VNIVERSITAT DE VALÈNCIA

MASTER DE INGENIERÍA BIOMÉDICA.

Métodos de ayuda al diagnóstico clínico.

Tema 4: Modelos lineales.

Emilio Soria, Antonio José Serrano y José David Martín Dpto Ingeniería Electrónica, ETSE
Sistemas de Ayuda a la Decisión Clínica, Curso 2009-2010

Objetivos del tema

Conocer los parámetros que indican la posible relación lineal entre variables.

Aprender la base de los métodos de mínimos cuadrados.

Implementar y analizar un modelo lineal.

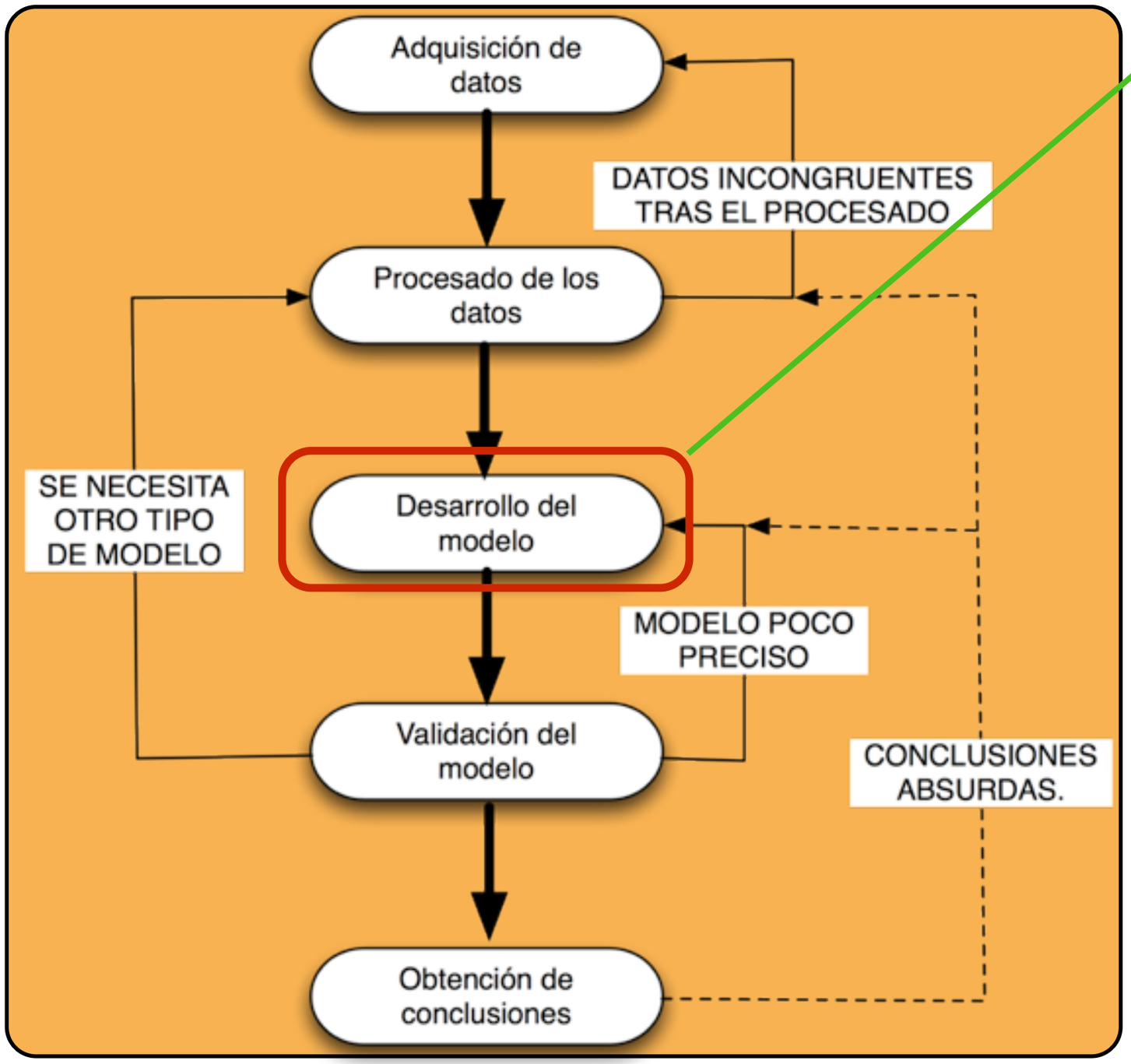
Interpretar los resultados obtenidos con un modelo lineal.

Conocer cómo se puede aumentar la potencia de un modelo manteniendo la linealidad en sus parámetros.

Conocer y desarrollar una regresión logística. Limitaciones/ventajas.

Aprender las limitaciones de los modelos lineales y las posibles formas que se tienen de superar dichas limitaciones.

Dónde estamos



Ya se ha realizado el procesado de los datos previos (se han completado, seleccionado variables, EDA y clustering); el siguiente paso es desarrollar un modelo que resuelva el problema planteado.

Los tipos de problemas que se abordan en un problema de decisión clínica son:

Problema	***Descripción***
Clasificación	$\vec{X}_n \Rightarrow \begin{cases} C_1 \\ \dots \\ C_N \end{cases}$
Modelización	$\vec{X}_n \Rightarrow \vec{Y}_n$
Predicción.	$\vec{X}_n \Rightarrow \vec{X}_{n+p}$

En los tres casos hay que establecer una correspondencia entre unos datos que llamaremos de entrada al modelo y otras que llamaremos de salida Y=g(X).

Empezaremos por los modelos más simples: LINEALES EN LOS PARÁMETROS.

Covarianza y coeficiente de correlación (I).

Dadas dos variables se define el estimador de la ***covarianza*** entre dichas variables

$$Cov(x,y) = \frac{1}{N}\sum_{i=1}^{N}(x_i - m_x)\cdot(y_i - m_y)$$

Este parámetro depende del rango de las variables x e y; se plantea usar un parámetro que no tenga en cuenta dicho rango; aparece entonces lo que se conoce como coeficiente de correlación.

$$coefc(x,y) = r_{xy} = \frac{Cov(x,y)}{\sigma_x \cdot \sigma_y}$$

$$r_{xy} = \frac{\sum_{i=1}^{N}(x_i - m_x)\cdot(y_i - m_y)}{\sqrt{\sum_{i=1}^{N}(x_i - m_x)^2}\cdot\sqrt{\sum_{i=1}^{N}(y_i - m_y)^2}}$$

Aquí m_k es el valor medio de la variable k. Este parámetro indica la posible relación lineal existente entre las variables x e y. Si no se tiene dicha relación, *o es de tipo no lineal*, dicho parámetro toma un valor cercano a 0.

Las propiedades más importantes de este coeficiente son:

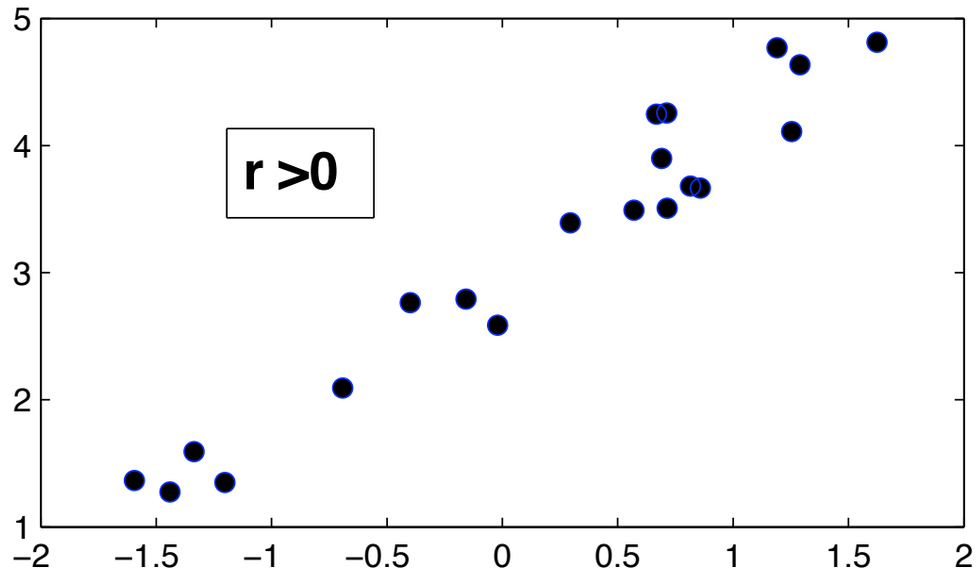
No tiene dimensiones

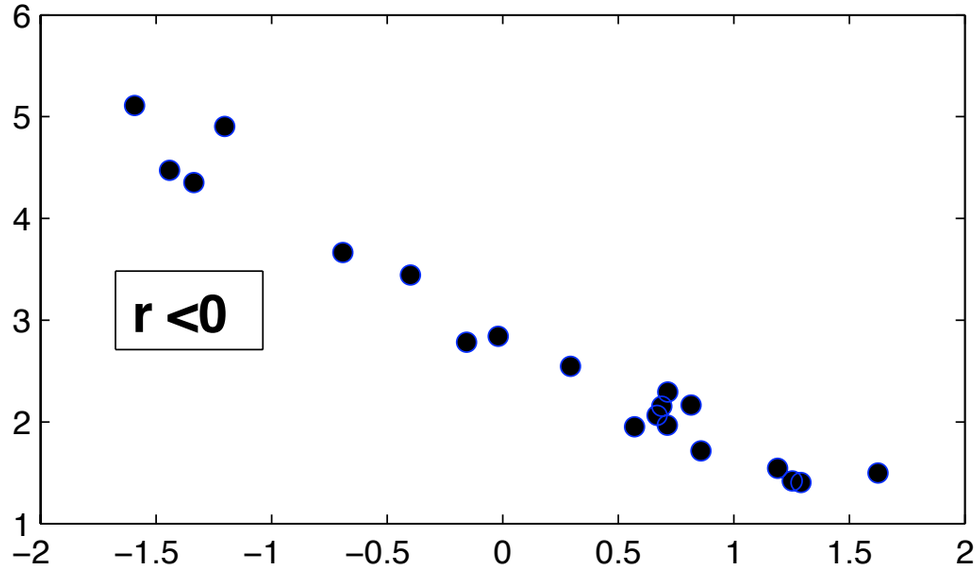
Es invariante bajo transformaciones lineales de las variables.

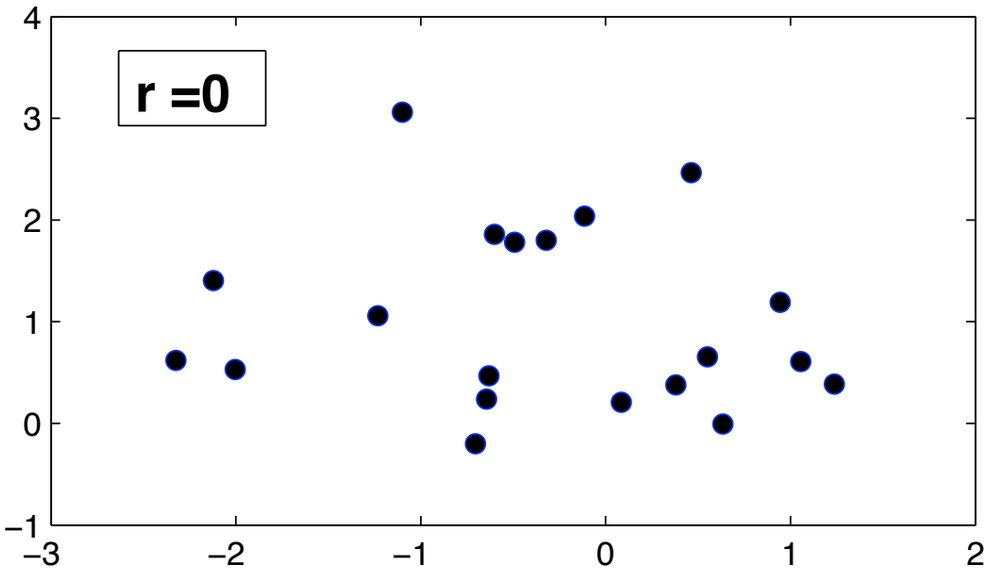
Está entre -1 y 1.

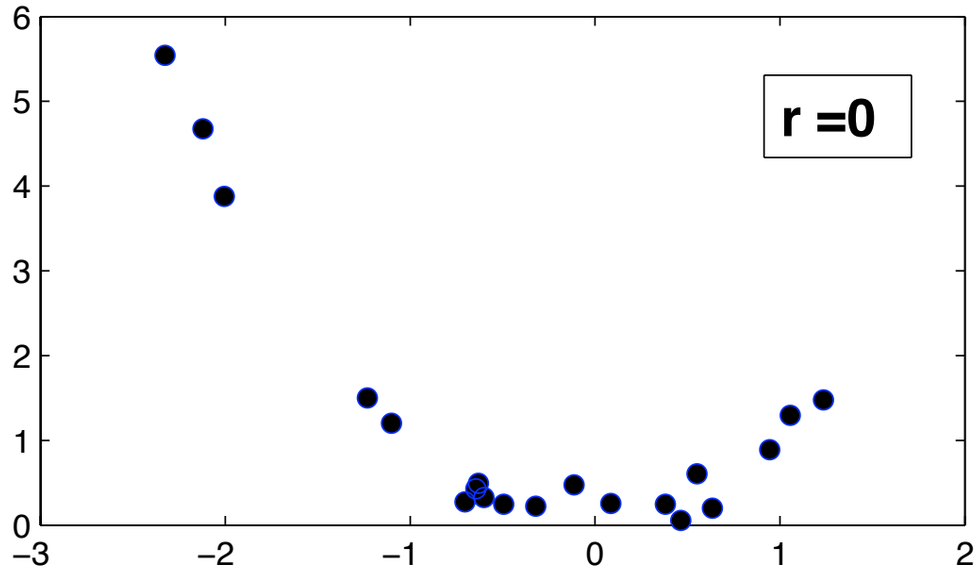
Su cercanía a 0 indica la ausencia de relación lineal.

Coeficiente de correlación (II)









Regresión simple (I)

En ese caso buscamos encontrar la relación lineal entre una variable que denominamos independiente (o variable respuesta) y otra que denominaremos dependiente

Finalidad	Objetivo
Descripción	Describir las relaciones entre las variables
Estimación.	Estimar la variable de salida dentro del rango que se tiene
Predicción	Estimar la variable de salida fuera del rango que se tiene
Testeo	Comprobar si existe tal relación mediante test estadísticos

