

e-UMAB

MODELOS LINEALES

Francesc Carmona Pontaque

Electronic-University Mathematical Books



Consejo editor:

T. Aluja

M.J. Bayarri

E. Carmona

C.M. Cuadras (coordinador)

F.R. Fernández

J. Fortiana

G. Gómez

W. González-Manteiga

M.J. Greenacre

J.M. Oller

J. Puerto

A. Satorra

e-UMAB

Electronic-University Mathematical Books

© EDICIONS DE LA UNIVERSITAT DE BARCELONA, 2004

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	13
1.1. Introducción	13
1.2. Un ejemplo	13
1.3. El modelo	15
1.4. El método de los mínimos cuadrados	16
1.5. Las condiciones de Gauss-Markov	18
1.6. Otros tipos de modelos lineales	19
1.7. Algunas preguntas	19
1.8. Ejemplos con R	20
1.9. Ejercicios	22
2. Estimación	25
2.1. Introducción	25
2.2. El modelo lineal	25
2.3. Suposiciones básicas del modelo lineal	27
2.4. Estimación de los parámetros	28
2.5. Estimación de la varianza	33
2.6. Distribuciones de los estimadores	34
2.7. Matriz de diseño reducida	36
2.8. Matrices de diseño de rango no máximo	38
2.8.1. Reducción a un modelo de rango máximo	38
2.8.2. Imposición de restricciones	39
2.9. Ejercicios	39
3. Funciones paramétricas estimables	43
3.1. Introducción	43
3.2. Teorema de Gauss-Markov	45
3.3. Varianza de la estimación y multicolinealidad	48
3.4. Sistemas de funciones paramétricas estimables	49
3.5. Intervalos de confianza	52
3.6. Ejercicios	53

4. Complementos de estimación	57
4.1. Ampliar un modelo con más variables regresoras	57
4.1.1. Una variable extra	57
4.1.2. Una interpretación	59
4.1.3. Más variables	61
4.2. Mínimos cuadrados generalizados	62
4.3. Otros métodos de estimación	64
4.3.1. Estimación sesgada	64
4.3.2. Estimación robusta	65
4.3.3. Más posibilidades	66
4.4. Ejercicios	66
5. Contraste de hipótesis lineales	67
5.1. Hipótesis lineales contrastables	67
5.2. El modelo lineal de la hipótesis	68
5.3. Teorema fundamental del Análisis de la Varianza	71
5.3.1. Un contraste más general	76
5.3.2. Test de la razón de verosimilitud	78
5.4. Cuando el test es significativo	79
5.5. Contraste de hipótesis sobre funciones paramétricas estimables	79
5.6. Elección entre dos modelos lineales	80
5.6.1. Sobre los modelos	80
5.6.2. Contraste de modelos	81
5.7. Ejemplos con R	83
5.8. Ejercicios	84
6. Regresión lineal simple	89
6.1. Estimación de los coeficientes de regresión	89
6.2. Medidas de ajuste	92
6.3. Inferencia sobre los parámetros de regresión	94
6.3.1. Hipótesis sobre la pendiente	94
6.3.2. Hipótesis sobre el punto de intercepción	95
6.3.3. Intervalos de confianza para los parámetros	95
6.3.4. Intervalo para la respuesta media	96
6.3.5. Predicción de nuevas observaciones	96
6.3.6. Región de confianza y intervalos de confianza simultáneos	97
6.4. Regresión pasando por el origen	97
6.5. Correlación	98
6.6. Carácter lineal de la regresión simple	99
6.7. Comparación de rectas	102
6.7.1. Dos rectas	102
6.7.2. Varias rectas	106
6.7.3. Contraste para la igualdad de varianzas	109
6.8. Un ejemplo para la reflexión	110
6.9. Ejemplos con R	113
6.10. Ejercicios	115

