IV Congreso de Economía Regional de Castilla y León Burgos 1994, Comunicaciones 2: 1146-1155

ANALISIS COMPARATIVO DE LAS PROVINCIAS DE CASTILLA Y LEON SEGUN LOS GRUPOS DE GASTO DE LA ENCUESTA DE PRESUPUESTOS FAMILIARES

Ramón ALVAREZ ESTEBAN
María GOMEZ RIOCEREZO
Carmen HUERGA CASTRO
Mª Jesús MURES QUINTANA
Universidad de León

1. INTRODUCCIÓN

El objetivo que nos proponemos con este trabajo consiste en analizar el gasto en bienes y servicios en las provincias de la Comunidad de Castilla y León a partir de la información suministrada por la Encuesta de Presupuestos Familiares correspondiente al año 1991. Concretamente, pretendemos determinar la existencia o no de diferencias significativas respecto a los gastos medios entre las nueve provincias de nuestra comunidad. También estudiaremos la posibilidad de establecer algún tipo de relación funcional entre los distintos grupos de gasto que aparecen desglosados en la Encuesta de Presupuestos Familiares.

La base metodológica más adecuada, a nuestro entender, para abordar los dos objetivos señalados son el Análisis de la varianza y el análisis de regresión múltiple, respectivamente.

El esquema que seguiremos en el desarrollo de este trabajo comienza exponiendo las características básicas de las técnicas empleadas para continuar con el análisis de los datos. Una vez tratados éstos con el soporte informático adecuado (SPSS/PC) especificaremos los resultados y presentaremos las conclusiones.

2. METODOLOGÍA

En el análisis de los datos vamos a utilizar dos técnicas estadísticas específicas dado que las hipótesis y características de las mismas se ajustan a la situación concreta que tratamos. Son el Análisis de la Varianza y el Análisis de Regresión Múltiple. Puesto que en el desarrollo del análisis de regresión se utiliza el análisis de la varianza, comenzaremos con la exposición de esta última.

2.1. Análisis de la Varianza

El análisis de la varianza pretende descomponer la variabilidad de una variable en componentes distintas que pueden asignarse a causas diferentes. La variable que se desea medir se denomina variable respuesta y los factores son las variables que pueden influir en la misma. Estos factores (variables independientes) suelen estar bajo el control del investigador y pueden ser cuantitativos o cualitativos.

Se definen los niveles o tratamientos de un factor como las condiciones de éste bajo el cual se observará la respuesta. Hoy en día el término tratamiento se utiliza generalmente para referirse a las diferentes clasificaciones del factor como por ejemplo zonas o regiones de un país.

Si en un experimento intervienen varios factores, el tratamiento es una combinación de los niveles de cada factor. La unidad experimental se considera como aquella que proporciona una respuesta después de aplicar un tratamiento. Generalmente existen factores externos difíciles de controlar por el investigador y que repercuten en la respuesta produciendo una variación no atribuible a un cambio de tratamiento. Esta situación se presenta frecuentemente en las ciencias socioeconómicas donde las desviaciones de las condiciones experimentales son muy comunes ya que el entorno económico no permite un control suficiente. En estos casos sólo es posible registrar la información a medida que se produce en el mundo real.

Otro aspecto importante se centra en la forma de asignar los tratamientos a cada unidad experimental. Si esta asignación se lleva a cabo de forma aleatoria y las unidades se suponen homogéneas se dice que estamos ante un diseño completamente aleatorio. En términos generales el uso de este tipo de diseño implica que las condiciones bajo las que se observa la respuesta se mantienen a lo largo del experimento.

Los métodos que incluye el análisis de la varianza son múltiples lo que conlleva que aquí sólo nos detengamos en aquellos más adecuados a nuestra situación. Vamos a considerar en principio el modelo más sencillo: una variable respuesta Y de la que se desea analizar su variación respecto a un factor con $k \geq 2$ niveles o tratamientos distintos. Según la forma de obtener los tratamientos del factor se distinguen el modelo de efectos fijos (el investigador selecciona los tratamientos) y el modelo de efectos aleatorios (los tratamientos se eligen de forma aleatoria entre todas las posibilidades del factor). Trataremos sólo el modelo de efectos fijos.