El modelo que se plantea es el siguiente

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

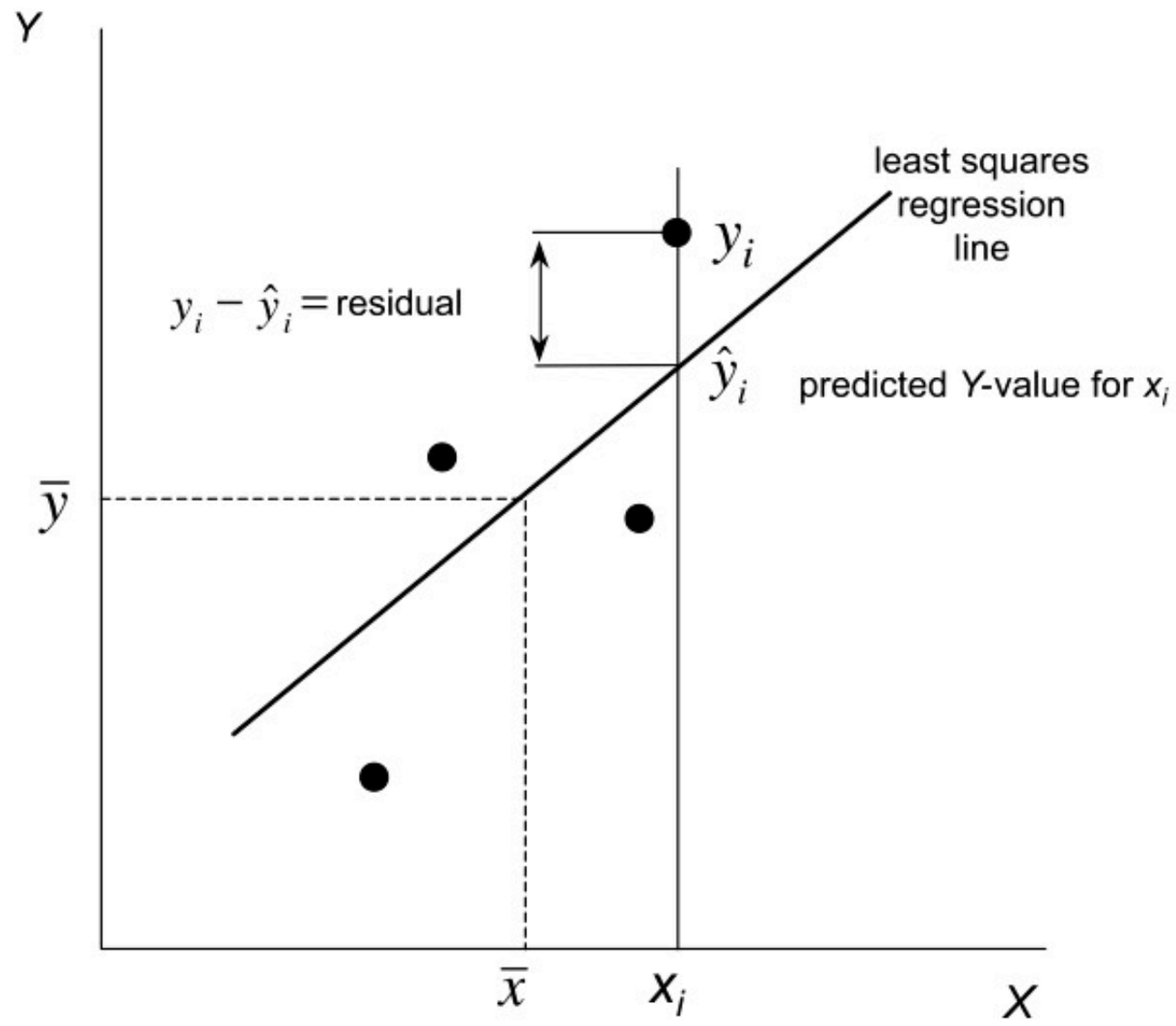
Donde x e y son las variables de interés y los parámetros a determinar son los β_k. Aquí los ϵ_i son variables aleatorias independientes e identicamente distribuidos, i.i.d, de forma normal con media cero y varianza σ^2 estas son la suposiciones que se tienen para este tipo de modelos.

Para determinar los parámetros se define lo que se conoce como una función de coste (lo mismo haremos en redes neuronales) que nos dará lo bueno que es el ajuste. La más usual es la que tiene en cuenta los errores cometidos por el modelo al cuadrado

$$J = \sum_{i=1}^{N} \left\{ y_i - \left[b_0 + b_1 \cdot x_i \right] \right\}^2$$

Regresión simple (II)

$$J = \sum_{i=1}^{N} \left\{ y_i - \left[b_0 + b_1 \cdot x_i \right] \right\}^2$$



Ahora hay que determinar los parámetros a partir de la función de coste anterior. Hay que fijarse que los términos que aparecen en el sumatorio son la diferencia entre el valor real y el valor estimado por el modelo. A esta diferencia se le conoce como **residuo**.

El objetivo es ***minimizar dicha función con respecto a los parámetros que se tienen*** (PROCEDIMIENTO MATEMÁTICO: derivar parcialmente con respecto a cada uno de los parámetros e igualar a cero). Se tiene entonces:

$$\frac{\partial J}{\partial b_1} = 0 \Rightarrow \sum_{i=1}^{N} \left\{ y_i - \left[b_0 + b_1 \cdot x_i \right] \right\} \cdot x_i = 0$$

$$\frac{\partial J}{\partial b_0} = 0 \Rightarrow \sum_{i=1}^{N} \left\{ y_i - \left[b_0 + b_1 \cdot x_i \right] \right\} = 0$$

Regresión simple (III)

Los parámetros obtenidos son

$$b_1 = \frac{\sum_{i=1}^{N} (x_i - m_x) \cdot (y_i - m_y)}{\sum_{i=1}^{N} (x_i - m_x)^2}$$

$$b_0 = m_y - b_1 \cdot m_x$$

A partir de los parámetros calculados se puede estimar la variable de respuesta para cada valor de x. La varianza de los residuos vendrá dada por

$$\sigma_{est}^2 = \frac{\sum_{i=1}^{N} \{y_i - [b_0 + b_1 \cdot x_i]\}^2}{N-2}$$

Una vez estimados los parámetros del modelo cabe preguntarse cómo de precisos son dichos parámetros. Para ello se hace uso del llamado ANOVA (análisis de la varianza) que considera tres términos para analizar (aquí y_{est} es el valor estimado por el modelo lineal):

$$SST = \sum_{i=1}^{N} \{y_i - m_y\}^2 \qquad SSR = \sum_{i=1}^{N} \{y_{iest} - m_y\}^2$$

$$SSE = \sum_{i=1}^{N} \{y_i - y_{iest}\}^2$$

Aquí SS hace referencia a suma de cuadrados y T,R y E a total, regresión y error

Cada uno de los tres parámetros define una característica; SST da la variabilidad de y si se estima usando el valor medio; SSE mide dicha variabilidad si se usa el modelo y SSR es la varianza del propio modelo. A partir del análisis de estos estadísticos y de relaciones entre ello se pede determinar la bondad del modelo lineal planteado

Regresión simple (IV)

La siguiente tabla muestra la forma típica de presentación de resultados en un ANOVA. Aquí n es el número de datos que se tienen

Suma de cuadrados	Expresión	Grados de libertad
Total (SST)	$\sum_{i=1}^{n}(y_i - m_y)^2$	n-1
Regresión (SSR)	$\sum_{i=1}^{n}(y_{iest} - m_y)^2$	1
Error (SSE)	$\sum_{i=1}^{n}(y_{iest} - y_i)^2$	n-2

