Mixtura de distribuciones normales incluyendo modelado conjunto de media y varianza desde un enfoque clásico

Xabier Fabian Roldan Figueredo
Lic en Matemáticas, Ms(c)
Código: 01832406

Universidad Nacional de Colombia
Facultad de Ciencias
Departamento de Estadística
Bogotá, D.C.
Agosto de 2014

Mixtura de distribuciones normales incluyendo modelado conjunto de media y varianza desde un enfoque clásico

Xabier Fabian Roldan Figueredo
Lic en Matemáticas, Ms(c)
Código: 01832406

Disertación presentada para optar al título de
Magister en Ciencias estadística

Director
Edilberto Cepeda, Ph.D.

Universidad Nacional de Colombia
Facultad de Ciencias
Departamento de Estadística
Bogotá, D.C.
Agosto de 2014

Título en español

Mixtura de distribuciones normales incluyendo modelado conjunto de media y varianza desde un enfoque clásico.

Title in English

Mixture of normal distributions, including joint modeling of mean an variance.

Resumen: En este trabajo se presenta un modelo de mixtura finita de distribuciones normales, condicionado al modelado de la media y la varianza, asumiendo heterocedasticidad de tipo multiplicativa en el modelo de la varianza. Se describe un procedimiento para hacer la estimación de máxima verosimilitud de los parámetros del modelo y para ello se emplea el algoritmo EM y Fisher Scoring en el paso M del algoritmo EM. Esta propuesta se desarrolla en tres capítulos: en el primer capítulo se analiza el modelo de regresión lineal clásico en presencia de heterocedasticidad con todas sus implicaciones, en el segundo capítulo se estudia el modelo de mixtura finita de distribuciones normales, inicialmente no condicionado y luego condicionado al modelado de la media y en el tercer capítulo se presenta el modelo de regresión de mixturas con modelado de varianza y el algoritmo para realizar su estimación. Como caso práctico para los capítulos 2 y 3 se toma un conjunto de datos correspondientes a los puntajes totales y a los puntajes específicos en la prueba de comprensión lectora obtenidos por una muestra de estudiantes de Economía en el Examen de Calidad en la Educación Superior (ECAES) en el 2009. Se ajusta un modelo de mixtura finita de distribuciones normales, un modelo de mixtura finita de distribuciones normales condicionado al modelado de media y luego un modelo de mixtura finita de distribuciones normales condicionado al modelado de media y varianza. Además se realiza una simulación del algoritmo propuesto para observar su desempeño en la estimación.

Abstract: We present a finite mixture model of normal distributions, conditioned by modeling for mean and variance, it assume multiplicative heteroscedasticity for variance model. It describe a procedure to estimate the maximum likelihood of the model's parameters, in order for this, the EM algorithm is applied with Fisher Scoring in M step. This proposal consist of three chapters: in the first one has an analyze about classical linear regression model with heteroscedasticity and its implications, in the second one we study the finite mixture model by normal distributions, no conditioned and conditioned and in third one a model of mixture regression with variance modeling is presented and the respective algorithm. As a practical application in chapters two and three we consider a data set from total and specific scoring of reading comprehension obtained in a sample of economics students in the Examen de Calidad en la Educación Superior (ECAES) in 2009. The finite mixture model by normal distributions, and finite mixture model, conditioned to mean modeling, and conditioned to mean and variance modeling, are fitting. In addition also an algorithm's simulation is performed to observe its performance in estimating.

Nota de aceptación

Trabajo de tesis
Aprobado

"Mención "

Jurado

Jurado

Jurado

Director
Edilberto Cepeda

Bogotá, D.C., Diciembre 31 de 2013

Dedicado a

Amados hermanos, cuando tengan que enfrentar problemas, considérenlo como un tiempo para alegrarse mucho porque ustedes saben que, siempre que se pone a prueba la fe, la constancia tiene una oportunidad para desarrollarse. Así que dejen que crezca, pues una vez que su constancia se haya desarrollado plenamente, serán perfectos y completos, y nos les faltara nada. Santiago 1, 2-4.

