e-UMAB

MODELOS LINEALES

Francesc Carmona Pontaque

Electronic-University Mathematical Books

UNIVERSITAT DE BARCELONA

e-UMAB
Electronic-University Mathematical Books

Copia impresa del libro electrónico con ISBN: XX-XXXX-XXX-X
D.L.: B-XX.XXX-2004
Impresión: Gráficas Rey, S.L.
Impreso en España / Printed in Spain

A la meva esposa Carme
i els nostres fills Mireia i Guillem.

"Soñemos con un mundo unido
sin ninguna otra soberanía
que la del pueblo universal.
No hacer daño nunca, nunca, a nadie."

José María de Llanos (Padre Llanos)

Prólogo

Las páginas que siguen constituyen una parte de las exposiciones teóricas y prácticas de asignaturas que se han impartido a lo largo de algunos años en varias licenciaturas y cursos de doctorado. En particular en la licenciatura de Matemáticas, la licenciatura de Biología y la diplomatura de Estadística de la Universidad de Barcelona. Se ha intentado un cierto equilibrio entre las explicaciones teóricas y los problemas prácticos. Sin embargo, nuestra intención siempre ha sido fundamentar sólidamente la utilización de los modelos lineales como base de las aplicaciones de la regresión, el análisis de la varianza y el diseño de experimentos. Por ello, en este libro la base matemática y estadística es considerable y creemos importante la correcta definición de los conceptos y la rigurosidad de las demostraciones. Una sólida base impedirá cometer ciertos errores, habituales cuando se aplican los procedimientos ciegamente.

Por otra parte, la aplicación práctica de los métodos de regresión y análisis de la varianza requiere la manipulación de muchos datos, a veces en gran cantidad, y el cálculo de algunas fórmulas matriciales o simples. Para ello es absolutamente imprescindible la utilización de algún programa de ordenador que nos facilite el trabajo. En una primera instancia es posible utilizar cualquier programa de hojas de cálculo que resulta sumamente didáctico. También se puede utilizar un paquete estadístico que seguramente estará preparado para ofrecer los resultados de cualquier modelo lineal estándar como ocurre con el paquete SPSS. En cambio, en este libro se ha optado por incluir algunos ejemplos con el programa R. Las razones son varias. En primer lugar, se trata de un programa que utiliza el lenguaje S, está orientado a objetos, tiene algunos módulos específicos para los modelos lineales y es programable. R utiliza un lenguaje de instrucciones y al principio puede resultar un poco duro en su aprendizaje, sin embargo superada la primera etapa de adaptación, su utilización abre todo un mundo de posibilidades, no sólo en los modelos lineales, sino en todo cálculo estadístico. Además, la razón más poderosa es que el proyecto R es GNU y, por tanto, de libre distribución. De modo que los estudiantes pueden instalar en su casa el programa R y practicar cuanto quieran sin coste económico alguno. Por otra parte, el paquete S-PLUS es una versión comercial con el mismo conjunto de instrucciones básicas.

El tratamiento de algunos temas tiene su origen en unos apuntes de C.M. Cuadras y Pedro Sánchez Algarra (1996) que amablemente han cedido para su actualización en este libro y a los que agradezco profundamente su colaboración. También es evidente que algunas demostraciones tienen su origen en el clásico libro de Seber [66].

Por último, este libro ha sido escrito mediante el procesador de textos científico LaTeX y presentado en formato electrónico. Gracias a ello se puede actualizar con relativa facilidad. Se agradecerá la comunicación de cualquier errata, error o sugerencia.

Barcelona, 6 de mayo de 2004.

Dr. Francesc Carmona
fcarmona@ub.edu

Índice general

1

Las condiciones

1.1. Introducción

Los métodos de la Matemática que estudian los fenómenos deterministas relacionan, por lo general, una variable dependiente con diversas variables independientes. El problema se reduce entonces a resolver un sistema lineal, una ecuación diferencial, un sistema no lineal, etc.. Sin embargo, la aplicación de los métodos cuantitativos a las Ciencias Experimentales ha revelado la poca fiabilidad de las relaciones deterministas. En tales Ciencias, el azar, la aleatoriedad, la variabilidad individual, las variables no controladas, etc. justifican el planteo, en términos muy generales, de la ecuación fundamental

"observación" = "modelo" + "error aleatorio"

El experimentador puede, fijando las condiciones de su experimento, especificar la estructura del modelo, pero siempre debe tener en cuenta el error aleatorio o desviación entre lo que observa y lo que espera observar según el modelo.

Los modelos de regresión utilizan la ecuación anterior fijando el modelo como una función lineal de unos parámetros. El objetivo consiste, casi siempre, en la predicción de valores mediante el modelo ajustado.

