Regresión Lineal Múltiple

El modelo, estimación de los parámetros, contrastes

Curso 2011-2012

Introducción

- Una extensión natural del modelo de regresión lineal simple consiste en considerar más de una variable explicativa.
- Los **modelo de regresión múltiple** estudian la relación entre
  - una variable de interés Y (variable respuesta o dependiente) y
  - un conjunto de variables explicativas o regresoras $X_1, X_2, \ldots, X_p$
- En el **modelo de regresión lineal múltiple** se supone que la función de regresión que relaciona la variable dependiente con las variables independientes es lineal, es decir:

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_p X_p + \varepsilon$$

Aplicaciones de la regresión múltiple

- Regresión múltiple como herramienta para predecir

Supongamos que estamos interesados en encontrar un hábitat adecuado para la familia de los escarabajos tigre (*Cicindela dorsalis dorsalis*), que viven en playas arenosas de la costa atlántica de Norteamérica.

Un posible procedimiento a seguir sería acudir a diferentes playas en las que habitase la especie y medir en ellas la densidad del escarabajo (Y) junto con distintos factores bióticos y abióticos (exposición al oleaje, tamaño del grano de arena, densidad de otros organismos,...)
Un modelo de regresión múltiple nos daría una ecuación para relacionar la densidad del escarabajo con el resto de variables, de modo que si acudimos a una nueva playa en la que no hay escarabajos y medimos el resto de factores podríamos predecir la densidad esperada de escarabajos al introducir la especie.[1]

[1] Handbook of Biological Statistics (http://udel.edu/~mcdonald/statintro.html)

Aplicaciones de la regresión múltiple

- Regresión múltiple como herramienta para detectar causalidad

La regresión múltiple también nos puede servir para entender la relación funcional entre la variable dependiente y las variables independientes y estudiar cuáles pueden ser las causas de la variación de Y.

Por ejemplo, si planteamos un modelo de regresión lineal simple que explique la densidad de escarabajo en función del tamaño de la arena, seguramente encontremos una relación significativa entre ambas variables. Y lo mismo si planteamos un modelo de regresión lineal simple que explique la densidad en función de la exposición al oleaje (pese a que seguramente el oleaje no sea el causante de los cambios en la densidad del escarabajo y lo que esté pasando es que la exposición al oleaje esté altamente correlacionada con el tamaño de la arena).
La regresión múltiple nos permite controlar este tipo de situaciones ya que podremos determinar si, manteniendo las mismas condiciones en el tamaño de arena, la exposición al oleaje realmente afecta a la densidad de la especie. [2]

[2] Handbook of Biological Statistics (http://udel.edu/~mcdonald/statintro.html)

Ejemplos

Una búsqueda rápida en artículos de revistas especializadas de Biología

How Microbial Community Composition Regulates Coral Disease Development.

J. Mao Jones, K. B. Ritchie, L. E. Jones, S. P. Ellner

PLoS Biology, Volume 8. (2003)

Abstract: Reef coral cover is in rapid decline worldwide, in part due to bleaching (expulsion of photosynthetic symbionts) and outbreaks of infectious disease. One important factor associated with bleaching and in disease transmission is a shift in the composition of the microbial community in the mucus layer surrounding the coral: the resident microbial community which is critical to the healthy functioning of the coral holobiont is replaced by pathogenic microbes, often species of Vibrio. In this paper we develop computational models for microbial community dynamics in the mucus layer in order to understand how the surface microbial community responds to changes in environmental conditions, and under what circumstances it becomes vulnerable to overgrowth by pathogens. Some of our models assumptions and parameter values are based on Vibrio spp. as a model system for other established and emerging coral pathogens. We find that the pattern of interactions in the surface microbial community facilitates the existence of alternate stable states, one dominated by antibiotic-producing beneficial microbes and the other pathogen dominated. A shift to pathogen dominance under transient stressful conditions, such as a brief warming spell, may persist long after environmental conditions have returned to normal. This prediction is consistent with experimental findings that antibiotic properties of Acropora palmata mucus did not return to normal long after temperatures had fallen. Long-term loss of antibiotic activity eliminates a critical component in coral defense against disease, giving pathogens an extended opportunity to infect and spread within the host, elevating the risk of coral bleaching, disease, and mortality.

Ejemplos

Una búsqueda rápida en artículos de revistas especializadas de Biología

Human Population Density and Extinction Risk in the World's Carnivores.

M. Cardillo, A. Purvis, W. Sechrest, J. L. Gittleman, J. Bielby, G. M. Mace

PLoS Biology, Volume 2. (2004)

Abstract: Understanding why some species are at high risk of extinction, while others remain relatively safe, is central to the development of a predictive conservation science. Recent studies have shown that a species' extinction risk may be determined by two types of factors: intrinsic biological traits and exposure to external anthropogenic threats. However, little is known about the relative and interacting effects of intrinsic and external variables on extinction risk. Using phylogenetic comparative methods, we show that extinction risk in the mammal order Carnivora is predicted more strongly by biology than exposure to high-density human populations. However, biology interacts with human population density to determine extinction risk: biological traits explain 80 % of variation in risk for carnivore species with high levels of exposure to human populations, compared to 45 % for carnivores generally. The results suggest that biology will become a more critical determinant of risk as human populations expand. We demonstrate how a model predicting extinction risk from biology can be combined with projected human population density to identify species likely to move most rapidly towards extinction by the year 2030. African viverrid species are particularly likely to become threatened, even though most are currently considered relatively safe. We suggest that a preemptive approach to species conservation is needed to identify and protect species that may not be threatened at present but may become so in the near future.

Ejemplos

Una búsqueda rápida en artículos de revistas especializadas de Biología

Selection for the compactness of highly expressed genes in *Gallus gallus*.

Y. S. Rao, Z. F. Wang, X. W. Chai, G. Z. Wu, M. Zhou, Q. H. Nie, X. Q. Zhang

Biology Direct, 5:35. (2010)

Abstract: Coding sequence (CDS) length, gene size, and intron length vary within a genome and among genomes. Previous studies in diverse organisms, including human, D. Melanogaster, C. elegans, S. cerevisiae, and Arabidopsis thaliana, indicated that there are negative relationships between expression level and gene size, CDS length as well as intron length. Different models such as selection for economy model, genomic design model, and mutational bias hypotheses have been proposed to explain such observation. The debate of which model is a superior one to explain the observation has not been settled down. The chicken (Gallus gallus) is an important model organism that bridges the evolutionary gap between mammals and other vertebrates. As D. Melanogaster, chicken has a larger effective population size, selection for chicken genome is expected to be more effective in increasing protein synthesis efficiency. Therefore, in this study the chicken was used as a model organism to elucidate the interaction between gene features and expression pattern upon selection pressure.