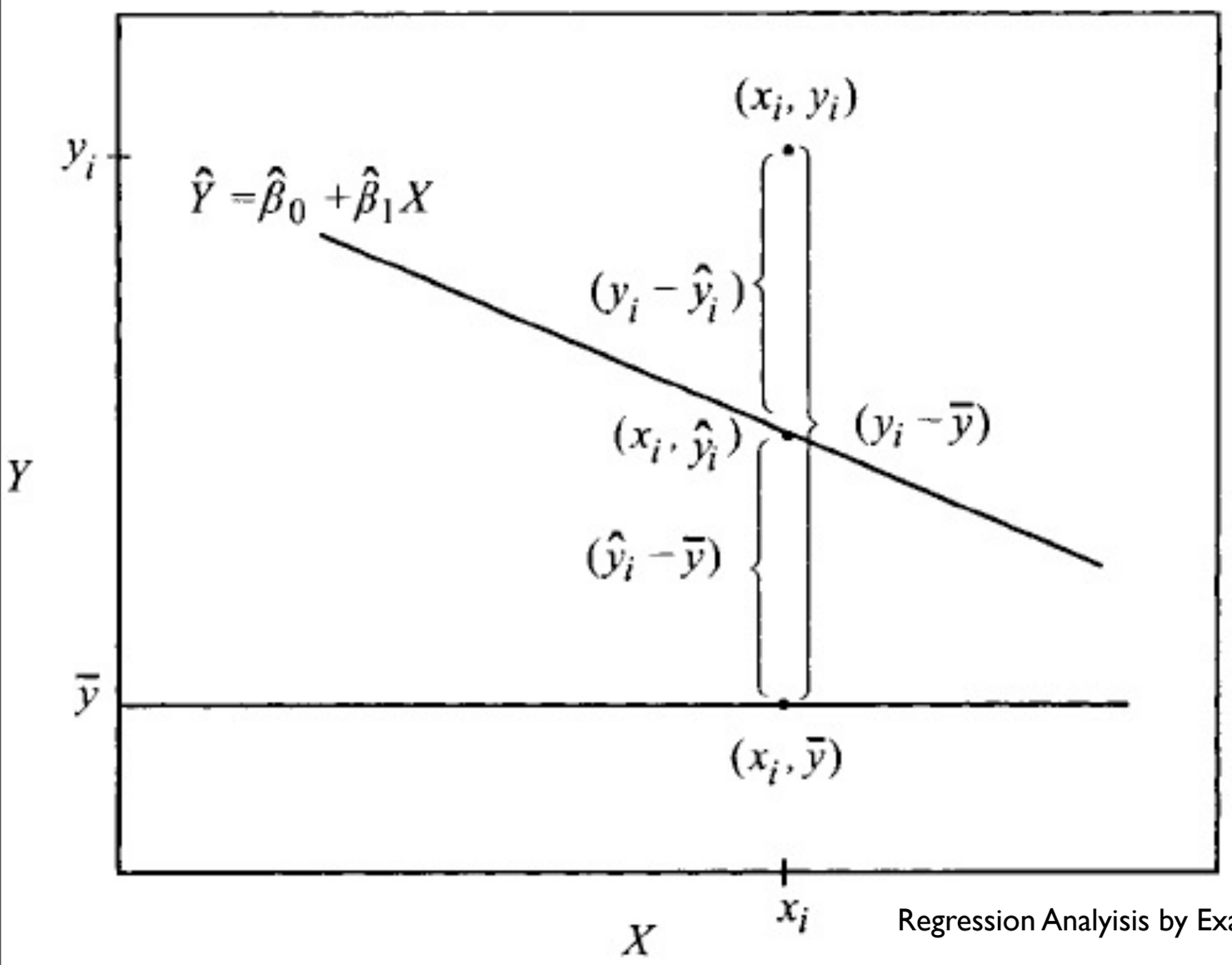
Fuente	Grados de libertad (gl)	SS	MS	F
Regresion	1	SSR	SSR/gl	MSR/MSE
Error	n-2	SSE	SSE/gl	
Total	n-1	SST		

SST=SSR+SSE

$$F = \frac{MSR}{MSE} \approx F_{1,n-2}; H_0 : \beta_1 = 0$$

Distribución F con 1,n-2 grados de libertad.

Regresión simple (V)



Regression Analyisis by Example, Wiley.

$$y_i - \bar{y} = (\hat{y}_i - \bar{y}) + (y_i - \hat{y}_i)$$

A partir de los términos que aparecen en el ANOVA se determina otro parámetro que se conoce como R^2

$$R^2 = \frac{SSR}{SST} = 1 - \frac{SSE}{SST}$$

R^2 da el % de SST que refleja el modelo lineal. Este parámetro irá entre 0 y 1; cercano a 1 indica un ajuste bueno. Otra forma de entenderlo es que indica la mejora del modelo de regresión frente al modelo más "tonto" considerar como estimación de la variable de salida el valor medio SIEMPRE.

Este parámetro es igual al coeficiente de correlación al cuadrado.

Regresión múltiple (I)

El caso analizado hasta ahora, la regresión simple no es el más usual, normalmente se busca establecer relaciones entre una determinada variable de salida y más de una de entrada; tenemos entonces lo que se conoce como ***regresión múltiple***.

$$y_i = \beta_0 + \beta_1 \cdot x_{1i} + ...\beta_p \cdot x_{pi} + \varepsilon_i$$

Las condiciones de la regresión simple se mantienen, esto es, términos ε_k se corresponden a variables aleatorias normales con valor medio 0 y varianza σ^2 no estando relacionados entre ellos (independientes).

Aquí se pueden plantear los siguientes objetivos para el modelo

- Inferencia (estimación y testeo) de los parámetros que forman el modelo
- Estimación/predicción de y
- Selección de qué variables son las mejores para formar el modelo.

El procedimiento de cálculo es el mismo lo que ocurre ahora es que el número de parámetros a calcular crece. Buscamos minimizar la suma de los residuos al cuadrado, esto es....

$$J = \sum_{i=1}^{N} \left\{ y_i - \left[b_0 + b_1 \cdot x_{1i} + ...b_p \cdot x_{pi} \right] \right\}^2$$

Ahora aparecen más derivadas parciales, m+1, que se convierten en un sistema de p+1 ecuaciones lineales con coeficientes constantes.

Regresión múltiple (II)

Se puede analizar el problema desde un punto de vista matricial lo que simplifica la notación. Para ello se definen los siguientes vectores y matrices

$$Y = \begin{bmatrix} y_1 \\ y_2 \\ .... \\ y_N \end{bmatrix} \quad \varepsilon = \begin{bmatrix} \varepsilon_1 \\ \varepsilon_2 \\ .... \\ \varepsilon_N \end{bmatrix} \quad \beta = \begin{bmatrix} \beta_0 \\ \beta_1 \\ .... \\ \beta_p \end{bmatrix}$$

$$X = \begin{bmatrix} 1 & x_{11} & ... & x_{p1} \\ 1 & x_{12} & ... & x_{p2} \\ ... & ... & ... & ... \\ 1 & x_{1N} & ... & x_{pN} \end{bmatrix}$$

$$y_i = \beta_0 + \beta_1 \cdot x_{1i} + ...\beta_p \cdot x_{pi} + \varepsilon_i$$

$$Y = X \cdot \beta + \varepsilon$$

$$J = \varepsilon^t \cdot \varepsilon = Y^t \cdot Y + b^t \cdot X^t \cdot X \cdot b - 2 \cdot Y^t \cdot X \cdot b$$

$$b = \left[\left(X^t \cdot X \right)^{-1} \right] \cdot X^t \cdot Y$$

$$Y_{est} = X \cdot \left[\left(X^t \cdot X \right)^{-1} \right] \cdot X^t \cdot Y = H \cdot Y$$

Regresión múltiple (III)

Recordemos que, al igual que ocurría en la regresión simple los parámetros obtenidos son una estimación de los parámetros que se tienen en realidad. Se tendrá entonces que dicha estimación tendrá una valor esperado y una varianza

$$E[b] = \beta \qquad \mathbf{var}[b_i] = \left[\left(X^t \cdot X\right)^{-1}\right]_{ii} \cdot \sigma^2$$

Si no se conoce la varianza de los errores (que será el caso más habitual) hay que estimar dicha varianza mediante la siguiente expresión:

$$\sigma^2_{est} = \frac{\sum_{i=1}^{N} \left(y_i - y_{iest}\right)^2}{N - p - 1}$$

Al igual que en la regresión simple aquí se puede realizar un Análisis de Varianza (ANOVA) y se puede calcular el parámetro R^2. En la regresión múltiple se tiene otro índice conocido como R^2 residual que refleja el mayor/menor número de parámetros en el modelo así el ajuste obtenido; viene dado por

$$R^2_{adj} = 1 - \frac{SSE/(N-p-1)}{SST/(N-1)} = 1 - \left(\frac{N-1}{N-p-1}\right) \cdot \left(1 - R^2\right)$$

Al aumentar el número de parámetros R^2 ajustado puede aumentar O DISMINUIR (puede ser hasta negativo). Esto puede servir para escoger o no determinadas variables de entrada

Regresión múltiple (IV)

Multiple correlation coefficient.

Indicates that 36% of the variance can be predicted from the independent variables.

Model Summary[b]

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate
1	.654[a]	.427	.362	5.32327

ANOVA[b]

Model		Sum of Squares	df	Mean Square	F	Sig.
1	Regression	1290.267	7	184.324	6.505	.000[a]
	Residual	1728.571	61	28.337		
	Total	3018.838	68			

a. Predictors: (Constant), gender, pleasure scale, grades in h.s., Motivation scale, mother's education, Competence scale, father's education

b. Dependent Variable: math achievement test

Indicates that the combination of these variables significantly ($p < .001$) predicts the dependent variable.