Índice general

Índice general **I**

Índice de tablas **III**

Índice de figuras **IV**

Introducción **V**

1. Modelo de regresión lineal con heterocedasticidad **1**

1.1. Introducción 1

1.2. Modelo Clásico de regresión lineal 2

1.2.1. Supuestos en el modelo de Regresión Lineal 3

1.2.2. Modelo de Regresión Lineal Normal 3

1.2.3. Heterocedasticidad y sus consecuencias en el Modelo Lineal Clásico . 4

1.3. Causas de la heterocedasticidad 5

1.4. Pruebas de homocedasticidad 6

1.4.1. Prueba de Goldfeld-Quandt 7

1.4.2. Prueba de Park 8

1.4.3. Prueba de Rutemiller-Bowers 8

1.4.4. Prueba de Glejser 8

1.4.5. Prueba de White 9

1.4.6. Prueba de Breusch Pagan 10

1.5. Estimación del Modelo de regresión lineal en presencia de heterocedasticidad 11

1.6. Heterocedasticidad Multiplicativa 14

1.6.1. Estimación del modelo de regresión lineal normal con heterocedasticidad multiplicativa 15

1.7. Aplicación 17

2. Modelo de Mixtura Finita de Distribuciones Normales 22

2.1. Introducción 22

2.2. Modelo no condicionado 23

2.2.1. Introducción 23

2.2.2. El modelo 25

2.2.3. Estimación del modelo vía algoritmo EM 26

2.2.4. Estimación de máxima verosimilitud y datos completos 27

2.2.5. EL algoritmo EM para modelos de mixtura finita con componentes normales 29

2.2.6. Aplicación 31

2.3. Modelo condicionado a la media 32

2.3.1. introducción 32

2.3.2. El modelo 39

2.3.3. Estimación del modelo vía algoritmo EM 39

2.3.4. Aplicación 40

3. Mixtura Finita de distribuciones normales con modelado de media y varianza 43

3.1. Introducción 43

3.2. Propuestas Anteriores 44

3.3. El modelo 44

3.4. Estimación del modelo vía algoritmos EM y Fisher Scoring 45

3.5. Valores Iniciales 48

3.6. Simulación 50

3.6.1. Diseño del Estudio 50

3.6.2. Resultados de la Simulación 53

3.7. Aplicación 54

Conclusiones 58

Bibliografía 59

Índice de tablas

1.1. Modelos propuestos para la estimación del gasto con tarjetas de crédito . . . 18

1.3. Significancia conjunta de las variables haciendo uso de matriz de White, $HC1$, $HC3$ y $HC4$ . . . 20

1.2. Estimación del modelo utilizando MCO y errores estándar bajo matrices de White, $HC1$, $HC3$ y $HC4$ . . . 20

1.4. Modelo de Heterocedasticidad Multiplicativa . . . 21

2.1. Valores de los criterios para la determinación del número de componentes de la mixtura de los puntajes de la prueba ECAES 2009 en economía . . . 31

2.2. Estimación de parámetros de la mixtura de dos componentes normales vía algoritmo EM para los puntajes de la prueba ECAES 2009 de economía. . . . 32

2.3. Criterios de Información para evaluar la cantidad de componentes del modelo de mixtura de regresiones de los puntajes totales de la prueba ECAES 2009 de economía . . . 41

2.4. Estimación de parámetros de la mixtura de regresiones de dos componentes normales vía algoritmo EM de los puntajes totales de la prueba ECAES 2009 de economía . . . 42

3.1. Modelos de Mixtura para la Simulación . . . 50

3.2. Resultados de desempeño del algoritmo EM-Fisher Scoring en las iteraciones de la estimación de parámetros . . . 54

3.3. Resultados de desempeño del algoritmo EM-Fisher Scoring en la estimación de parámetros . . . 54

3.4. Estimación de parámetros de la mixtura de dos distribuciones normales con modelado de media y varianza vía algoritmo EM-Fisher Scoring de los puntajes totales de la prueba ECAES 2009 de economía . . . 55