Ejemplos

Una búsqueda rápida en artículos de revistas especializadas de Biología

The evolution of egg colour and patterning in birds.

R. M. Kilner

Biological Reviews, 81. (2006)

Abstract: Avian eggs differ so much in their colour and patterning from species to species that any attempt to account for this diversity might initially seem doomed to failure. Here I present a critical review of the literature which, when combined with the results of some comparative analyses, suggests that just a few selective agents can explain much of the variation in egg appearance. Ancestrally, bird eggs were probably white and immaculate. Ancient diversification in nest location, and hence in the clutchs vulnerability to attack by predators, can explain basic differences between bird families in egg appearance. The ancestral white egg has been retained by species whose nests are safe from attack by predators, while those that have moved to a more vulnerable nest site are now more likely to lay brown eggs, covered in speckles, just as Wallace hypothesized more than a century ago. Even blue eggs might be cryptic in a subset of nests built in vegetation. It is possible that some species have subsequently turned these ancient adaptations to new functions, for example to signal female quality, to protect eggs from damaging solar radiation, or to add structural strength to shells when calcium is in short supply. The threat of predation, together with the use of varying nest sites, appears to have increased the diversity of egg colouring seen among species within families, and among clutches within species. Brood parasites and their hosts have probably secondarily influenced the diversity of egg appearance. Each drives the evolution of the others egg colour and patterning, as hosts attempt to avoid exploitation by rejecting odd-looking eggs from their nests, and parasites attempt to outwit their hosts by laying eggs that will escape detection. This co-evolutionary arms race has increased variation in egg appearance both within and between species, in parasites and in hosts, sometimes resulting in the evolution of egg colour polymorphisms. It has also reduced variation in egg appearance within host clutches, although the benefit thus gained by hosts is not clear.

Ejemplos

- Uno de los problemas que se tratan en disciplinas como la Ecología o la Biología de la Conservación es el de identificar factores que influyen en variables como la riqueza de una especie (medida como el número de individuos de la especie en un área dada)

Ejemplo: En un estudio sobre la población de un parásito se hizo un recuento de parásitos en 15 localizaciones con diversas condiciones ambientales.
Los datos obtenidos son los siguientes:

Temperatura	15	16	24	13	21	16	22	18	20	16	28	27	13	22	23
Humedad	70	65	71	64	84	86	72	84	71	75	84	79	80	76	88
Recuento	156	157	177	145	197	184	172	187	157	169	200	193	167	170	192

Regresión lineal múltiple

parasito.sav [Conjunto_de_datos6] - PASW Statistics Editor de datos

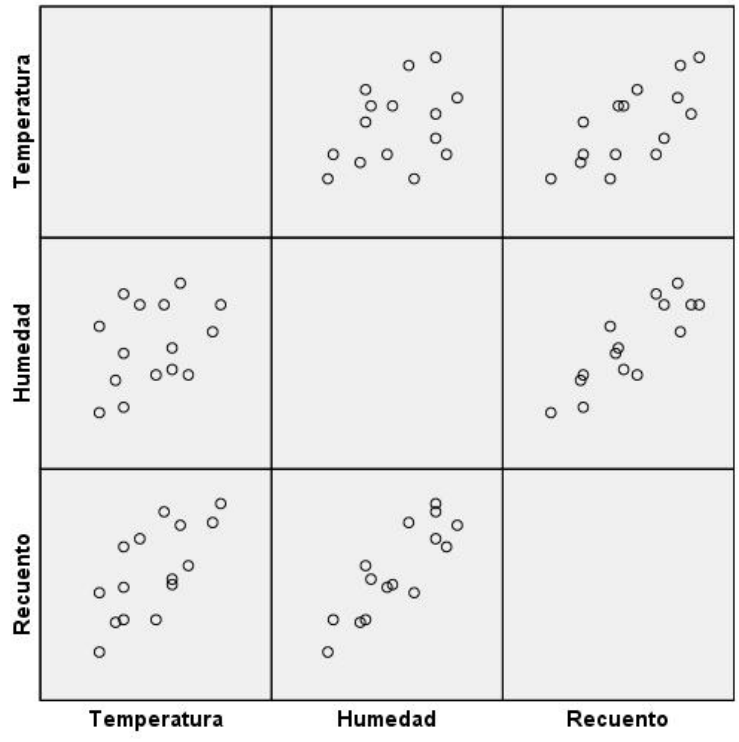
Archivo Edición Ver Datos Transformar Analizar Gráficos Utilidades Ventana Ayuda

Visible: 3 de 3 variables

	Temperatura	Humedad	Recuento	var	var
1	15,00	70,00	156,00		
2	16,00	65,00	157,00		
3	24,00	71,00	177,00		
4	13,00	64,00	145,00		
5	21,00	84,00	197,00		
6	16,00	86,00	184,00		
7	22,00	72,00	172,00		
8	18,00	84,00	187,00		
9	20,00	71,00	157,00		
10	16,00	75,00	169,00		
11	28,00	84,00	200,00		
12	27,00	79,00	193,00		
13	13,00	80,00	167,00		
14	22,00	76,00	170,00		
15	23,00	88,00	192,00		
16					
17					