Ejemplo obtenido en Use of Intermediate Statistics: Use and Interpretation, Lawrence & Erlbaum,

Regresión múltiple (V)

Aquí se plantean dos test estadísticos que usan diferente distribuciones de probabilidad según la hipótesis que se quiere comprobar.

$$H_0; \beta_0 = \beta_1 = ... = \beta_p = 0;$$

$$H_1; al\ menos\ existe\ un\ \beta_k \neq 0$$

Para esta hipótesis se determina el siguiente estadístico.

$$F_0 = \frac{SSR/(p)}{SSE/(N-p-1)} = \frac{R^2/p}{(1-R^2)/(N-p-1)}$$

El siguiente paso es fijar un nivel de significancia (α=0.05) y aceptar/rechazar la hipótesis nula según el siguiente criterio

$$F_o > F_{\alpha,p,n-p-1} \Rightarrow H_o\ se\ rechaza$$

Aquí la distribución a usar es la F.

Ahora lo que queremos comprobar es la hipótesis coeficiente a coeficiente

$$H_0; \beta_k = 0;$$

$$H_1; \beta_k \neq 0$$

Se calcula el siguiente factor

$$t_k = \frac{b_k}{\sqrt{var(b_k)}} \qquad var[b_k] = \left[\left(X^t \cdot X\right)^{-1}\right]_{kk} \cdot \sigma^2$$

$$\sigma^2_{est} = \frac{\sum_{i=1}^{N} (y_i - y_{iest})^2}{N-p-1}$$

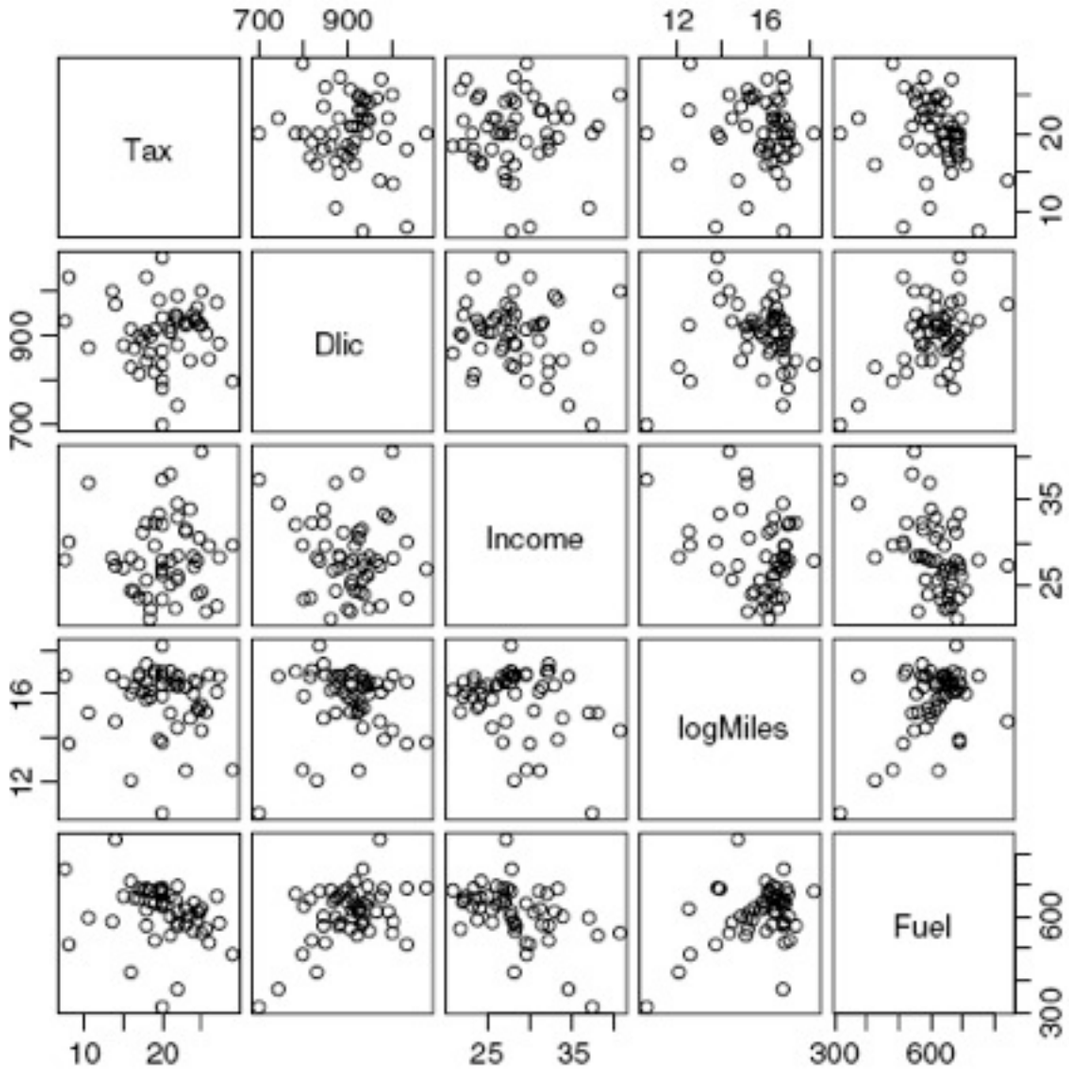
El siguiente paso es fijar un nivel de significancia (α=0.05) y aceptar/rechazar la hipótesis nula según el siguiente criterio

$$|t_k| > t_{\alpha/2,N-p-1}$$

Regresión múltiple (VI)

A la hora de desarrollar una regresión múltiple hay que tener en cuenta una serie de cuestiones sobre las suposiciones de partida

¿Existe la relación buscada y=g(x)?; usamos lo que se conoce como scatterplot; en este tipo de representación se tienen gráficas de la variables agrupadas por pares.



Applied Linear Regression, Wiley.

Otra opción es usar la matriz de correlaciones,

Sample Correlations

	Tax	Dlic	Income	logMiles	Fuel
Tax	1.0000	-0.0858	-0.0107	-0.0437	-0.2594
Dlic	-0.0858	1.0000	-0.1760	0.0306	0.4685
Income	-0.0107	-0.1760	1.0000	-0.2959	-0.4644
logMiles	-0.0437	0.0306	-0.2959	1.0000	0.4220
Fuel	-0.2594	0.4685	-0.4644	0.4220	1.0000

Sobre la selección de la variables en el modelo de regresión múltiple se tienen varias aproximaciones.

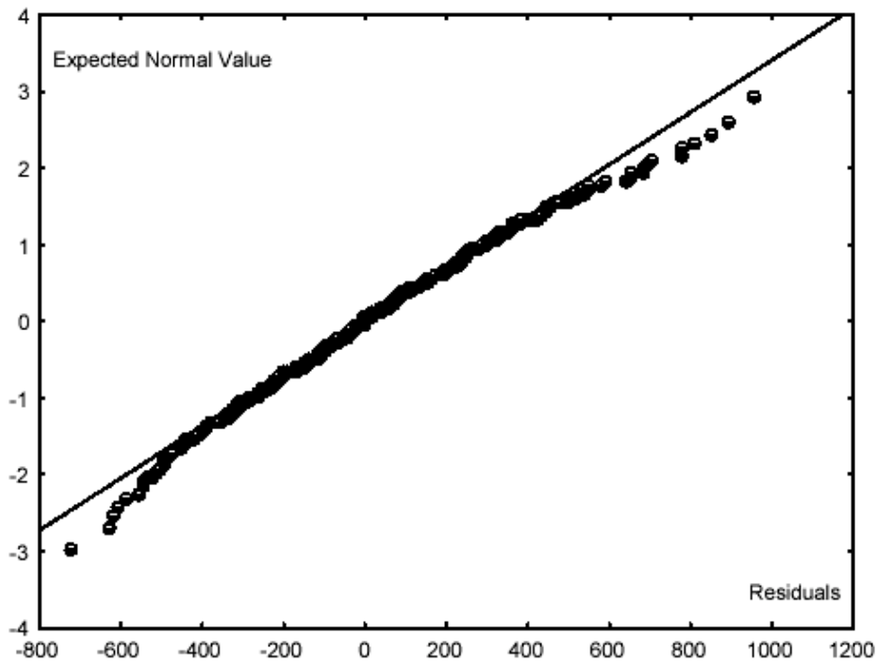
Forward Selection. Se considera primero un modelo simple con cada variable. Se escoge la que mejor ajuste ofrezca y se repite el procedimiento otra vez manteniendo esa variable. Así se continúa hasta no tener mejoras en el ajuste.

