beta_k$ son los coeficientes parciales de la regresión tal que β_1 mide el cambio en Y por cada cambio unitario en X_1, manteniendo $X_2, X_3, \ldots, X_k$ constantes y así sucesivamente.
  - ε es el error de observación debido a variables no controladas.

Estructura de datos

Y	X_1	X_2	…	X_k
y_1	x_{11}	x_{21}	…	x_{k1}
y_2	x_{12}	x_{22}	…	x_{k2}
⋮	⋮	⋮		⋮
y_n	y_{1n}	x_{2n}	…	x_{kn}

Análisis de regresión

Parámetro	Estimación	Error estándar	Estadístico	Valor-p
Intercepción	$\hat{\beta}_0$	$\sqrt{CM_E C_{11}}$	$\frac{\hat{\beta}_0}{\sqrt{CM_E C_{11}}}$	$\Pr(T > \lvert t_0 \rvert)$
β_1	$\hat{\beta}_1$	$\sqrt{CM_E C_{22}}$	$\frac{\hat{\beta}_1}{\sqrt{CM_E C_{22}}}$	$\Pr(T > \lvert t_0 \rvert)$
⋮	⋮	⋮	⋮	⋮
β_k	$\hat{\beta}_k$	$\sqrt{CM_E C_{k+1,k+1}}$	$\frac{\hat{\beta}_k}{\sqrt{CM_E C_{k+1,k+1}}}$	$\Pr(T > \lvert t_0 \rvert)$

Prueba de hipótesis

$H_0 : \beta_1 = \beta_2 = \ldots \beta_k = 0$

$H_A : \beta_j \neq 0$ para al menos un $j = 1, 2, \ldots, k$

ANOVA

Fuente de variación	Suma de cuadrados	Grados de libertad	Cuadrado medio	F_0	Valor-p
Regresión	$SC_R = \hat{\beta}'\mathbf{X}'\mathbf{y} - \frac{(\Sigma_{i=1}^n y_i)^2}{n}$	k	CM_R	CM_R/CM_E	$\Pr(F > F_0)$
Error o residuo	$SC_E = \mathbf{y}'\mathbf{y} - \hat{\beta}'\mathbf{X}'\mathbf{y}$	$n-k-1$	CM_E		
Total	$S_{yy} = \mathbf{y}'\mathbf{y} - \frac{(\Sigma_{i=1}^n y_i)^2}{n}$	$n-1$			

Ejemplo

- Se tiene una muestra de 8 mujeres, donde se quiere predecir el peso en función a una serie de variables. La información se muestra a continuación:

Registro	estatura	pie	l_brazo	a_espalda	d_cráneo	peso
1	158	36	68	43	55	43
2	152	34	66	40	55	45
3	168	39	73	41	54	48
4	159	36	69	42	57	49
5	158	36	69	44	57	50
6	164	36	71	45	54	51
7	156	36	67	36	56	52
8	167	37	73	42	58	52

Análisis de regresión lineal

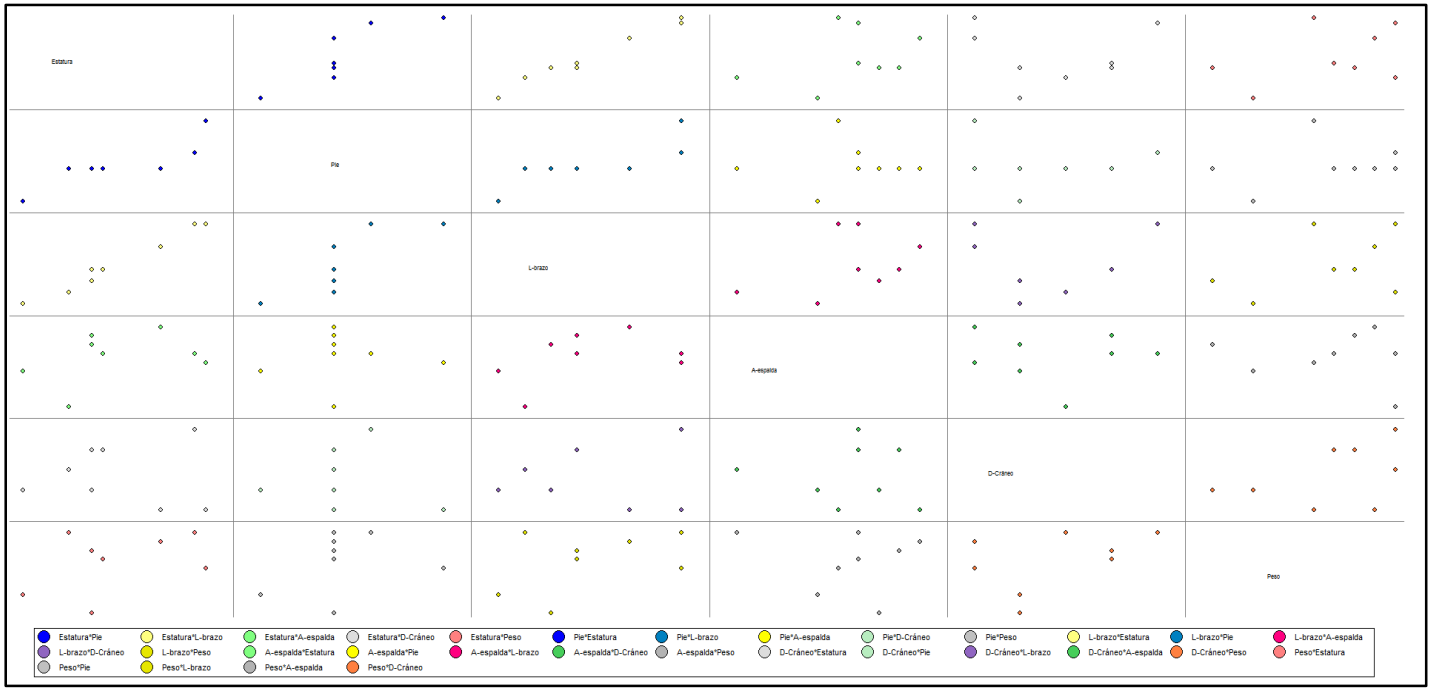
Variable	N	R^2	R^2 Aj	ECMP	AIC	BIC
Peso	8	0.50	0.00	1056.81	49.10	49.66