Vista de datos | Vista de variables

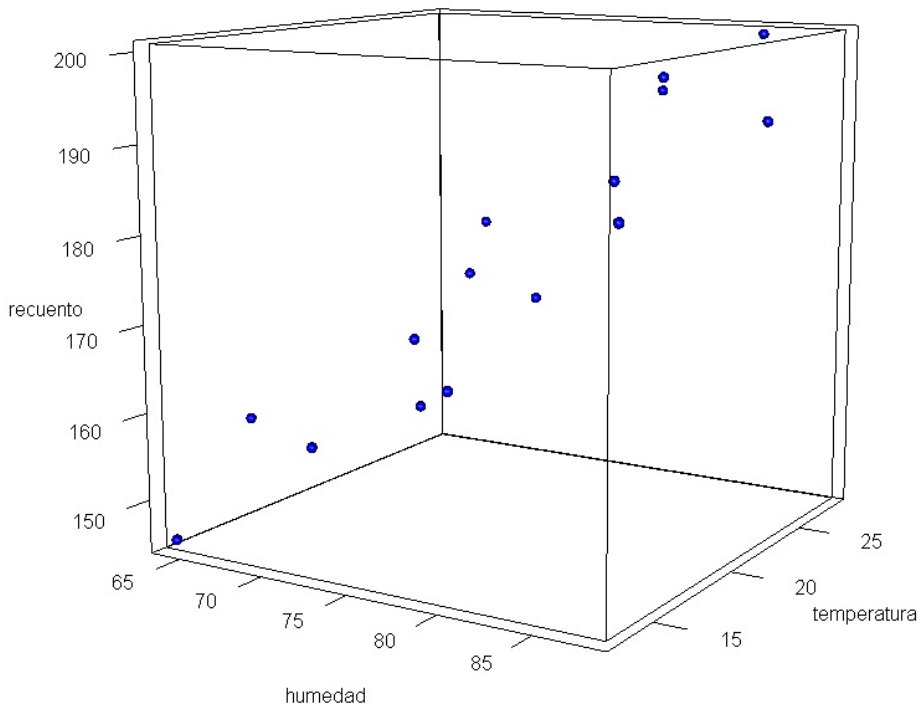
PASW Statistics Processor está listo



Regresión lineal múltiple

- Parece que la humedad y la temperatura son dos factores que afectan a la riqueza de la especie. ¿por qué no utilizamos toda la información que tenemos e intentamos explicar el comportamiento de la riqueza de parásitos a partir de ambas variables?

$$\text{Recuento} = \beta_0 + \beta_1 \text{Temperatura} + \beta_2 \text{Humedad} + \epsilon$$



Regresión lineal múltiple

Formulación del modelo

- En el **modelo de regresión lineal múltiple** se supone que la función de regresión que relaciona la variable dependiente con las variables independientes es lineal, es decir:

$$Y = \beta_0 + \beta_1 X_1 + \cdots + \beta_p X_p + \varepsilon$$

- β_0 es el término independiente. Es el valor esperado de Y cuando $X_1, \ldots, X_p$ son cero.
- $\beta_1, \beta_2, \ldots \beta_p$ son los coeficientes parciales de la regresión:
  - β_1 mide el cambio en Y por cada cambio unitario en X_1, manteniendo $X_2, X_3, \ldots, X_p$ constantes.
  - β_2 mide el cambio en Y por cada cambio unitario en X_2, manteniendo $X_1, X_3, \ldots, X_p$ constantes.
  - ...
  - β_p mide el cambio en Y por cada cambio unitario en X_p, manteniendo $X_1, \ldots, X_{p-1}$ constantes.
- ε es el error de observación debido a variables no controladas.

Regresión lineal múltiple

Modelo de regresión

$$Y = \beta_0 + \beta_1 X_1 + \cdots + \beta_p X_p + \varepsilon$$

Información muestral

$$\begin{array}{cccc} y_1 & x_{11} & \cdots & x_{1p} \\ y_2 & x_{21} & \cdots & x_{2p} \\ \vdots & \vdots & \ddots & \vdots \\ y_n & x_{n1} & \cdots & x_{np} \end{array}$$

De la expresión matemática del modelo de regresión lineal general se deduce

$$y_i = \beta_0 + \beta_1 x_{i1} + \cdots + \beta_p x_{ip} + \varepsilon_i, \quad i = 1, 2, \ldots, n$$

- Asumimos que los errores $\varepsilon_1, \varepsilon_2, \ldots, \varepsilon_n$ tienen distribución normal de media cero y varianza σ^2, y que son independientes.
- Las variables explicativas son linealmente independientes entre si.

Objetivo

Obtener a partir de la muestra estimadores:

- De los coeficientes $\hat{\beta}_0, \hat{\beta}_1, \ldots, \hat{\beta}_p$
- De la varianza del error $\hat{\sigma}^2$.

Regresión lineal múltiple

El modelo en forma matricial

- Podemos plantear el modelo en forma matricial de la siguiente manera:

$$\begin{pmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{pmatrix} = \begin{pmatrix} 1 & x_{11} & \cdots & x_{1p} \\ 1 & x_{21} & \cdots & x_{2p} \\ \vdots & \vdots & \ddots & \vdots \\ 1 & x_{n1} & \cdots & x_{np} \end{pmatrix} \cdot \begin{pmatrix} \beta_0 \\ \beta_1 \\ \vdots \\ \beta_p \end{pmatrix} + \begin{pmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \vdots \\ \varepsilon_n \end{pmatrix}$$

- Asignando la notación a las matrices respectivas, podríamos escribir la expresión anterior así:

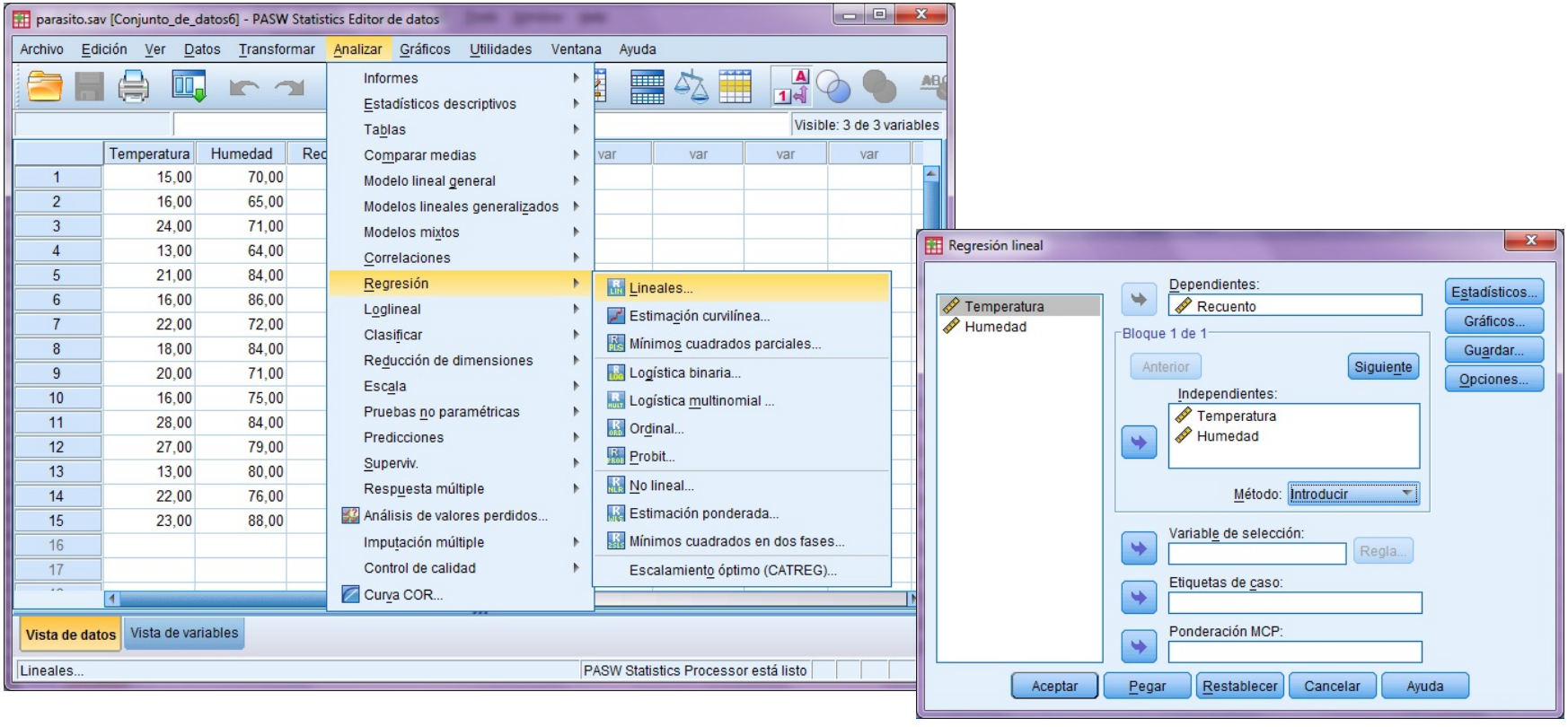
$$\mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \varepsilon$$