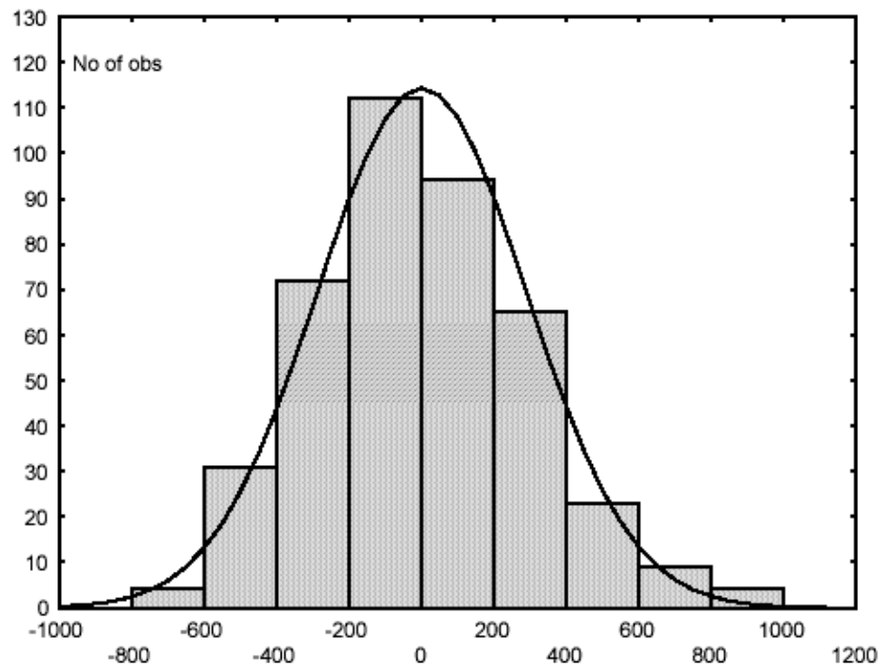
Backward Selection. Es el procedimiento inverso al anterior, se parte de un modelo con todas las variables y se van eliminando. Se detiene el proceso cuando no se tienen mejoras en el ajuste.

Stepwise Selection. Es una mezcla de los anteriores; Se comienza con un modelo simple como el primero y después de cada inclusión se plantea una eliminación de variables hasta que no se consigue una mejora apreciable.

Regresión múltiple (VII)

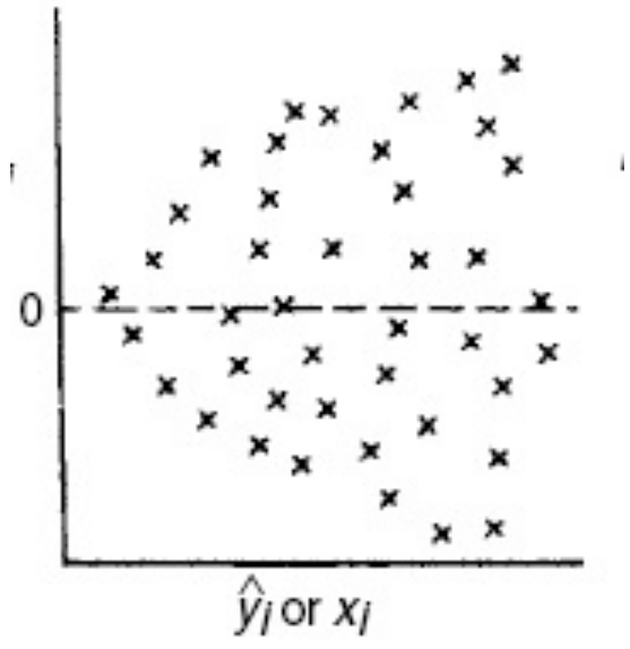
¿Los residuos son i.i.d según una distribución normal de varianza constante y valor medio cero?.
Se utiliza representaciones gráficas así como la determinación de los estadísticos para comprobar esta suposición





Residuals Statistics(a)

	Minimum	Maximum	Mean	Std. Deviation	N
Predicted Value	$13,966.34	$132,960.17	$34,419.57	$15,654.379	474
Residual	-$23,904.615	$47,558.465	$.000	$6,820.458	474
Std. Predicted Value	-1,307	6,295	,000	1,000	474
Std. Residual	-3,486	6,936	,000	,995	474
a Dependent Variable: Current Salary					



Applied Statistics Using SPSS, STATISTICA, MATLAB and R, Springer.

Applied Linear Regression, Wiley

Regresión múltiple (VIII)

Otro punto a tener en cuenta es el evitar altas correlaciones entre las variables de entrada. Este problema se conoce como ***colinealidad***. Este problema supone que pequeñas variaciones en los datos pueden provocar grandes cambios en los coeficientes del modelo.

Existen varias formas de comprobar este hecho. La primera opción es usando lo que se conoce como matriz de correlaciones. Sin embargo esta matriz no detecta de forma correcta si se tienen variables de entrada combinaciones lineales de otras.

Otro parámetro para detectar la colinealidad y que no tiene el problema de la matriz de correlaciones es el conocido como VIF (Variance Inflation Factor) definido como

$$VIF_k = \frac{1}{1 - R_k^2}$$

Aquí el índice de ajuste que aparece es el correspondiente al ajuste de la variable de entrada k usando el resto de variables de entrada. Se da como regla el eliminar aquella variable que tiene un valor superior a 10 de este parámetro.

Se conoce con este nombre porque la varianza de los coeficientes de la regresíon múltiple es propocional a este valor. Si el parámetro de ajuste R es próximo a 1 el factor VIF aumenta y, por tanto, la varianza de los coeficientes también aumenta

Regresión múltiple (VIII)

Coefficients[a]

Model		Unstandardized Coefficients		Standardized Coefficients	t	Sig.	Collinearity Statistics	
		B	Std. Error	Beta			Tolerance	VIF
1	(Constant)	-6.912	4.749		-1.455	.151		
	Motivation scale	1.639	1.233	.154	1.330	.188	.698	1.432
	Competence scale	1.424E-02	1.412	.001	.010	.992	.539	1.854
	pleasure scale	.953	1.119	.096	.852	.398	.746	1.340
	grades in h.s.	1.921	.480	.453	4.001	.000	.731	1.368
	father's education	.303	.331	.126	.915	.364	.497	2.013
	mother's education	.333	.406	.109	.820	.415	.529	1.892
	gender	-3.497	1.424	-.264	-2.455	.017	.814	1.228

a. Dependent Variable: math achievement test

Only *grades* and *gender* are significantly contributing to the equation. However, all of the variables need to be included to obtain this result, since the overall F value was computed with all the variables in the equation.

Tolerance and VIF give the same information. (Tolerance = 1/VIF) They tell us if there is multicollinearity. If the Tolerance value is low (< 1-R^2), then there is probably a problem with multicollinearity. In this case, since adjusted R^2 is .36, and 1- R^2 is about .64, then tolerances are low for *competence*, *mother's* and *father's*

This tells you how much each variable is contributing to any collinearity in the model.