El *Análisis de la Varianza* es un método estadístico introducido por R.A. Fisher de gran utilidad en las Ciencias Experimentales, que permite controlar diferentes variables cualitativas y cuantitativas (llamadas factores), a través de un modelo lineal, suponiendo normalidad para el error aleatorio. Fisher(1938) definió este método como "la separación de la varianza atribuible a un grupo de la varianza atribuible a otros grupos". Como veremos, los tests en Análisis de la Varianza se construyen mediante estimaciones independientes de la varianza del error.

Ambos conjuntos de modelos se pueden abordar con una teoría común: los modelos lineales.

Iniciaremos este capítulo con un ejemplo de modelización de un problema y su aplicación práctica. A continuación explicaremos en qué consiste esencialmente el método de los mínimos cuadrados y estableceremos las condiciones para que este método sea válido para su utilización en Estadística.

1.2. Un ejemplo

En el libro de Sen and Srivastava en [67, pág. 2] se explica este ejemplo que nosotros hemos adaptado a las medidas europeas.

Sabemos que cuantos más coches circulan por una carretera, menor es la velocidad del tráfico. El estudio de este problema tiene como objetivo la mejora del transporte y la reducción del tiempo de viaje.

La tabla adjunta proporciona los datos de la densidad (en vehículos por km) y su correspondiente velocidad (en km por hora).

Dato	Densidad	Velocidad	Dato	Densidad	Velocidad
1	12,7	62,4	13	18,3	51,2
2	17,0	50,7	14	19,1	50,8
3	66,0	17,1	15	16,5	54,7
4	50,0	25,9	16	22,2	46,5
5	87,8	12,4	17	18,6	46,3
6	81,4	13,4	18	66,0	16,9
7	75,6	13,7	19	60,3	19,8
8	66,2	17,9	20	56,0	21,2
9	81,1	13,8	21	66,3	18,3
10	62,8	17,9	22	61,7	18,0
11	77,0	15,8	23	66,6	16,6
12	89,6	12,6	24	67,8	18,3

Cuadro 1.1: Datos del problema de tráfico

Como la congestión afecta a la velocidad, estamos interesados en determinar el efecto de la densidad en la velocidad. Por razones que explicaremos más adelante (ver ejercicio 9.2), tomaremos como variable dependiente la raíz cuadrada de la velocidad.

El gráfico 1.1 presenta la nube de puntos o diagrama de dispersión (*scatter plot*) con la variable independiente (densidad) en el eje horizontal y la variable dependiente (raíz cuadrada de la velocidad) en el eje vertical.

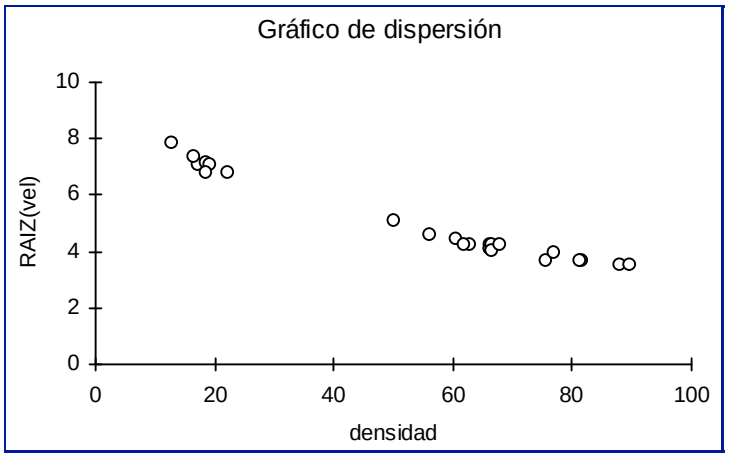

Figura 1.1: Nube de puntos del problema de tráfico

Como primera aproximación podríamos tomar, como modelo de ajuste, la recta que une dos puntos representativos, por ejemplo, los puntos $(12,7,\sqrt{62,4})$ y $(87,8,\sqrt{12,4})$. Dicha recta es $y = 8,6397 - 0,0583x$.

Inmediatamente nos proponemos hallar la mejor de las rectas, según algún criterio. Como veremos, el método de los mínimos cuadrados proporciona una recta, llamada recta de regresión, que goza de muy buenas propiedades. Este método consiste en hallar a y b tales que se minimice la suma de los errores al cuadrado.

$$\sum_{i=1}^{n} (y_i - (a + bx_i))^2$$

En este caso la recta de regresión es $y = 8,0898 - 0,0566x$.

Para estudiar la bondad del ajuste se utilizan los residuos

$$e_i = y_i - \hat{y}_i$$

donde $\hat{y}_i = 8,0898 - 0,0566x_i$. Los gráficos de la figura 1.2 nos muestran estos residuos.