Los datos se ordenan como se muestra en la tabla siguiente:

	Tratamientos						
	1	2	...	j	...	k	
	y_{11}	y_{12}	...	y_{1j}	...	y_{1k}	
	y_{21}	y_{22}	...	y_{2j}	...	y_{2k}	
	.	.	...	.	...	.	
	y_{i1}	y_{i2}	...	y_{ij}	...	y_{ik}	
	.	.	...	.	...	.	
	y_{n1}	y_{n2}	...	y_{nj}	...	y_{nk}	
Total	$T_{.1}$	$T_{.2}$	...	$T_{.j}$	...	$Y_{.k}$	$T_{..}$
Media	$\bar{y}_{.1}$	$\bar{y}_{.2}$	...	$\bar{y}_{.}$	...	$\bar{y}_{.k}$	$\bar{y}_{..}$

Cada y_{ij} representa la i-ésima observación del tratamiento j. En este caso obtenemos muestras de tamaño n de la variable respuesta para cada tratamiento pero estos tamaños pueden ser distintos. Nos interesa comprobar si la variable respuesta se ve afectada por el factor, es decir, si llamamos μ_j a la media de Y en el tratamiento j debemos contrastar:

H_0: $\mu_1=\ldots=\mu_k$ (la media de Y es igual independientemente del tratamiento)

H_1: al menos dos de las medias no son iguales

Ahora bien, cada observación puede escribirse en la forma:

$$y_{ij} = \mu + \alpha_j + \varepsilon_{ij} \qquad [1]$$

siendo μ la media global de todas las k poblaciones, α_j el efecto del j-ésimo tratamiento y ε_{ij} el error experimental para la i-ésima observación bajo el tratamiento j. Como el modelo es de efectos fijos:

$$\sum_{j=1}^{k} \alpha_j = 0$$

Además se deben verificar dos condiciones básicas: la varianza dentro de cada tratamiento es constante (homocedasticidad) y las ε_{ij} son variables aleatorias independientes y $N(0,\sigma^2)$. Entonces la hipótesis nula propuesta anteriormente se puede reemplazar por:

H_0: $\alpha_1 = \alpha_2 = \ldots = \alpha_k = 0$

H_1: al menos una de las α_j no es igual a cero

De este modo H_0 establece que no existe ningún efecto de los tratamientos sobre la respuesta, lo que a su vez implica que las k medias poblacionales son iguales entre sí y cada observación consiste en una media común más un error aleatorio.

El modelo expresado en [1] indica que cualquier desviación de una observación respecto a la media global se puede deber a dos causas: a la diferencia en el tratamiento o a un error aleatorio, ya que se puede escribir como:

$$y_{ij} = \mu + (\mu_j-\mu) + (y_{ij}-\mu_j) \Leftrightarrow y_{ij} - \mu = (\mu_j-\mu) + (y_{ij}-\mu_j)$$

Después de una serie de transformaciones puramente matemáticas se llega a la ecuación fundamental del análisis de la varianza:

$$\sum_{j=1}^{k}\sum_{i=1}^{n}(y_{ij} - \overline{y}_{..})^2 = n\sum_{j=1}^{k}(\overline{y}_{.j} - \overline{y}_{..})^2 + \sum_{j=1}^{k}\sum_{i=1}^{n}(y_{ij} - \overline{y}_{.j})^2$$

SCT = SCTR + SCE

Suma total de cuadrados = suma de cuadrados de los tratamientos + suma de cuadrados de los errores

Se comprueba que $SCTR/\sigma^2$ y SCE/σ^2 son dos variables aleatorias independientes con distribuciones chi-cuadrado con (k-1) y (N-k) grados de libertad respectivamente y que el estadístico:

$$F = \frac{\dfrac{SCTR / \sigma^2}{k-1}}{\dfrac{SCE / \sigma^2}{N-k}} = \frac{SCTR / (k-1)}{SCE / (N-k)}$$

sigue una distribución F de Snedecor con k-1 y N-k grados de libertad que permite contrastar la hipótesis nula propuesta.

Toda la información para el análisis de la varianza se resume en la siguiente tabla:

Fuente de variación	Suma de cuadrados	Grados de libertad	Cuadrados medios	Estadístico F
Tratamiento	SCTR	k-1	SCTR/(k-1)=CMTR	CMTR/SME=F
Error	SCE	N-k	SCE/(N-k)=SME	
Total	SCT	N-1		

Notemos que si el estadístico F permite rechazar la hipótesis nula no es posible especificar qué medias son diferentes. Se necesita un estudio más completo para analizar las diferencias entre las medias. Se han propuesto varios métodos: rangos estudentizados de Tuckey, rangos múltiples de Duncan y el método de Scheffé.