- Para estimar el vector de parámetros β podemos aplicar el método de mínimos cuadrados, igual que en el modelo lineal simple, y como resultado se obtiene el siguiente estimador:

$$\hat{\boldsymbol{\beta}} = \left(\mathbf{X}^t\mathbf{X}\right)^{-1}\mathbf{X}^t Y$$

donde $\mathbf{X}^t$ denota a la matriz transpuesta de $\mathbf{X}$.

Regresión lineal múltiple



Coeficientes[a]

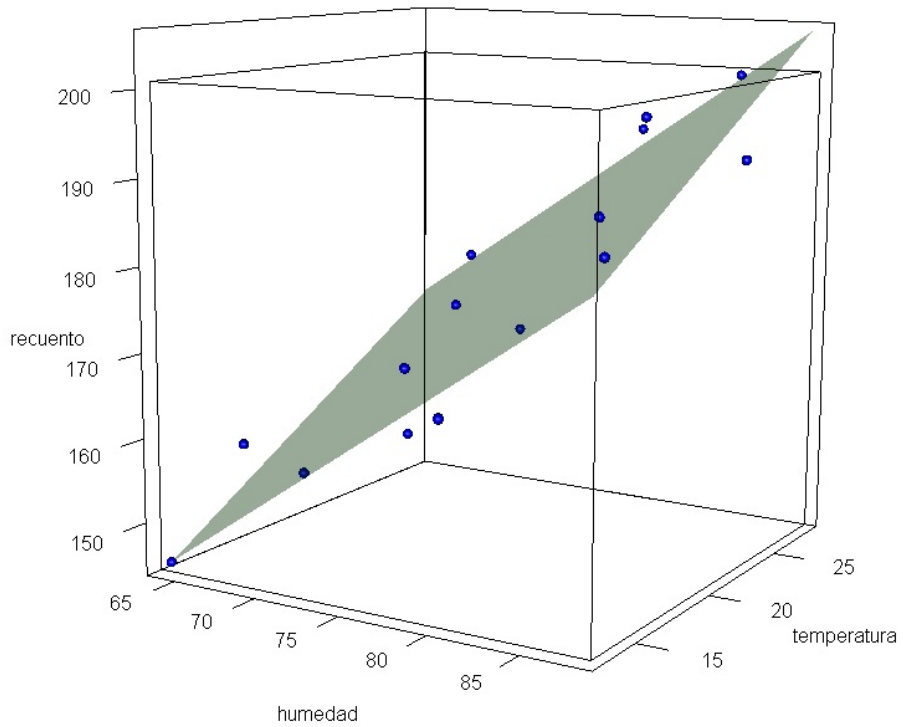
Modelo		Coeficientes no estandarizados		Coeficientes tipificados		
		B	Error típ.	Beta	t	Sig.
1	(Constante)	25,712	14,372		1,789	,099
	Temperatura	1,582	,320	,447	4,939	,000
	Humedad	1,542	,200	,700	7,731	,000

a. Variable dependiente: Recuento

Regresión lineal múltiple

Según el ajuste anterior:

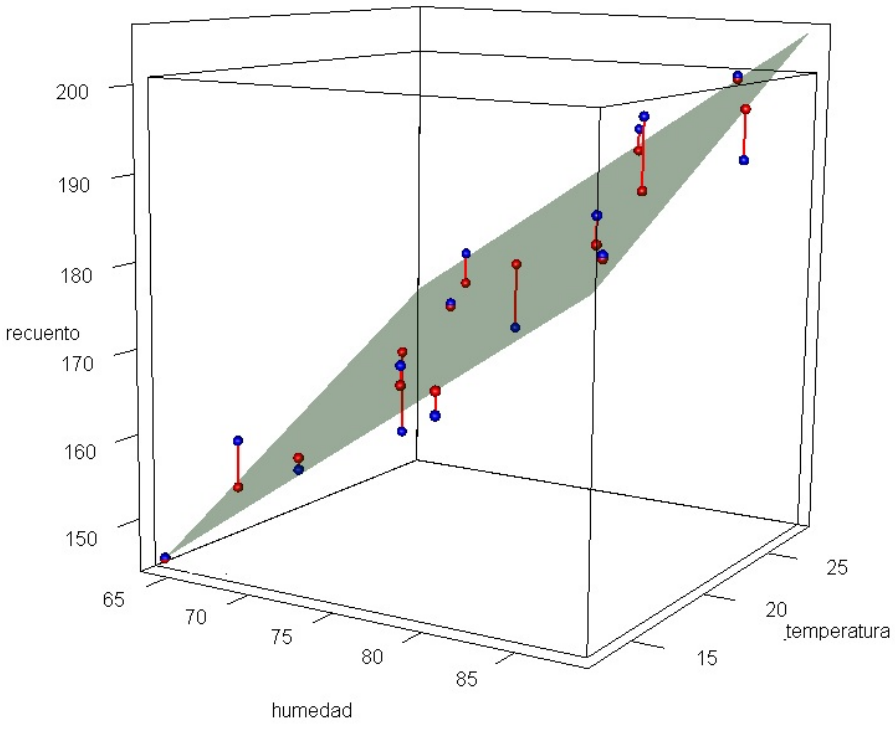
$$\text{Recuento} = 25.7115 + 1.5818\text{Temperatura} + 1.5424\text{Humedad}$$