3.5. Criterios de Información para evaluar los modelos de mixtura con modelado de media y varianza propuestos para el puntaje total de la prueba ECAES 2009 de economía . . . 57

Índice de figuras

1.1. Gráfico de residuos frente a las variables 19

2.1. Histograma de 45 puntajes totales de la prueba saber pro de economía del año 2009 31

2.2. Gráfico de dispersión de una muestra de los puntajes de comprensión lectora y puntaje total en el ECAES 2009 de economía 41

2.3. Rectas de regresión ajustadas al modelo de regresión de mixturas del puntaje total obtenido por una muestra de 45 estudiantes en la prueba ECAES de Economía de 2009 42

3.1. Modelo 1 y 2 de estudio de simulación 51

3.2. Modelo 3 de estudio de simulación 52

3.3. Modelo 4 y 5 de estudio de simulación 53

3.4. Rectas de regresión ajustadas al modelo de regresión de mixturas del puntaje total obtenido por una muestra de 45 estudiantes en la prueba ECAES de Economía 2009, con modelado conjunto de media y varianza. Con la propuesta 1 para la varianza 55

3.5. Rectas de regresión ajustadas al modelo de regresión de mixturas del puntaje total obtenido por una muestra de 45 estudiantes en la prueba ECAES de Economía 2009, con modelado conjunto de media y varianza. Con la propuesta 2 para la varianza 56

Introducción

En algunos fenómenos aleatorios, el análisis exploratorio de los datos puede mostrar cómo en estos se forman grupos o clúster, en el que es posible aplicar un modelo de regresión para cada grupo. Este tipo de situaciones motivaron la propuesta de modelos que proporcionen entre otros, la identificación del número de grupos o componentes en la muestra, la pertenencia o no de una observación a cada componente y la explicación de la variabilidad presente en cada grupo. En (Quandt, 1972) se introduce el término **Cambio de Regresión** a un problema asociado con la estimación de un conjunto de regresiones para todas las observaciones procedentes de un número posiblemente desconocido de clases. Modelos relacionados con este mismo escenario son trabajados posteriormente en (Quandt & Ramsey, 1978), en el que consideran que un sistema de cambio de regresión es a menudo comparado con un cambio estructural en el sistema. Otro es el problema combinatorio de clusterwise tratado por (Späth, 1979) y la regresión clusterwise presentada por (DeSarbo, 1988). En el trabajo de (Wedel & DeSarbo, 1995) aparece la regresión de clases latentes asociada al mismo contexto de los trabajos anteriores.

El contexto en el que están enmarcadas estas propuestas es el de regresión de mixturas. Estos modelos son un tipo de mixturas finitas en el que se define un modelo de regresión para la media de cada distribución que compone la mixtura. Cada grupo de observaciones tienen una función de densidad asociada y corresponden a una proporción de la muestra. La mixtura combina de forma lineal esas funciones de densidad. (McLachlan G, 2000) considera a las mixturas finitas de distribuciones como un método extremadamente flexible de modelado estadístico, cuyas aplicaciones se han ampliado considerablemente y son base de una variedad de métodos estadísticos.

En el momento de trabajar con mixturas de regresiones, un problema a resolver es el que se refiere a la estimación de los parámetros. Son amplias las metodologías para la estimación del conjunto de parámetros de este tipo de modelos. Los métodos clásicos implementan, entre otros, la maximización de la función de logverosimilitud, utilizando para ello algoritmos como el EM, un ejemplo es el trabajo de (Aitkin & Tunnicliffe, 1980), donde se propone la detección de valores atípicos en el conjunto de datos usando regresión de mixtura de normales. Otros métodos que también se han utilizado es la función generatriz de momentos como en (Quandt & Ramsey, 1978) o programación matemática no lineal con restricciones lineales como en (Lau et al., 1999) donde se propone un modelo generalizado clusterwise. Desarrollos computacionales también han mejorado la implementación de los algoritmos de estimación, permitiendo comparar sus desempeños.