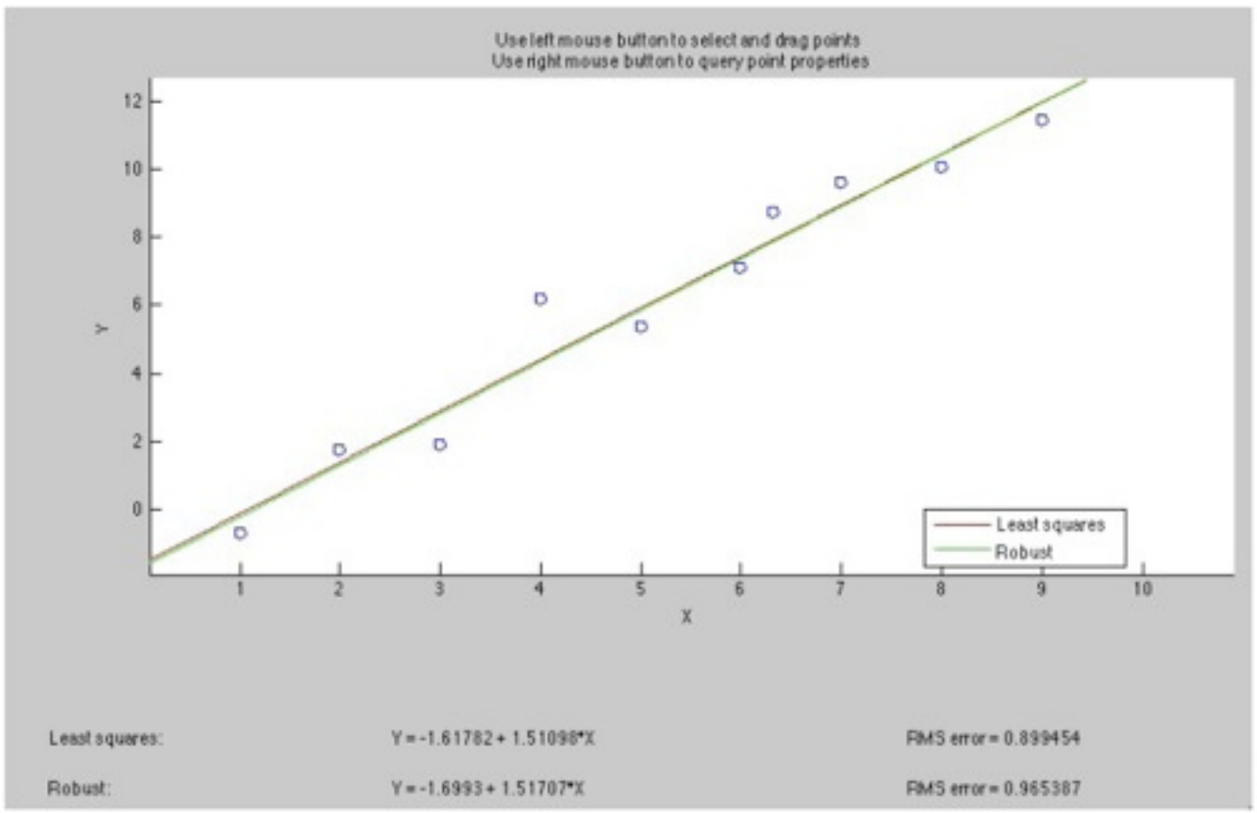
Collinearity Diagnostics[a]

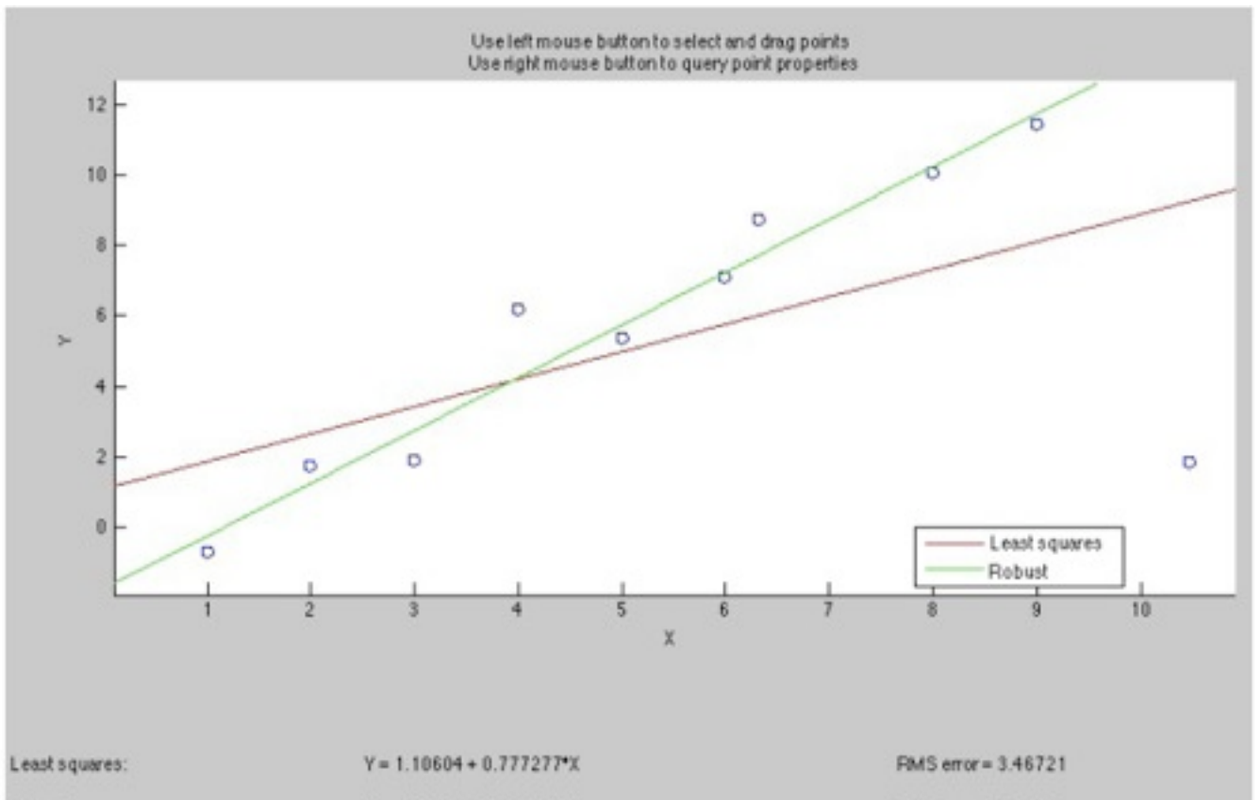
Model	Dimension	Eigenvalue	Condition Index	Variance Proportions							
				(Constant)	Motivation scale	Competence scale	pleasure scale	grades in h.s.	father's education	mother's education	gender
1	1	7.035	1.000	.00	.00	.00	.00	.00	.00	.00	.00
	2	.550	3.577	.00	.00	.00	.00	.00	.04	.02	.49
	3	.215	5.722	.00	.02	.01	.01	.00	.18	.09	.32
	4	8.635E-02	9.026	.00	.00	.00	.00	.06	.45	.78	.01
	5	5.650E-02	11.159	.00	.01	.00	.10	.60	.23	.04	.08
	6	2.911E-02	15.545	.01	.59	.00	.43	.05	.01	.00	.08
	7	1.528E-02	21.456	.70	.00	.46	.02	.02	.05	.04	.01
	8	1.290E-02	23.350	.29	.38	.53	.44	.28	.03	.02	.00

a. Dependent Variable: math achievement test

Ejemplo obtenido en Use of Intermediate Statistics: Use and Interpretation, Lawrence & Erlbaum,

Funciones de coste. Regresión robusta (I).





El problema de las regresiones analizadas (simple y múltiple) estriba en que se intenta minimizar la suma de los residuos al cuadrado. Este hecho provoca que los outliers tengan mucha importancia en el ajuste final.

Dos posibles soluciones a este problema son: a) definir una nueva función a minimizar (función de coste) y b) "pesar" cada uno de los términos que aparecen en la función minimizar ("weighted regression")

Funciones de coste. Regresión robusta (II).

Funciones de coste robustas.

La función que se ha planteado ha sido la siguiente:

$$J(n) = \frac{1}{2} \cdot e^2(n)$$

Las funciones robustas más usadas son:

$$J(n) = |e(n)|$$

$$J(n) = \log\{\cosh[e(n)]\}$$

$$J(n) = \begin{cases} e^2(n) & |e(n)| < c \\ c \cdot |e(n)| & |e(n)| \geq c \end{cases}$$

Todas estas funciones tienen en común que presentan bajos valores cuando la variable de entrada es muy grande en relación al valor que tomaría la función de coste cuadrática.