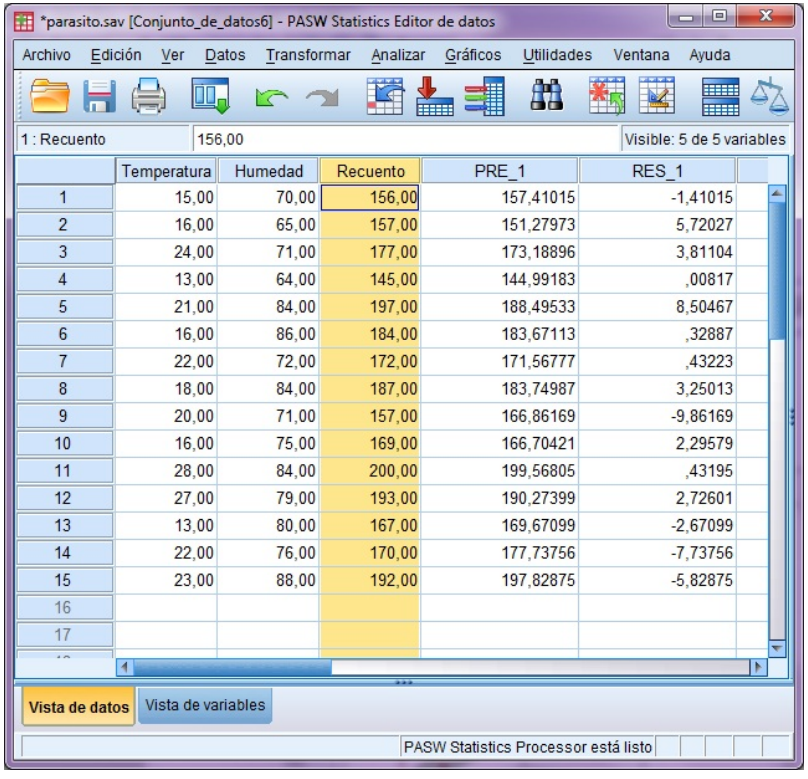


Regresión lineal múltiple

Según el ajuste anterior:

$$\text{Recuento} = 25.7115 + 1.5818\text{Temperatura} + 1.5424\text{Humedad}$$





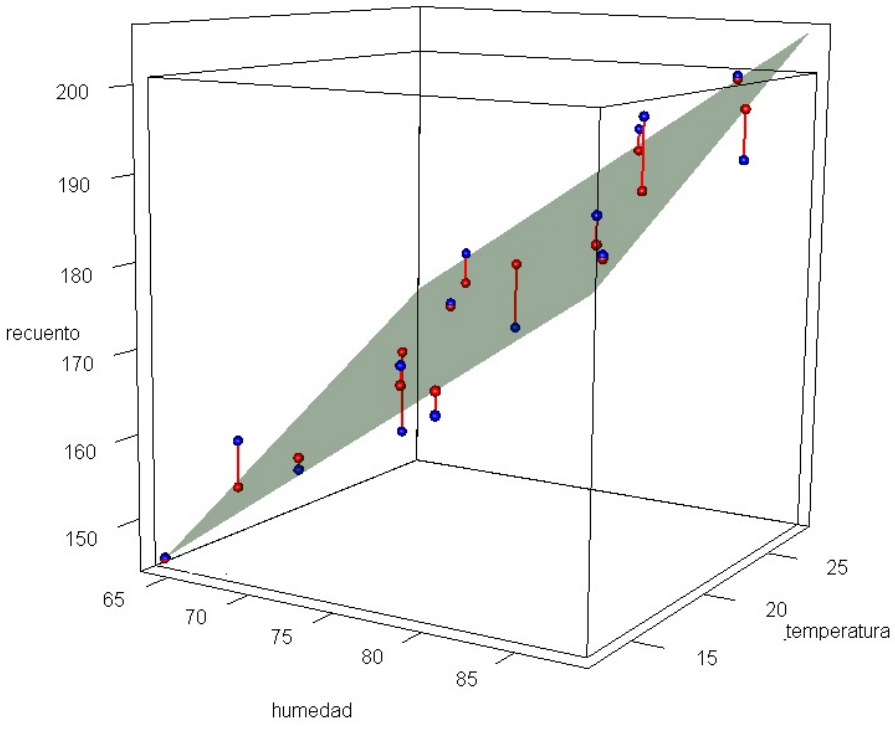
	Temperatura	Humedad	Recuento	PRE_1	RES_1
1	15,00	70,00	156,00	157,41015	-1,41015
2	16,00	65,00	157,00	151,27973	5,72027
3	24,00	71,00	177,00	173,18896	3,81104
4	13,00	64,00	145,00	144,99183	,00817
5	21,00	84,00	197,00	188,49533	8,50467
6	16,00	86,00	184,00	183,67113	,32887
7	22,00	72,00	172,00	171,56777	,43223
8	18,00	84,00	187,00	183,74987	3,25013
9	20,00	71,00	157,00	166,86169	-9,86169
10	16,00	75,00	169,00	166,70421	2,29579
11	28,00	84,00	200,00	199,56805	,43195
12	27,00	79,00	193,00	190,27399	2,72601
13	13,00	80,00	167,00	169,67099	-2,67099
14	22,00	76,00	170,00	177,73756	-7,73756
15	23,00	88,00	192,00	197,82875	-5,82875
16					
17					