Coeficientes de regresión y estadísticos asociados

Coef	Est.	E.E.	LI(95%)	LS(95%)	T	p-valor	CpMallows	VIF
const	-32.60	111.52	-512.44	447.24	-0.29	0.7976		
Estatura	0.83	2.61	-10.41	12.08	0.32	0.7803	4.10	78.80
Pie	-1.48	2.72	-13.17	10.20	-0.55	0.6394	4.30	5.28
L-brazo	-0.38	5.43	-23.72	22.97	-0.07	0.9510	4.00	74.95
A-espalda	-0.52	0.79	-3.94	2.90	-0.66	0.5787	4.43	1.80
D-Cráneo	0.89	1.36	-4.97	6.75	0.65	0.5802	4.43	1.53

Cuadro de Análisis de la Varianza (SC tipo III)

F.V.	SC	gl	CM	F	p-valor
Modelo.	37.80	5	7.56	0.40	0.8226
Estatura	1.91	1	1.91	0.10	0.7803
Pie	5.63	1	5.63	0.30	0.6394
L-brazo	0.09	1	0.09	4.8E-03	0.9510
A-espalda	8.14	1	8.14	0.43	0.5787
D-Cráneo	8.06	1	8.06	0.43	0.5802
Error	37.70	2	18.85		
Total	75.50	7			



Ejemplo:

- Se presenta un experimento secuencial para optimizar la producción de un colorante natural. En la etapa final se delimitó una zona de experimentación donde se sospecha que se encuentran las condiciones óptimas para la producción de este colorante en función de la concentración de carbono ($X1$) y temperatura ($X2$).

Y	X1	X2
5707	9	17
5940	13	17
3015	9	25
2673	13	25
5804	8	21
6700	13	21
5310	11	15
7250	11	26
7521	11	21
7642	11	21
7500	11	21
7545	11	21