Existe otro modelo importante en el estudio del análisis de la varianza: el diseño en bloques completamente aleatorizados. Este caso se presenta cuando consideramos un factor sobre el cual no queremos obtener información sino tan solo conseguir que no afecte al experimento (variable bloque). Suponiendo entonces un factor con k tratamientos y n valores de la variable bloque, el modelo establecido es:

$$y_{ij} = \mu + \beta_i + \alpha_j + \varepsilon_{ij} \qquad i=1...n \quad j=1...k$$

donde y_{ij} es la respuesta del i-ésimo bloque bajo el j-ésimo tratamiento, μ la media global, β_i es afecto del i-ésimo bloque, α_j el efecto del j-ésimo tratamiento y ε_{ij} el error aleatorio. De nuevo las hipótesis básicas son las mismas exigiendo además que los bloques y los tratamientos no interactúen entre sí.

En la tabla de datos se debe añadir una columna a la izquierda donde aparezcan los n bloques. la hipótesis nula que se plantea sigue siendo la misma: H_0: $\alpha_j=0$ j=1...k, pero ahora en la ecuación fundamental del análisis de la varianza aparece otro sumando que representa la suma de los cuadrados debida a los bloques: SCB. En la tabla ANOVA correspondiente se incluye otra fuente de variación (la de los bloques) con (n-1) grados de libertad y el estadístico F para el contraste tiene ahora (k-1) y (n-1)(k-1) grados de libertad.

2.2. Análisis de Regresión Múltiple

El análisis de regresión múltiple, por su parte, trata de expresar la relación entre una variable dependiente Y y un conjunto de variables explicativas independientes $X_1, X_2...X_k$ a través de un modelo adecuado. El caso más sencillo en que la relación planteada es lineal se expresa de la forma siguiente:

$$Y_i = \beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + ... + \beta_k X_{ik} + \varepsilon_i \qquad i=1...n$$

Y_i es la i-ésima observación de la respuesta para un conjunto de valores fijos $X_{i1}...X_{ik}$ de las variables explicativas; ε_i es el error aleatorio no observable asociado con Y_i y $\beta_0....\beta_k$ son los parámetros desconocidos que habrá que determinar.

Cada parámetro β_j, j=1...k representa el cambio en la respuesta promedio para un cambio igual a una unidad de la correspondiente variable de predicción X_j, cuando el resto de ellas permanecen constantes.

Los supuestos básicos del modelo planteado son los siguientes:

1. Las variables X_j, j=1...k no son variables aleatorias y no existe relación lineal entre dos o más variables independientes.
2. Los errores correspondientes a distintas observaciones están incorreladas.
3. Los errores aleatorios ε_i son $N(0,\sigma^2)$ independientes.

De los supuestos anteriores se deduce que las Y_i son variables aleatorias independientes normalmente distribuidas con:

$$E(Y_i) = \beta_0 + \beta_1 X_{i1} + \ldots + \beta_k X_{ik}$$
$$Var(Y_i) = \sigma^2 \qquad \forall i=1\ldots n$$

Para obtener los estimadores de los parámetros $\beta_0 \ldots \beta_k$ se utiliza el criterio de los mínimos cuadrados:

$$\text{mín SCE} = \text{mín} \sum (Y - \hat{Y})^2$$

donde $\hat{Y}$ expresa el modelo estimado.

En su desarrollo resulta más sencillo presentar las ecuaciones en forma matricial escribiendo el modelo de la forma:

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

El vector de parámetros estimados de la regresión y la estimación (insesgada) de la varianza del error son respectivamente:

$$\hat{\beta} = \left[X^T X\right]^{-1} X^T Y$$

$$S^2 = \hat{\sigma}^2 = \frac{SCE}{n - (k + 1)} = \frac{Y^T Y - \hat{\beta}^T X^T Y}{n - (k + 1)}$$

Una vez estimados los parámetros es posible construir intervalos de confianza y tests de hipótesis para cada uno de los mismos por separado puesto que tienen una distribución de probabilidad conocida. Sin embargo, nuestro interés recae en establecer un test global para contrastar la bondad de la regresión entre Y y las variables X_j, j=1...k de la forma siguiente:

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

H_1: $\beta_j \neq 0$ para algún j=1...k

Evidentemente, la hipótesis nula establece que no existe relación lineal entre la variable respuesta y el conjunto de variables de predicción. La técnica de análisis de la varianza proporciona un camino adecuado para probar esta hipótesis. También aquí es posible descomponer la suma total de cuadrados (variación total de Y respecto a su media) en dos partes: la suma de cuadrados de la regresión y la suma de cuadrados de los errores.

$$SCT = SCR + SCE$$

El número de grados de libertad que afectan a cada uno de los sumandos son: (n-1), k y n-(k+1) respectivamente, lo que da lugar a que bajo H_0, el cociente:

$$F = \frac{SCR / k}{SCE / [n - (k + 1)]}$$

se distribuya según una F de Snedecor con k y n-(k+1) grados de libertad. El test consiste en calcular el valor experimental de F y rechazar la hipótesis nula a nivel α si $F > f_{\alpha,k,n-(k+1)}$

En el análisis de la regresión lineal múltiple se define el coeficiente de determinación múltiple R^2 como:

$$R^2 = \frac{SCR}{SCT} = 1 - \frac{SCE}{SCT} \qquad 0 \leq R^2 \leq 1$$

Dicho coeficiente mide la proporción de la variación total de las observaciones con respecto a su media que es atribuible a la ecuación de regresión estimada. Es decir, R^2 expresa en qué medida las variables de predicción incluidas en el modelo explican la variación de las observaciones. A la raíz cuadrada positiva de este coeficiente se le denomina generalmente correlación múltiple:

$$R = +\sqrt{R^2}$$

y proporciona una medida de la relación lineal entre Y y $X_1 ... X_k$.

También es conveniente determinar los coeficientes de correlación lineal entre dos variables cualesquiera X_i, X_j de las que intervienen en el estudio cuando se elimina la parte de la relación que pueda ser debida a la influencia del resto de las variables. Estos coeficientes r_{ij} aparecen dispuestos en una matriz cuadrada simétrica ($r_{ij}=r_{ji}$) y cuya diagonal está formada por unos ($r_{ii}=1\ \forall i$).

Esta matriz nos permitirá detectar un problema frecuente que se plantea en la regresión lineal múltiple: la existencia de correlación lineal entre alguna de las variables explicativas, problema que se conoce con el nombre de multicolinealidad. Este hecho da lugar a resultados menos precisos puesto que los parámetros β_j no miden correctamente los efectos individuales sobre la respuesta, sino que reflejan un efecto parcial sobre la misma, sujeto a todo lo que pase con las demás variables de predicción en la ecuación de regresión. Si en la matriz de correlaciones entre las variables explicativas aparecen correlaciones significativas una solución posible (entre las varias que existen) consiste en calcular las ecuaciones de regresión eliminando las variables que se encuentran correlacionadas linealmente y valorar la bondad de dichas regresiones. Existen criterios para seleccionar la mejor ecuación de regresión entre las que se encuentra la regresión por pasos, pero en cualquier caso, es conveniente realizar un análisis de los residuos para detectar las deficiencias del modelo o la posible violación de las hipótesis con el fin de obtener conclusiones acertadas.

3. ANÁLISIS DE DATOS Y RESULTADOS

Los datos utilizados para el análisis se han obtenido de la última encuesta de Presupuestos Familiares publicada por Instituto Nacional de Estadística realizada entre abril de 1990 y marzo de 1991.

La variable estudiada es el "gasto anual" de los hogares de Castilla y León para los distintos grupos de gasto de consumo que refleja la encuesta:

GRUPO 1: Alimentación, bebidas y tabaco.
GRUPO 2: Vestido y calzado.
GRUPO 3: Vivienda, calefacción y alumbrado.
GRUPO 4: Artículos de mobiliario, menaje y conservación del hogar.
GRUPO 5: Servicios médicos y gastos sanitarios.
GRUPO 6: Transportes y comunicaciones.
GRUPO 7: Esparcimiento, enseñanza y cultura.
GRUPO 8: Otros bienes y servicios.
GRUPO 9: Otros gastos no mencionados anteriormente..