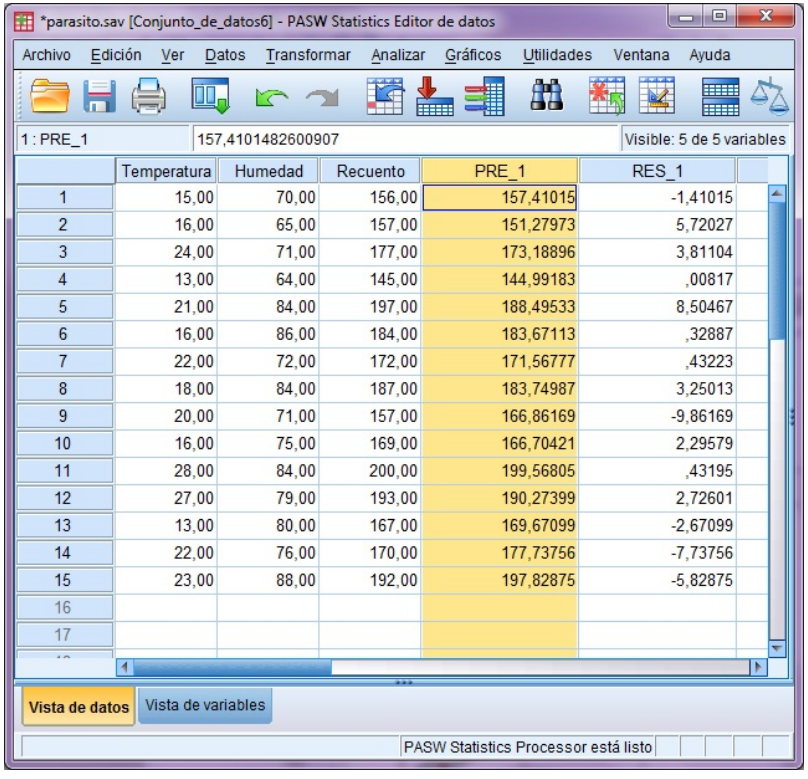
Valores observados y_i, $i = 1, \ldots, n$

Regresión lineal múltiple

Según el ajuste anterior:

$$\text{Recuento} = 25.7115 + 1.5818\text{Temperatura} + 1.5424\text{Humedad}$$





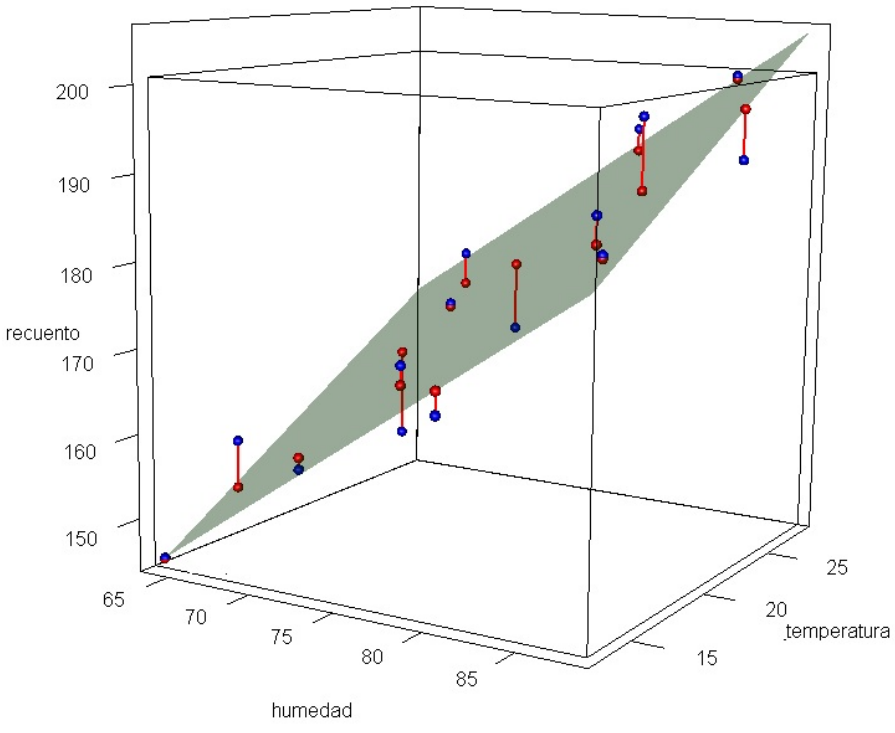
	Temperatura	Humedad	Recuento	PRE_1	RES_1
1	15,00	70,00	156,00	157,41015	-1,41015
2	16,00	65,00	157,00	151,27973	5,72027
3	24,00	71,00	177,00	173,18896	3,81104
4	13,00	64,00	145,00	144,99183	,00817
5	21,00	84,00	197,00	188,49533	8,50467
6	16,00	86,00	184,00	183,67113	,32887
7	22,00	72,00	172,00	171,56777	,43223
8	18,00	84,00	187,00	183,74987	3,25013
9	20,00	71,00	157,00	166,86169	-9,86169
10	16,00	75,00	169,00	166,70421	2,29579
11	28,00	84,00	200,00	199,56805	,43195
12	27,00	79,00	193,00	190,27399	2,72601
13	13,00	80,00	167,00	169,67099	-2,67099
14	22,00	76,00	170,00	177,73756	-7,73756
15	23,00	88,00	192,00	197,82875	-5,82875
16					
17					

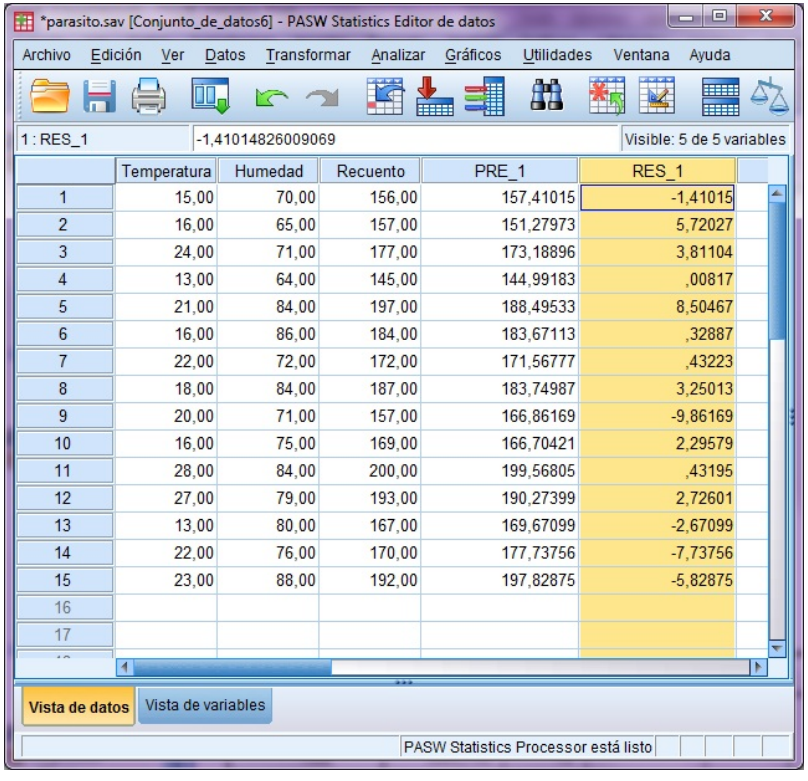
Valores ajustados $\hat{y}_i$, $i = 1, \ldots, n$