Análisis de regresión lineal

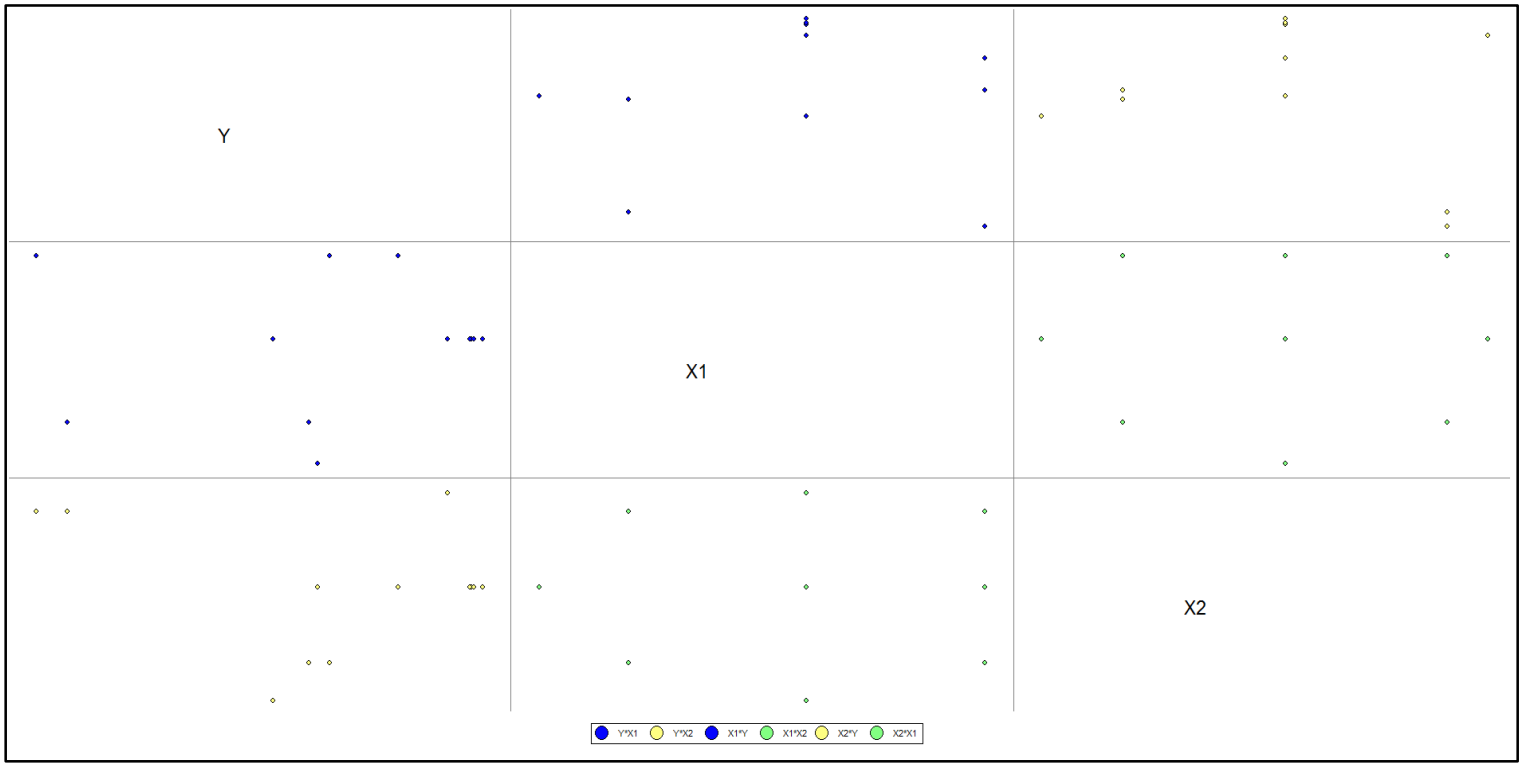
Variable	N	R^2	R^2 Aj	ECMP	AIC	BIC
Y	12	0.05	0.00	6864357.82	219.11	221.05

Coeficientes de regresión y estadísticos asociados

Coef	Est.	E.E.	LI(95%)	LS(95%)	T	p-valor	CpMallows	VIF
const	7608.77	5130.96	-3998.28	19215.81	1.48	0.1722		
X1	62.65	343.47	-714.33	839.63	0.18	0.8593	1.03	1.00
X2	-107.19	165.25	-481.02	266.64	-0.65	0.5328	1.42	1.00

Cuadro de Análisis de la Varianza (SC tipo III)

F.V.	SC	gl	CM	F	p-valor
Modelo.	1549956.17	2	774978.09	0.23	0.8012
X1	113501.03	1	113501.03	0.03	0.8593
X2	1435332.87	1	1435332.87	0.42	0.5328
Error	30701588.75	9	3411287.64		
Total	32251544.92	11			

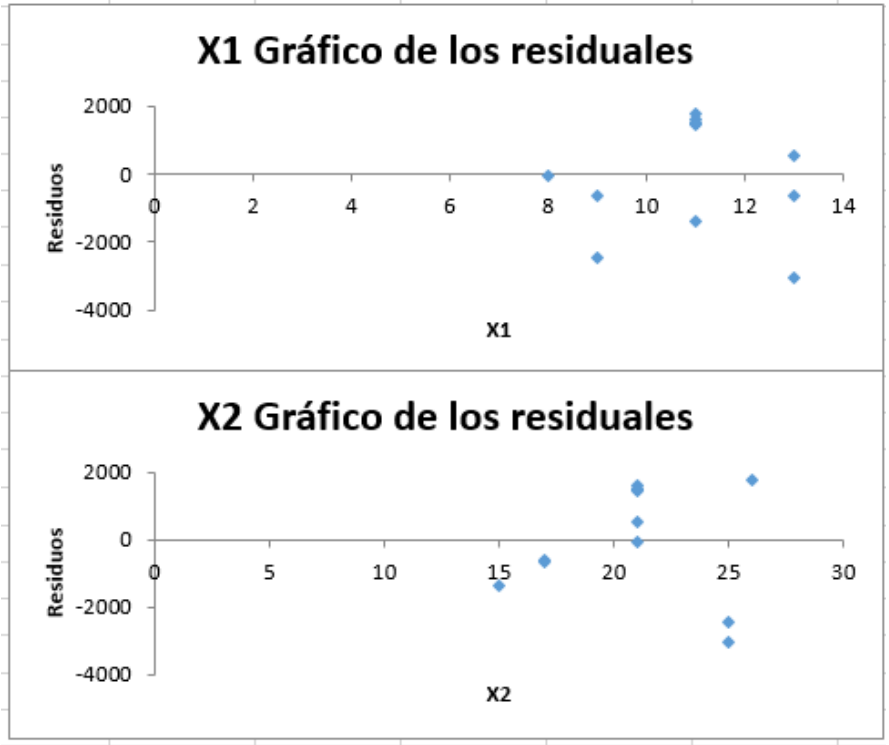


Estadísticas de la regresión	
Coeficiente de corr	0.219222157
Coeficiente de dete	0.048058354
R^2 ajustado	-0.163484234
Error típico	1846.967146
Observaciones	12