El problema de todas ellas es que no existe una solución directa y hay que aplicar procedimiento iterativos (REGLA DELTA).

Regresión robusta (residuos pesados)

Aquí el problema que se tiene es el impacto que tienen los outliers sobre el modelo final. Podemos reducir dicha influencia si le asignamos a esos datos un valor pequeño en la función final.

$$J = \sum_{i=1}^{N} w_i \cdot \{y_i - [b_0 + b_1 \cdot x_{1i} + \ldots\ldots b_p \cdot x_{pi}]\}^2$$

Esta ecuación puesta en forma matricial (aquí W es una matriz diagonal)

$$J = \varepsilon^t \cdot \varepsilon = [Y - X \cdot b]^t \cdot W \cdot [Y - X \cdot b]$$

Desarrollando se llega a

$$J = \varepsilon^t \cdot \varepsilon = Y^t \cdot W \cdot Y - 2 \cdot b^t \cdot X^t \cdot W \cdot Y + b^t \cdot X^t \cdot X \cdot b$$

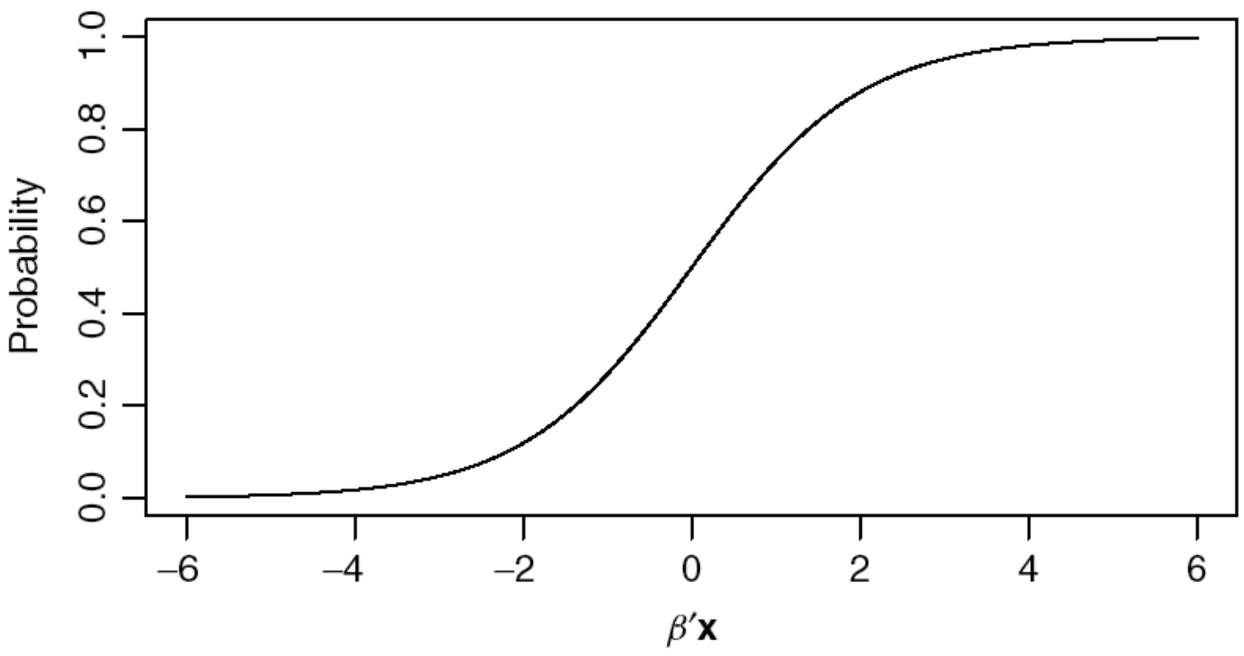
El mínimo de J es

$$b = \left[\left(X^t \cdot W \cdot X\right)^{-1}\right] \cdot X^t \cdot W \cdot Y$$

Regresión logística (I)

Hasta ahora se han desarrollado modelos para determinar el valor de una determinada variable continua (modelo de regresión); ¿qué modelo aplicamos cuando tenemos un problema de clasificación binaria?. Usamos entonces lo que se conoce como una regresión logística

$$y_i = \frac{1}{1 + e^{-\left[b_0 + b_1 \cdot x_{1i} + \ldots\ldots b_p \cdot x_{pi}\right]}}$$



Este modelo es una caso particular de lo que se conoce como MODELO LINEAL GENERALIZADO y consisten en la aplicación de una función no lineal a una regresión múltiple

$$y_i = g(\mu_i)$$

$$\mu_i = b_0 + b_1 \cdot x_{1i} + \ldots\ldots b_p \cdot x_{pi}$$

Si se hace la siguiente transformación

$$p_i = ln\left(\frac{y_i}{1 - y_i}\right)$$ → Odds

Se llega a la siguiente relación lineal

$$p_i = b_0 + b_1 \cdot x_{1i} + \ldots\ldots b_p \cdot x_{pi}$$

Podemos entonces aplicar todo lo aprendido de modelos lineales para la variable transformada.

Regresión logística (II)

Existe una forma de obtener la regresión logística de manera probabilística: supongamos que tenemos un problema donde queremos asignar el vector de entrada X a dos posibles clases A y B.

$$P(A \mid X) \Leftrightarrow P(B \mid X)$$

Aplicando Bayes.

$$P(A \mid X) = \frac{P(X \mid A) \cdot P(A)}{P(X \mid A) \cdot P(A) + P(X \mid B) \cdot P(B)}$$

$$P(A \mid X) = \frac{1}{1 + \frac{P(X \mid B) \cdot P(B)}{P(X \mid A) \cdot P(A)}}$$

Si las distribuciones condicionadas son gaussianas, esto es,

$$P(X \mid i) = K_1 \cdot e^{-\left(\frac{X-\mu_i}{\sigma_i^2}\right)^2}$$

Entonces, si las dos distribuciones TIENEN LA MISMA VARIANZA se llega la ecuación de una regresión logística para el valor de la probabilidad condicionada P(A|X). Aquí se tiene el significado de probabilidad que se le da a este modelo.

Hay que destacar que, si los dos clases no tuvieran la mima varianza (hecho que ocurre casi siempre) la aplicación de la regresión logística tiene poco sentido desde esta aproximación.

Regresión no lineal

La regresión logística es un claro ejemplo de que podemos realizar ajustes a funciones no lineales en las entradas pero sí lineales en los parámetros. Las siguientes funciones son ejemplos de este tipo de linealidad

$$Y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \beta_3 x_{i1} x_{i2} + \varepsilon_i$$

$$Y_i = \beta_0 + \beta_1 x_i + \beta_2 x_i^2 + \varepsilon_i \,.$$

$$Y_i = \beta_0 \exp(\beta_1 x_i) + \varepsilon_i$$

El problema que existe es que, a priori no conocemos el tipo de "mapeo" entre cada variable de entrada y la de salida.

La solución que se recurre más a menudo es establecer gráficas bidimensionales/ tridimensionales entre variables de entrada y la de salida manteniendo el resto de variables a un valor constante. Pero esto tiene un problema; dependiendo de ese valor podemos encontrar una relación u otra. A modo de ejemplo en la siguiente función

$$y_i = a_0 + a_1 \cdot (x_{1i} - 2) \cdot x_{2i} + a_2 \cdot (x_{3i} - 2)^2$$

Como no podemos representar las tres variables de entrada y la de salida fijamos por ejemplo x_1 a 2; ¿qué ocurre con x_2 ?. Y si tomamos x_3 igual a 2?.

Los modelos lineales se pueden obtener sin problemas sin embargo tienen serias limitaciones en cuanto a su capacidad de modelización de fenómenos como la saturación, histéresis, fenómenos de memoria, ec.

VNIVERSITAT DE VALÈNCIA

MASTER DE INGENIERÍA BIOMÉDICA.

Métodos de ayuda al diagnóstico clínico.

Tema 4: Modelos lineales.

Emilio Soria, Antonio José Serrano y José David Martín Dpto Ingeniería Electrónica, ETSE
Sistemas de Ayuda a la Decisión Clínica, Curso 2009-2010