Más concretamente la variable que se analiza hace referencia al gasto anual medio por hogar en las provincias que integran nuestra comunidad. El hecho de considerar el gasto medio se debe a que permite analizar mejor las posibles diferencias en los gastos de los hogares por provincia. Hay que tener en cuenta que la E.P.F. es una encuesta por muestreo y obviamente no se ha investigado el mismo número de hogares en cada provincia. La tabla 1 recoge la información de partida:

Tabla 1

	Avila	Burgos	León	Palencia	Salamanca	Segovia	Soria	Valladolid	Zamora
GRUPO 1	534,595	602,307	623,047	589,710	488,784	528,040	679,722	567,361	544527
GRUPO 2	242,609	255,567	245,240	206,048	167,814	184,840	232,673	223,201	178,835
GRUPO 3	388,063	600,121	500,414	449,113	400,049	455,368	503,695	566,433	402,354
GRUPO 4	92,808	162,166	136,030	113,714	78,217	103,446	129,768	140,573	99,953
GRUPO 5	47,035	51,308	41,667	34,787	24,476	46,337	55,000	46,111	32,146
GRUPO 6	254,563	280,023	333,066	248,620	195,065	217,156	272,249	254,216	227,163
GRUPO 7	74,522	132,161	119,657	115,825	69,846	91,436	117,587	149,041	70,283
GRUPO 8	250,853	301,813	267,506	294,068	193,056	259,705	300,014	327,774	231,577
GRUPO 9	82,061	111,224	146,434	109,264	54,442	116,303	120,803	98,430	125,332

En primer lugar hemos realizado un análisis de la varianza con el fin de determinar en qué medida la variable dependiente o variable respuesta "gasto medio por hogar" se ve afectada por la variable independiente o factor "provincia". Ahora bien, el hecho de que los gastos se clasifiquen también en función del "grupo de gasto", nos lleva a considerar este último factor como una variable bloque para eliminar su efecto y conseguir mayor información en la comparación de los tratamientos (las nueve provincias de Castilla y León). La tabla 2 recoge los resultados siguientes:

Tabla 2

ANALYSIS OF VARIANCE

	x2	Gasto medio por hogar
BY	x1	Provincia
	x3	Grupo de gasto

Source of Variation	Sum of Squares	DF	Mean Squaree	F	Signif of F
Main Effects	2390277474912	16	149392342182	180.134	.000
X1 Tratamientos	71973485148.2	8	8996685643.5	10.848	.000
X2 Bloques	2318303989763	8	289787998720	349.419	.000
Explained	2390277474912	16	149392342182	180.134	.000
Residual	53077887340.0	64	829341989.69		
Total	2443355362252	80	30541942028		

En la tabla ANOVA la variación debida a los efectos principales se divide entre la debida al factor "PROVINCIA" y la debida al bloque "GRUPO DE GASTO". Tan solo nos interesa el efecto de la provincia y como éste es significativo (significación de F=0.000), rechazamos la hipótesis nula de igualdad de medias entre provincias y concluimos que el gasto medio por hogar se ve afectado por el factor "PROVINCIA". Como el efecto del factor es significativo podemos obtener una clasificación múltiple que valora las diferencias entre los efectos de los nueve niveles del factor "PROVINCIA" (tabla 3):

Tabla 3

Grand Mean = 239619.926

Variable + Category	N	Unadjusted Dev'n	Unadjusted Eta	Adjusted for Independents Dev'n	Adjusted for Independents Beta
X1 PROVINCIAS					
1 Avila	9	-210052.26		-210052.26	
2 Burgos	9	37790.07		37790.07	
3 León	9	28497.96		28497.96	
4 Palencia	9	507.74		507.74	
5 Salamanca	9	-53870.04		-53870.04	
6 Segovia	9	-17105.26		-17105.26	
7 Soria	9	28325.74		28325.74	
8 Valladolid	9	24062.30		24062.30	
9 Zamora	9	-27156.26		-27156.26	
			.17		.17

Dicha tabla refleja las desviaciones típicas que presentan la media de cada provincia respecto a la media total. Como puede observarse las diferencias negativas se sitúan por debajo de la media total: 239619,926 y las positivas por encima. No obstante, dado que el coeficiente de asociación ETA, proporciona un valor de 0,17 las conclusiones deducidas no serán muy fiables.

Respecto a la validación de las hipótesis iniciales analizamos la homocedasticidad (igualdad de varianzas entre los tratamientos) a través del test de Bartlett-Box que proporciona un valor de 0,132 con una p=0,998, lo que indica que no se puede rechazar la hipótesis de homogeneidad de varianzas.

Se ha comprobado también la normalidad de las poblaciones en estudio (gasto medio por provincia) tomando como referencia el test no paramétrico de Kolmogorov-Smirnov, que indica que no se rechaza la hipótesis de normalidad en ninguno de los casos.