ANÁLISIS DE VARIANZA

	Grados de libertad	na de cuadrad	dio de los cua	F	Valor crítico de F
Regresión	2	1549956.17	774978.085	0.22718052	0.801210202
Residuos	9	30701588.7	3411287.64		
Total	11	32251544.9			

	Coeficientes	Error típico	Estadístico t	Probabilidad	Inferior 95%	Superior 95%	Inferior 95.0%	Superior 95.0%
Intercepción	7608.766922	5130.96375	1.48291184	0.17224599	-3998.279483	19215.81333	-3998.279483	19215.81333
X1	62.65074055	343.467361	0.18240668	0.85930614	-714.3264105	839.6278916	-714.3264105	839.6278916
X2	-107.1930282	165.253016	-0.64866004	0.53275799	-481.0213219	266.6352655	-481.0213219	266.6352655



Regresión no lineal o de orden superior

- Los parámetros son lineales cuando cada término del modelo es aditivo y contiene solo un parámetro que multiplica el término.
- Es un problema de inferencia para un modelo del tipo

$$y = f(x, \theta) + \varepsilon$$

- Está fundamentado en datos multidimensionales *x*, *y*, donde f es una función no lineal de algunos parámetros desconocidos θ.
- Se pretende obtener los valores de los parámetros de tal manera que se obtenga la mejor curva de ajuste.
- Se utiliza cuando no pueda modelarse adecuadamente la relación con parámetros lineales.

- Se quiere probar un modelo de segundo orden:

$$y_i = \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + \beta_{12} x_{1i} x_{2i} + \beta_{11} x_{1i}^2 + \beta_{22} x_{2i}^2 + \varepsilon_i$$

y	X1	x2	x1x2	$x1^2$	$x2^2$
5707	9	17	153	81	289
5940	13	17	221	169	289
3015	9	25	225	81	625
2673	13	25	325	169	625
5804	8	21	168	64	441
6700	13	21	273	169	441
5310	11	15	165	121	225
7250	11	26	286	121	676
7521	11	21	231	121	441
7642	11	21	231	121	441
7500	11	21	231	121	441
7545	11	21	231	121	441

Resumen									
Estadísticas de la regresión									
Coeficiente (	0.79266847								
Coeficiente (	0.62832331								
R^2 ajustado	0.31859273								
Error típico	1413.45602								
Observacion	12								
ANÁLISIS DE VARIANZA									
	Grados de liberta	*a de cuadrad*	*io de los cua*	*F*	*alor crítico de F*				
Regresión	5	20264397.4	4052879.47	2.02861245	0.20715042				
Residuos	6	11987147.6	1997857.93						
Total	11	32251544.9							
	Coeficientes	*Error típico*	*Estadístico t*	*Probabilidad*	*Inferior 95%*	*Superior 95%*	*nferior 95.0%*	*uperior 95.0%*	
Intercepción	-71007.2838	32101.5037	-2.2119613	0.06895682	-149556.834	7542.26602	-149556.834	7542.26602	
X1	8091.25305	3863.2781	2.09440088	0.08110317	-1361.84793	17544.354	-1361.84793	17544.354	
x2	3473.11924	1771.74197	1.96028502	0.09765656	-862.17719	7808.41567	-862.17719	7808.41567	
x1x2	-17.96875	88.3410014	-0.20340215	0.84554363	-234.131393	198.193893	-234.131393	198.193893	
X1sq	-355.726778	157.612486	-2.2569708	0.06481733	-741.390638	29.937082	-741.390638	29.937082	
X2sq	-81.6321728	35.6846521	-2.28759895	0.06214839	-168.949371	5.68502528	-168.949371	5.68502528	

Análisis de regresión lineal

Variable	N	R^2	R^2 Aj	ECMP	AIC	BIC
Y	12	0.63	0.32	16032059.28	213.83	217.22

Coeficientes de regresión y estadísticos asociados

Coef	Est.	E.E.	LI(95%)	LS(95%)	T	p-valor	CpMallows	VIF
const	-71007.28	32101.50	-149556.82	7542.25	-2.21	0.0690		
X1	8091.25	3863.28	-1361.85	17544.35	2.09	0.0811	8.39	216.02
X2	3473.12	1771.74	-862.18	7808.41	1.96	0.0977	7.84	196.27
X1X2	-17.97	88.34	-234.13	198.19	-0.20	0.8455	4.04	109.71
X1sq	-355.73	157.61	-741.39	29.94	-2.26	0.0648	9.09	166.08
X2sq	-81.63	35.68	-168.95	5.69	-2.29	0.0621	9.23	137.18