7. Una recta resistente	119
7.1. Recta resistente de los tres grupos	119
7.1.1. Formación de los tres grupos	119
7.1.2. Pendiente e intercepción	120
7.1.3. Ajuste de los residuos e iteraciones	121
7.1.4. Mejora del método de ajuste	125
7.2. Métodos que dividen los datos en grupos	125
7.3. Métodos que ofrecen resistencia	127
7.4. Ejercicios	129
8. Regresión lineal múltiple	131
8.1. El modelo	131
8.2. Medidas de ajuste	132
8.3. Inferencia sobre los coeficientes de regresión	134
8.4. Coeficientes de regresión estandarizados	139
8.5. Multicolinealidad	142
8.6. Regresión polinómica	143
8.6.1. Polinomios ortogonales	145
8.6.2. Elección del grado	146
8.7. Comparación de curvas experimentales	148
8.7.1. Comparación global	148
8.7.2. Test de paralelismo	149
8.8. Ejemplos con R	150
8.9. Ejercicios	154
9. Diagnóstico del modelo	159
9.1. Residuos	159
9.1.1. Estandarización interna	159
9.1.2. Estandarización externa	161
9.1.3. Gráficos	162
9.2. Diagnóstico de la influencia	164
9.2.1. Nivel de un punto	164
9.2.2. Influencia en los coeficientes de regresión	165
9.2.3. Influencia en las predicciones	166
9.3. Selección de variables	167
9.3.1. Coeficiente de determinación ajustado	167
9.3.2. Criterio C_p de Mallows	168
9.3.3. Selección paso a paso	168
9.4. Ejemplos con R	168
9.5. Ejercicios	170
10. Regresión robusta	173
10.1. Minimizar una función objetivo	173
10.1.1. Funciones objetivo	174
10.2. Regresión robusta mínimo-cuadrada recortada	176

10.3. Ejemplos con S-PLUS	177
10.4. Ejercicios	181
11. Análisis de la Varianza	183
11.1. Introducción	183
11.2. Diseño de un factor	184
11.2.1. Comparación de medias	184
11.2.2. Un modelo equivalente	187
11.3. Diseño de dos factores sin interacción	190
11.4. Diseño de dos factores con interacción	196
11.5. Descomposición ortogonal de la variabilidad	201
11.5.1. Descomposición de la variabilidad en algunos diseños	203
11.5.2. Estimación de parámetros y cálculo del residuo	205
11.6. Diagnóstico del modelo	207
11.7. Diseños no balanceados y observaciones faltantes	210
11.8. Ejemplos con R	211
11.9. Ejercicios	217
12. Análisis de Componentes de la Varianza	221
12.1. Introducción	221
12.2. Contraste de hipótesis	222
12.2.1. Los test F	224
12.2.2. Estimación de los componentes de la varianza	225
12.3. Los modelos más sencillos	226
12.3.1. Diseño de un factor con efectos fijos	226
12.3.2. Diseño de un factor con efectos aleatorios	229
12.3.3. Diseño de dos factores sin interacción con efectos fijos	233
12.3.4. Diseño de dos factores sin interacción con efectos aleatorios	236
12.3.5. Diseño de dos factores aleatorios con interacción	238
12.3.6. Diseño de tres factores aleatorios y réplicas	239
12.3.7. Diseño anidado de dos factores aleatorios	240
12.3.8. Resumen	242
12.4. Correlación intraclásica	243
12.5. Ejemplos con R	244
12.6. Ejercicios	245
A. Matrices	249
A.1. Inversa generalizada	249
A.2. Derivación matricial	250
A.3. Matrices idempotentes	250
A.4. Matrices mal condicionadas	251
B. Proyecciones ortogonales	253
B.1. Descomposición ortogonal de vectores	253
B.2. Proyecciones en subespacios	255

C. Estadística multivariante	257
C.1. Esperanza, varianza y covarianza	257
C.2. Normal multivariante	258
Bibliografía	259
Índice alfabético	263



Las condiciones

I.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

$$\text{“observación”} = \text{“modelo”} + \text{“error aleatorio”}$$

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.

I.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.

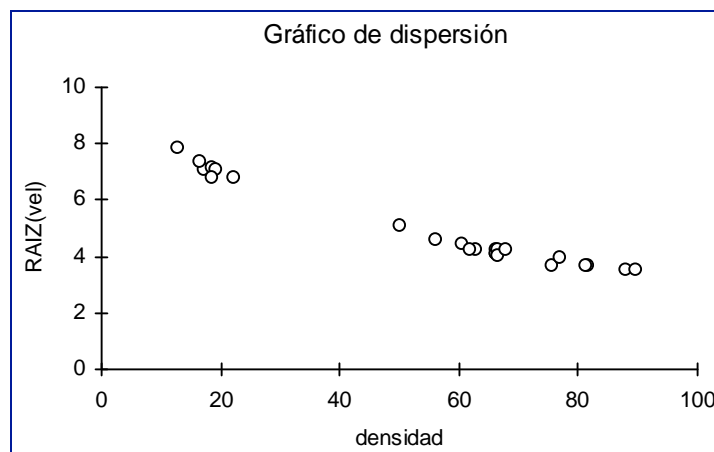


Figura 1.1: Nube de puntos del problema de tráfico

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.