En cuanto a la posibilidad de establecer relaciones funcionales entre los diferentes grupos de gasto en las nueve provincias (considerando como variable dependiente el gasto en el grupo 1), se ha obtenido un coeficiente de determinación múltiple de $R^2=1$, lo que indicaría que el porcentaje de variación explicada del gasto en el grupo 1 por el resto de los grupos de gasto sería del 100%. Además, la suma de cuadrados de los errores es nula: SCE=0 lo que hace que el estadístico F no esté definido. Sin embargo, al analizar la matriz de correlaciones observamos que hay variables que están frecuentemente correlacionadas: los gastos medios en los grupos 3 (vivienda, calefacción y alumbrado), grupo 4 (artículos de mobiliario, menaje y conservación del hogar) y grupo 7 (esparcimiento, enseñanza y cultura). Esto indica que existe multicolinealidad, lo cual no permite considerar el coeficiente R^2 todo lo bueno que parece ser.

Por ello, se realiza un análisis de regresión eliminando en primer lugar la variable "gasto en el grupo 4", y a continuación eliminando las variables" gasto en el grupo 4" y "gasto en el grupo 7". En ambos casos no podemos rechazar la hipótesis de que todos los coeficientes de regresión son nulos ya que tanto los contrastes globales con el test de la F, como los contrastes individuales para los distintos coeficientes de regresión con el test de la t de Student no son significativos.

Evidentemente, en estos dos últimos casos, los coeficientes de determinación múltiple han disminuido y son respectivamente 0.82562 y 0.74351. Sin embargo, esto no significa que las variables explicativas seleccionadas no influyan en la variable considerada como dependiente: gasto en el grupo 1, sino solamente que sus efectos no se pueden detectar en la muestra.

4. CONCLUSIONES

* En primer lugar señalar la utilidad del tratamiento estadístico para deducir resultados sobre los gastos de consumo, variable con gran incidencia en el funcionamiento económico.

* La diferencia significativa de medias de gasto en las provincias de Castilla y León lo que indica una clara diferencia entre las provincias con respecto al gasto analizado. Por tanto, no se puede caracterizar de la misma forma a todas las provincias de esta Comunidad.

* Las provincias que se separan por encima de la media global de gasto son: Burgos, León, Palencia, (muy débilmente), Soria y Valladolid, mientras que las que se desvían por debajo son: Avila, Salamanca, Segovia y Zamora.

* Una fuerte correlación entre las variables gasto en el grupo 3, gasto en el grupo 4 y gasto en el grupo 7. Aún prescindiendo de las variables mencionadas, el modelo de regresión lineal no es adecuado para relacionar la variable que hemos considerado como dependiente "gasto en el grupo 1" con el resto de los grupos de gasto.

* No podemos concluir que las variables seleccionadas como explicativas (gasto en el grupo 2, grupo 3, grupo 5, grupo 6, grupo 8, grupo 9) no influyan en la variable dependiente, sino que la información de la muestra no permite detectar sus efectos.

5. BIBLIOGRAFÍA

BISQUERRA, R., Introducción a la Estadística Aplicada a la Investigación Educativa. Un enfoque informático con los paquetes BMDP y SPSSX, Ed. PPU, Barcelona, 1987.

CANAVOS, G.C., Probabilidad y Estadística. Aplicaciones y Métodos, Ed. McGraw-Hill, Madrid, 1992.

IMAN, R.L.; CONOVER, W.J., Modern Business Statistics, Ed. John Wiley and Sons, New York, 1989.

INSTITUTO NACIONAL DE ESTADISTICA, Encuesta de Presupuestos Familiares 1990-91, Madrid, 1992.

PEÑA SANCHEZ DE RIVERA, D., Estadística. Modelos y Métodos. 2.Momentos lineales y series temporales, Ed. Alianza Universidad, Madrid, 1989.

RENTA NACIONAL DE ESPAÑA Y SU DISTRIBUCION PROVINCIAL. Avance 1990-1991, Ed. BBV, Bilbao, 1992.

SHIFFLER, R.E.; ADAMS, A.J., Introductory Business Statistics with Microcomputer Applications, Ed. PWS-KENT, Boston, 1990.

WALPOLE, R.E.; MYERS, R.H., Probabilidad y Estadística, Ed. McGraw-Hill, Mexico, 1992, 4ª edición.