Cuadro de Análisis de la Varianza (SC tipo III)

F.V.	SC	gl	CM	F	p-valor
Modelo.	20264397.36	5	4052879.47	2.03	0.2072
X1	8763633.84	1	8763633.84	4.39	0.0811
X2	7677203.30	1	7677203.30	3.84	0.0977
X1X2	82656.25	1	82656.25	0.04	0.8455
X1sq	10176922.80	1	10176922.80	5.09	0.0648
X2sq	10455008.25	1	10455008.25	5.23	0.0621
Error	11987147.56	6	1997857.93		
Total	32251544.92	11			

RegMultiple

Caso	Y	X1	X2
1	5707	9	17
2	5940	13	17
3	3015	9	25
4	2673	13	25
5	5804	8	21
6	6700	13	21
7	5310	11	15
8	7250	11	26
9	7521	11	21
10	7642	11	21
11	7500	11	21
12	7545	11	21

Análisis de regresión no lineal

Caso

1(0)

Seleccionar si contiene..

Variables | Particiones ...

Variable dependiente

Análisis de regresión no lineal

Y=

Sólo Nelder-Mead

Verificar la sintaxis del modelo

+ − * / ^

B0+B1*X1+B2*X2+B2*X1*X2+B4*X1*X1+B5*X2*X2

Regresoras

X1
X2

Covarianzas y correlacione
Mostrar solo los parámetros
Graficar

Parámetros en el modelo

B0
B1
B2
B4
B5

Guardar.....

Residuos
Residuos Estandarizados
Predichos
Sobrescribir

Aceptar

Cancelar

Ayuda

Análisis de regresión no lineal

*Modelo Y B0+B1*X1+B2*X2+B3*X1*X2+B4*X1*X1+B5*X2*X2*

Variable	N	CMError	Sigma	AIC	BIC	Iteración
Y	12	1997857.93	1413.46	213.83	217.22	6

Parámetros	Cota inf.	Cota sup.	Val.Ini.	Estimación	E.E.	T	p-valor
B0	-1E30	1E30	1.0E-03	-71007.63	32100.94	-2.21	0.0690
B1	-1E30	1E30	1.0E-03	8091.26	3863.12	2.09	0.0811
B2	-1E30	1E30	1.0E-03	3473.15	1771.73	1.96	0.0977
B3	-1E30	1E30	1.0E-03	-17.97	88.34	-0.20	0.8455
B4	-1E30	1E30	1.0E-03	-355.73	157.61	-2.26	0.0648
B5	-1E30	1E30	1.0E-03	-81.63	35.68	-2.29	0.0621

Matriz de covarianzas de las estimaciones

	B0	B1	B2	B3	B4	B5
B0	1030470195.14	-104267571.81	-45208055.73	1802725.12	3054078.46	604159.01
B1	-104267571.81	14923694.49	2366701.15	-163879.50	-532437.43	-13393.28
B2	-45208055.73	2366701.15	3139015.23	-85844.72	-25437.33	-52671.67
B3	1802725.12	-163879.50	-85844.72	7804.13	-0.35	-0.02
B4	3054078.46	-532437.43	-25437.33	-0.35	24839.60	603.80
B5	604159.01	-13393.28	-52671.67	-0.02	603.80	1273.37

Matriz de correlación de las estimaciones

	B0	B1	B2	B3	B4	B5
B0	1.00	-0.84	-0.79	0.64	0.60	0.53
B1	-0.84	1.00	0.35	-0.48	-0.87	-0.10
B2	-0.79	0.35	1.00	-0.55	-0.09	-0.83
B3	0.64	-0.48	-0.55	1.00	-2.5E-05	-5.8E-06
B4	0.60	-0.87	-0.09	-2.5E-05	1.00	0.11
B5	0.53	-0.10	-0.83	-5.8E-06	0.11	1.00

Estrategias de modelado

- En un análisis predictivo el mejor modelo es el que produce predicciones más *fiables* para una nueva observación, mientras que en un análisis estimativo el mejor modelo es el que produce estimaciones más *precisas* para el coeficiente de la variable de interés.
- En ambos casos se prefiere el modelo más sencillo posible (a este modo de seleccionar modelos se le denomina *parsimonia*).
- Sin embargo, hay una serie de pasos que deben realizarse siempre:
  - Especificación del modelo máximo.
  - Especificación de un criterio de comparación de modelos y definición de una estrategia para realizarla.
  - Evaluación de la fiabilidad del modelo.