Regresión lineal múltiple

Según el ajuste anterior:

$$\text{Recuento} = 25.7115 + 1.5818\text{Temperatura} + 1.5424\text{Humedad}$$





	Temperatura	Humedad	Recuento	PRE_1	RES_1
1	15,00	70,00	156,00	157,41015	-1,41015
2	16,00	65,00	157,00	151,27973	5,72027
3	24,00	71,00	177,00	173,18896	3,81104
4	13,00	64,00	145,00	144,99183	,00817
5	21,00	84,00	197,00	188,49533	8,50467
6	16,00	86,00	184,00	183,67113	,32887
7	22,00	72,00	172,00	171,56777	,43223
8	18,00	84,00	187,00	183,74987	3,25013
9	20,00	71,00	157,00	166,86169	-9,86169
10	16,00	75,00	169,00	166,70421	2,29579
11	28,00	84,00	200,00	199,56805	,43195
12	27,00	79,00	193,00	190,27399	2,72601
13	13,00	80,00	167,00	169,67099	-2,67099
14	22,00	76,00	170,00	177,73756	-7,73756
15	23,00	88,00	192,00	197,82875	-5,82875
16					
17					

Residuos $\hat{\varepsilon}_i = y_i - \hat{y}_i$, $i = 1, \ldots, n$

Regresión lineal múltiple

Estimadores de los parámetros del modelo

- En resumen:

$$\hat{\boldsymbol{\beta}} = \left(\mathbf{X}^t\mathbf{X}\right)^{-1}\mathbf{X}^t Y$$

- Como estimador de la varianza del error se puede emplear:

$$\hat{\sigma}^2 = \frac{1}{n-(p+1)}\sum_{i=1}^{n}\hat{\varepsilon}_i^2 = \frac{1}{n-(p+1)}\sum_{i=1}^{n}(y_i-\hat{y}_i)^2$$

- En el ejemplo, $\hat{\sigma}^2 = \frac{343.542}{12} = 28.628$. Por tanto, $\hat{\sigma} = 5.35056$

Resumen del modelo[b]

Modelo	R	R cuadrado	R cuadrado corregida	Error típ. de la estimación
1	,956[a]	,914	,900	5,35056

a. Variables predictoras: (Constante), Humedad, Temperatura

b. Variable dependiente: Recuento

ANOVA[b]

Modelo		Suma de cuadrados	gl	Media cuadrática	F	Sig.
1	Regresión	3650,192	2	1825,096	63,751	,000[a]
	Residual	343,542	12	28,628		
	Total	3993,733	14			

a. Variables predictoras: (Constante), Humedad, Temperatura

b. Variable dependiente: Recuento

Regresión lineal múltiple

Descomposición de la variabilidad: Tabla ANOVA

- La variabilidad de toda la muestra se denomina **variabilidad total (VT)**.

$$\text{VT} = \sum_{i=1}^{n} (y_i - \bar{y})^2 .$$

- Al igual que en el modelo de regresión lineal simple, podemos descomponer la variabilidad total de Y en dos sumandos:
  - La variabilidad explicada (VE).

$$\text{VE} = \sum_{i=1}^{n} (\hat{y}_i - \bar{y})^2 .$$

  - La variabilidad no explicada (VNE) por la regresión.

$$\text{VNE} = \sum_{i=1}^{n} (y_i - \hat{y}_i)^2 .$$

Descomposición de la variabilidad

$$\text{VT} = \text{VE} + \text{VNE}.$$

Regresión lineal múltiple

Descomposición de la variabilidad: Tabla ANOVA

Fuente de variación	Suma de cuadrados	Grados de libertad
Regresión (VE)	$\sum_{i=1}^{n}(\hat{y}_i - \bar{y}_i)^2$	p
Residual (VNE)	$\sum_{i=1}^{n}(y_i - \hat{y}_i)^2$	$n-(p+1)$
Total (VT)	$\sum_{i=1}^{n}(y_i - \bar{y})^2$	$n-1$

Cuadro: Tabla ANOVA para el modelo de regresión lineal múltiple con constante, p variables explicativas y n observaciones

ANOVA[b]

Modelo		Suma de cuadrados	gl	Media cuadrática	F	Sig.
1	Regresión	3650,192	2	1825,096	63,751	,000[a]
	Residual	343,542	12	28,628		
	Total	3993,733	14			

a. Variables predictoras: (Constante), Humedad, Temperatura

b. Variable dependiente: Recuento

Regresión lineal múltiple

Variabilidad explicada

- El **coeficiente de determinación** (R^2) se define como la proporción de variabilidad de la variable dependiente que es explicada por la regresión

Coeficiente de determinación

$$R^2 = \frac{\text{VE}}{\text{VT}} = 1 - \frac{\text{VNE}}{\text{VT}}.$$

- El coeficiente de determinación presenta el inconveniente de aumentar siempre que aumenta el número de variables regresoras (algunas veces de forma artificial)
- Por ello y para penalizar el número de variables regresoras que se incluyen en el modelo de regresión, es conveniente utilizar el coeficiente de determinación corregido por el número de grados de libertad

Coeficiente de determinación ajustado

$$R^2 \ \textit{ajustado} = 1 - \frac{\text{VNE}/(n-(p+1))}{\text{VT}/(n-1)}$$

Resumen del modelo[b]

Modelo	R	R cuadrado	R cuadrado corregida	Error típ. de la estimación
1	,956[a]	,914	,900	5,35056

a. Variables predictoras: (Constante), Humedad, Temperatura

b. Variable dependiente: Recuento

Regresión lineal múltiple

Inferencia sobre los parámetros del modelo

- Los estimadores tienen el siguiente comportamiento probabilístico:

$$\hat{\beta} \in \mathrm{N}_{p+1}\left(\beta, (X^t X)^{-1}\sigma^2\right) \qquad \frac{(n-(p+1))\hat{\sigma}^2}{\sigma^2} \in \chi^2_{n-(p+1)}$$

  y además son independientes.

- En base a estas propiedades, se pueden efectuar las mismas tareas de inferencia que hemos realizado en el modelo lineal simple, salvo que en algunos casos aumenta la complejidad por el carácter multidimensional de los elementos del problema.