Especificando el modelo máximo

- El criterio para decidir qué variables forman el modelo máximo se establece en función de sus objetivos y del conocimiento teórico que se tenga sobre el problema, evidentemente cuanto menor sea el conocimiento previo mayor tenderá a ser el modelo máximo.
- Un modelo máximo grande minimiza la probabilidad de error tipo II o *infra-ajuste*, que en un análisis de regresión consiste en no considerar una variable que realmente tiene un coeficiente de regresión distinto de cero.
- Un modelo máximo pequeño minimiza la probabilidad de error tipo I o *sobreajuste*(incluir en el modelo una variable independiente cuyo coeficiente de regresión realmente sea cero).
- Aunque el sobreajuste no introduce sesgos en la estimación de los coeficientes, un infra-ajuste puede producirlos, pero que un modelo máximo grande aumenta la probabilidad de problemas de colinealidad.

Comparando los modelos

- Debe establecerse cómo y con qué se comparan los modelos.
- Si bien hay varios estadísticos sugeridos para comparar modelos, el más frecuentemente usado es la F parcial, recordando que cuando los dos modelos sólo difieren en una variable, el contraste sobre la F parcial es exactamente el mismo que el realizado con la t sobre el coeficiente de regresión
- Por otro lado a veces interesa contrastar varias variables conjuntamente mejor que una a una (por ejemplo todos los términos no lineales) o, incluso, es necesario hacerlo (por ejemplo para variables indicadoras).
- En un análisis estimativo el criterio para incluir o excluir variables distintas a las de interés, es sobre todo los cambios en los coeficientes y no los cambios en la significación del modelo.

Selección de las variables regresoras

- La selección de variables regresoras es un procedimiento estadístico importante por que no todas las variables regresoras tienen igual importancia.
- Algunas variables regresoras pueden perjudicar la confiabilidad del modelo, en especial si están correlacionadas entre ellas.
- Computacionalmente es más fácil trabajar con un conjunto pequeño de variables regresoras.
- Es más económico recolectar información para un modelo con pocas variables.
- Si se reduce el número de variables entonces el modelo cumple con el principio de la parsimonia.

Algunos métodos de selección de las variables

- La idea de estos métodos es elegir el mejor modelo en forma secuencial pero incluyendo o excluyendo una sola variable regresora en cada paso de acuerdo a ciertos criterios.
- El proceso secuencial termina cuando se satisface una regla de parada establecida.
- Hay tres algoritmos usados: Backward Elimination, Forward Selection y Stepwise Selección.

Backward elimination

- Se comienza con el modelo completo y en cada paso se va eliminando una variable. Toda variable que sale, no puede entrar.
- Si todas las variables regresoras son importantes, es decir tienen p-value pequeños para la prueba t, entonces el mejor modelo es el que tiene todas las variables regresoras disponibles.
- En caso contrario, en cada paso la variable que se elimina del modelo es aquella que satisface cualquiera de los siguientes requisitos equivalentes entre sí:
  - Aquella variable que tiene el estadístico de t, en valor absoluto, más pequeño entre las variables incluidas aún en el modelo.
  - Aquella variable que produce la menor disminución en el R^2 al ser eliminada del modelo.
  - Aquella variable que tiene la correlación parcial (en valor absoluto) más pequeña con la variable de respuesta, tomando en cuenta las variables aún presentes en el modelo.
- El proceso termina cuando se llega a un modelo con un número prefijado $p*$ de variables regresoras.

Forward elimination

- Se empieza con aquella variable regresora que tiene la más alta correlación con la variable respuesta.
- En este caso, toda variable que entra no puede salir.
- En el siguiente paso se añade al modelo la variable que reúne cualquiera de estos requisitos equivalentes entre sí.
  - Aquella variable que produce el mayor incremento en el R^2 al ser añadida al modelo.
  - El proceso termina cuando se llega a un modelo con un número prejado $p*$ de variables predictoras.

Stepwise selection

- Se empieza con un modelo de regresión simple y en cada paso se puede añadir una variable en forma similar al método forward
- Se coteja si alguna de las variables que ya están presentes en el modelo puede ser eliminada.
- El proceso termina cuando ninguna de las variables fuera del modelo tiene importancia suficiente como para ingresar al modelo.

Evaluando la fiabilidad del modelo

- Una vez encontrado el mejor modelo hay que evaluar su fiabilidad, es decir, evaluar si se comporta igual en otras muestras extraídas de la misma población.
- Evidentemente, el modo más completo de evaluarlo será repetir el estudio con otra muestra y comprobar que se obtienen los mismos resultados, aunque generalmente esta aproximación resulta excesivamente costosa.
- Otra aproximación alternativa consiste en partir aleatoriamente la muestra en dos grupos y ajustar el modelo con cada uno de ellos y si se obtienen los mismos resultados se considera que el modelo es fiable.
- Esta aproximación es demasiado estricta ya que, en la práctica, casi nunca se obtienen los mismos resultados.
- Una validación menos estricta consiste en ajustar el modelo sobre uno de los grupos (grupo de trabajo) y calcular su R^2, que se puede interpretar como el cuadrado del coeficiente de correlación simple entre la variable dependiente y las estimaciones obtenidas en la regresión.

Se tiene un análisis de la eficiencia en peso, de un proceso de producción. Se escogieron seis variables que se cree pueden regir el proceso. Determine el mejor modelo de regresión

N	Y	X1	X2	X3	X4	X5	X6
1	138.00	8.00	534.00	107.39	138.08	690.00	138.08
2	58.00	12.00	182.00	43.80	72.40	290.00	72.40
3	30.00	10.00	546.00	0.43	35.55	150.00	35.55
4	30.00	13.00	367.00	12.06	30.19	150.00	30.19
5	69.00	5.00	478.00	64.41	63.01	345.00	-63.01
6	49.00	6.00	330.00	25.89	46.63	245.00	-46.63
7	30.00	7.00	165.00	18.29	32.20	150.00	32.20
8	136.00	12.00	184.00	86.63	139.33	680.00	139.33
9	119.00	11.00	355.00	104.37	97.54	595.00	-97.54
10	93.00	5.00	469.00	62.52	98.09	465.00	-98.09
11	19.00	8.00	246.00	16.65	21.30	95.00	-21.30
12	59.00	8.00	323.00	53.97	45.33	295.00	-45.33
13	143.00	12.00	395.00	115.52	142.41	715.00	-142.41
14	107.00	12.00	423.00	84.25	121.20	535.00	-121.20
15	81.00	8.00	525.00	56.16	69.19	405.00	-69.19
16	71.00	15.00	184.00	58.55	61.44	355.00	-61.44
17	103.00	5.00	160.00	3.66	100.49	515.00	-100.49
18	112.00	8.00	259.00	98.80	113.76	560.00	113.76
19	76.00	8.00	395.00	4.63	80.45	380.00	80.45
20	149.00	5.00	476.00	79.87	148.40	745.00	-148.40
21	126.00	11.00	173.00	111.57	122.65	630.00	122.65
22	47.00	7.00	175.00	19.65	46.43	235.00	-46.43
23	140.00	8.00	209.00	113.07	142.50	700.00	142.50
24	59.00	10.00	292.00	39.36	50.62	295.00	-50.62
25	85.00	12.00	220.00	34.39	85.51	425.00	-85.51
26	19.00	10.00	251.00	6.47	3.57	95.00	-3.57
27	6.00	8.00	354.00	5.91	0.10	30.00	0.10
28	78.00	14.00	222.00	2.56	59.08	390.00	59.08
29	12.00	14.00	560.00	4.39	7.52	60.00	7.52
30	26.00	11.00	501.00	4.70	19.68	130.00	19.68

Análisis de regresión lineal

Variable	N	R^2	R^2 Aj	ECMP	AIC	BIC
y	30	0.23	0.03	3041.60	318.61	329.82

Coeficientes de regresión y estadísticos asociados

Coef	Est.	E.E.	LI(95%)	LS(95%)	T	p-valor	CpMallows	VIF
const	72.37	40.73	-11.88	156.62	1.78	0.0888		
x1	-2.38	2.85	-8.27	3.52	-0.83	0.4128	5.70	1.08
x2	0.01	0.06	-0.12	0.13	0.09	0.9293	5.01	1.08
x3	-0.42	0.37	-1.19	0.35	-1.13	0.2702	6.28	3.55
x4	-2.07	0.99	-4.11	-0.03	-2.10	0.0470	9.41	32.05
x5	0.50	0.21	0.06	0.94	2.34	0.0280	10.50	34.76
x6	-0.02	0.10	-0.22	0.18	-0.22	0.8257	5.05	1.13

Cuadro de Análisis de la Varianza (SC tipo III)

F.V.	SC	gl	CM	F	p-valor
Modelo.	12861.59	6	2143.60	1.17	0.3570
x1	1276.67	1	1276.67	0.70	0.4128
x2	14.75	1	14.75	0.01	0.9293
x3	2342.26	1	2342.26	1.28	0.2702
x4	8085.20	1	8085.20	4.41	0.0470
x5	10090.01	1	10090.01	5.50	0.0280
x6	91.00	1	91.00	0.05	0.8257
Error	42204.58	23	1834.98		
Total	55066.17	29			

Análisis de regresión lineal

Variable	N	R^2	R^2 Aj	ECMP	AIC	BIC
y	30	0.15	0.09	2078.14	313.68	319.28

Eliminación backward. Máximo p-valor para retener: 0.15
Número original de regresoras: 6,regresoras retenidas en el modelo 2

Coeficientes de regresión y estadísticos asociados

Coef	Est.	E.E.	LI(95%)	LS(95%)	T	p-valor	CpMallows	VIF
const	55.59	15.82	23.12	88.06	3.51	0.0016		
x4	-2.00	0.93	-3.91	-0.09	-2.15	0.0408	5.62	30.32
x5	0.42	0.19	0.03	0.82	2.19	0.0373	5.80	30.32