Regresión lineal múltiple

Contrastes de la regresión

- Suponiendo que se cumple el modelo de regresión lineal múltiple, estamos interesados en determinar si el modelo es o no explicativo.

$$H_0 : \beta_1 = \beta_2 = \ldots = \beta_p = 0$$
$$H_1 : \beta_j \neq 0 \text{ para algún } j = 1, \ldots, p$$

- Si se acepta la hipótesis nula ($\beta_1 = \beta_2 = \ldots = \beta_p = 0$), el modelo no es explicativo, es decir, ninguna de las variables explicativas influye en la variable respuesta Y.
- Si se rechaza la hipótesis nula, el modelo es explicativo, es decir, al menos una de las variables explicativas influye en la respuesta Y.
- Calculamos el estadístico

$$F = \frac{\frac{\text{VE}}{p}}{\frac{\text{VR}}{n-(p+1)}}$$

- Bajo la hipótesis nula ($\beta_1 = \beta_2 = \ldots = \beta_p = 0$) el estadístico F sigue una distribución $F_{p,n-(p+1)}$.

Regresión lineal múltiple

Contrastes de la regresión

- Suponiendo que se cumple el modelo de regresión lineal múltiple, estamos interesados en determinar si el modelo es o no explicativo.

$$H_0 : \beta_1 = \beta_2 = \ldots = \beta_p = 0$$
$$H_1 : \beta_j \neq 0 \text{ para algún } j = 1, \ldots, p$$

- Bajo la hipótesis nula ($\beta_1 = \beta_2 = \ldots = \beta_p = 0$) el estadístico F sigue una distribución $F_{p,n-(p+1)}$.

ANOVA[b]

Modelo		Suma de cuadrados	gl	Media cuadrática	F	Sig.
1	Regresión	3650,192	2	1825,096	63,751	,000[a]
	Residual	343,542	12	28,628		
	Total	3993,733	14			

a. Variables predictoras: (Constante), Humedad, Temperatura

b. Variable dependiente: Recuento

Regresión lineal múltiple

Contrastes de hipótesis individuales sobre los coeficientes

- Suponiendo que se cumple el modelo de regresión lineal múltiple, estamos interesados en determinar qué variables X_j son significativas para explicar la variable respuesta Y.

$$
\begin{aligned}
H_0 &: \beta_j = 0 \quad (X_j \text{ no influye sobre } Y) \\
H_1 &: \beta_j \neq 0 \quad (X_j \text{ influye sobre } Y)
\end{aligned}
$$

Regresión lineal múltiple

Contrastes de hipótesis individuales sobre los coeficientes (basados en la t de Student)

- Suponiendo que se cumple el modelo de regresión lineal múltiple, estamos interesados en determinar qué variables X_j son significativas para explicar la variable respuesta Y.

$$H_0 : \beta_j = 0 \quad (X_j \text{ no influye sobre } Y)$$
$$H_1 : \beta_j \neq 0 \quad (X_j \text{ influye sobre } Y)$$

- Para $j = 1, \ldots, p$:

$$\frac{\hat{\beta}_j - \beta_j}{\text{error típico de } \hat{\beta}_j} \sim t_{n-(p+1)}$$

Coeficientes[a]

Modelo		Coeficientes no estandarizados		Coeficientes tipificados		
		B	Error típ.	Beta	t	Sig.
1	(Constante)	25,712	14,372		1,789	,099
	Temperatura	1,582	,320	,447	4,939	,000
	Humedad	1,542	,200	,700	7,731	,000

a. Variable dependiente: Recuento

Regresión lineal múltiple

Contrastes de hipótesis individuales sobre los coeficientes (basados en la F de Snedecor)

- Suponiendo que se cumple el modelo de regresión lineal múltiple, estamos interesados en determinar qué variables X_j son significativas para explicar la variable respuesta Y.

$$\begin{aligned} H_0 &: \beta_j = 0 \quad (X_j \text{ no influye sobre } Y) \\ H_1 &: \beta_j \neq 0 \quad (X_j \text{ influye sobre } Y) \end{aligned}$$

- El contraste individual de la t de Student permite contrastar la influencia individual de la variable X_j
- Esta influencia también puede estudiarse por medio de una tabla ANOVA, analizando el incremento que se produce en la suma de cuadrados explicada por el modelo al introducir la variable regresora X_j

Regresión lineal múltiple

Contrastes de hipótesis individuales sobre los coeficientes (basados en la F de Snedecor)

- Si queremos contrastar la influencia de la variable X_j podemos:
  1. Ajustar el modelo de regresión completo, con X_1,..., X_p como variables regresoras y calcular la variabilidad explicada por el modelo VE(p)
  2. Ajusta el modelo de regresión con todas las variables excepto X_j y calcular la variabilidad explicada por este modelo VE($p-1$)
  3. Calcular la diferencia entre VE(p) $-$ VE($p-1$), que indica el aumento de la variabilidad explicada por el modelo al introducir la variable X_j
  4. Utilizar como estadístico de contraste

$$F = \frac{\frac{\text{VE}(p)-\text{VE}(p-1)}{1}}{\frac{\text{VR}(p)}{n-(p+1)}}$$

     donde VR(p) representa la variabilidad residual del modelo que incluye todas las variables regresoras
  5. Bajo la hipótesis nula ($\beta_j = 0$) el estadístico F sigue una distribución $F_{1,n-(p+1)}$

Regresión lineal múltiple

Contrastes de hipótesis individuales sobre los coeficientes (basados en la F de Snedecor)

- Este contraste proporciona exactamente el mismo resultado que el contraste individual de la t de Student (igual p-valor)
- Este método presenta la ventaja de poder utilizarse para contrastar la **influencia de un subconjunto de k variables explicativas**
- Por ejemplo, si queremos contrastar

$$H_0 : \beta_1 = \ldots = \beta_k = 0 \quad (X_1, \ldots X_k \text{ no influyen sobre } Y)$$
$$H_1 : \text{Alguno de los } \beta_j \neq 0, \quad j = 1, \ldots, k$$

calculamos el estadístico

$$F = \frac{\frac{\text{VE}(p) - \text{VE}(p-k)}{k}}{\frac{\text{VR}(p)}{n-(p+1)}}$$

- Bajo la hipótesis nula ($\beta_1 = \ldots = \beta_k = 0$) el estadístico F sigue una distribución $F_{k,n-(p+1)}$.