Error cuadrático medio: 1731.518962

Variable	N	R^2	R^2 Aj	ECMP	AIC	BIC
y	30	0.00	0.00	2032.05	314.59	317.39

Selección Forward. Máximo p-valor para entrar: 0.15
Número original de regresoras: 6,regresoras retenidas en el modelo 0

Coeficientes de regresión y estadísticos asociados

Coef	Est.	E.E.	LI(95%)	LS(95%)	T	p-valor	CpMallows	VIF
const	69.83	7.96	53.56	86.10	8.78	<0.0001		

Error cuadrático medio: 1898.833333

Variable	N	R^2	R^2 Aj	ECMP	AIC	BIC
y	30	0.00	0.00	2032.05	314.59	317.39

Selección Stepwise.
Máximo p-valor para entrar: 0.15
Máximo p-valor para retener: 0.15
Número original de regresoras: 6,regresoras retenidas en el modelo 0

Coeficientes de regresión y estadísticos asociados

Coef	Est.	E.E.	LI(95%)	LS(95%)	T	p-valor	CpMallows	VIF
const	69.83	7.96	53.56	86.10	8.78	<0.0001		

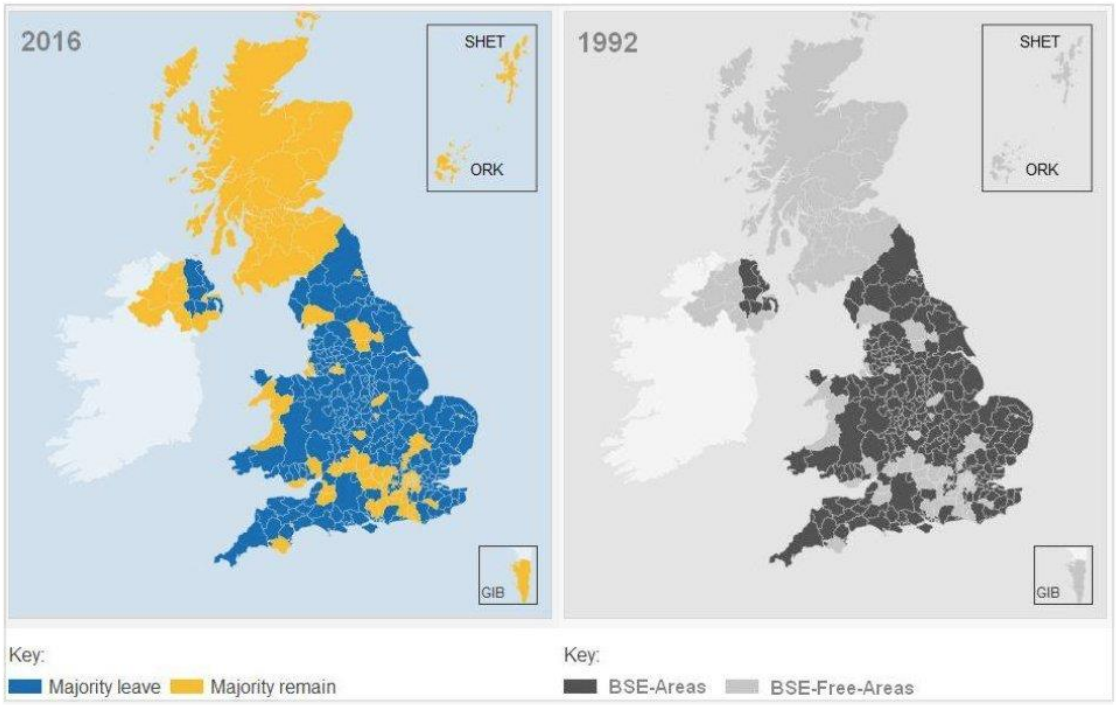
Error cuadrático medio: 1898.833333

Riesgos de la regresión

- La relación entre las variables puede ser espuria y el usuario es quien debe investigar si tal relación es de tipo causa-efecto. Se debe tomar en cuenta que algunas de las razones por las que las variables *X* y *Y* aparecen relacionadas de manera significativa son:
  - *X* influye sobre *Y*.
  - *Y* influye sobre *X*.
  - *X* y *Y* interactúan entre sí, una tercera variable *Z* influye sobre ambas y es lcausante de tal relación.
  - *X* y *Y* actúan en forma similar debido al azar.
  - *X* y *Y* aparecen relacionados debido a que la muestra no es representativa.
- Hacer extrapolaciones indiscriminadas con base en el modelo. Para no incurrir en esto se debe tener cuidado en cuanto a extrapolar más allá de la región que contienen las observaciones originales.

Ejemplo de relaciones espurias:

- http://www.tylervigen.com/spurious-correlations
- https://rodas5.us.es/items/ea5754ec-0941-2a2a-32e1-b34998971302/1/viewcontent/5742ce5f-de69-479e-8251-f507e4488223?_sl.t=true



Michael Von Freising
24 June at 19:23 ·

EU Referendum Local Results 2016 vs. Mad Cow Disease Outbreak Areas 1992

However, it would be a mistake to jump to conclusions.