(Ding, 2006), propone una mixtura de regresiones para el análisis de la relación entre el rendimiento de los estudiantes en matemáticas, el enfoque metodológico implementado,

el concepto propio de los estudiantes frente a las matemáticas y la calificación dada por el profesor sobre el dominio de las matemáticas en los estudiantes. El análisis exploratorio de los datos permite observar la presencia de subgrupos. Por ejemplo, es posible que los estudiantes con calificaciones altas en matemáticas puedan estar influenciados por el enfoque metodológico, su rendimiento y su propio concepto sobre las matemáticas, mientras otro grupo de estudiantes con calificaciones bajas en matemáticas, puedan estar influenciados únicamente por el enfoque metodológico. En esta aplicación, sí se asume que todos los individuos pertenecen a una única población y que las variables independientes tienen la misma influencia sobre la variable dependiente entonces es el modelo de regresión lineal clásica el que mejor se ajusta, implicando con ello que se haga caso omiso a la heterogeneidad presente en los datos. En este ejemplo se presta atención a las situaciones en los que la heterogeneidad de la población es inobservable; la pertenencia de un individuo a un grupo de individuos de la población es latente.

Los modelos de regresión de mixturas se pueden ver como una combinación del modelo de regresión convencional y los modelos de clases latentes; la variabilidad presente en los grupos es explicada a través del modelado de la media de las distribuciones. Una ampliación a este modelo es considerar además la presencia de heterocedasticidad en cada componente de la mixtura. En (Garrido et al., 2011) y (Garrido & Cepeda, 2012) se propone la estimación de un modelo de mixtura finita de distribuciones normal-exponenciales con modelado de media y varianza implementando metodología bayesiana para la estimación del mismo.

En esta tesis se propone un modelo de mixtura de distribuciones normales incluyendo el modelado conjunto de media y varianza, realizando la estimación de máxima verosimilitud de los parámetros del modelo vía algoritmo EM con Fisher Scoring para la maximización en el paso M. Se utiliza un conjunto de datos para las aplicaciones en el que es posible que una sola regresión normal no lo explique y se incluye una simulación del algoritmo para observar su desempeño.

Esta tesis se estructuró en tres capítulos de la siguiente manera:

En el capítulo 1 se presenta el problema de la heterocedasticidad en el modelo de regresión lineal clásico. En el capítulo 2 se revisan algunos de los modelos que han sido propuestos para los problemas de mixtura y de regresión de mixturas. Se caracteriza este capítulo por el estudio del modelo de mixtura finita de distribuciones normales no condicionado y del modelo de mixturas condicionado al modelado de media.

Finalmente, en el capítulo 3 se presenta una propuesta en la que se hace una extensión de los modelos de regresión de mixturas de distribuciones normales, en la cual se incluye heterocedasticidad en la regresión de mixturas. La heterocedasticidad es de tipo multiplicativo. La esencia de esta tesis se encuentra en este capítulo, en el que una vez hecha la propuesta del modelo, se describe un procedimiento para estimarlo y se exploran algunos aspectos teóricos y prácticos de este. Es en este capítulo donde se presentan los resultados de la simulación del algoritmo.

CAPÍTULO 1

Modelo de regresión lineal con heterocedasticidad

1.1. Introducción

Con frecuencia y en particular en el campo de la economía se requiere especificar y estimar un modelo que relacione variables económicas. Se recoge información a partir de una muestra, con la que se pretende cuantificar la relación entre dichas variables.

Se trata entonces el problema de explicar una variable utilizando cierto número de variables explicativas y una variable aleatoria no observable. La relación que se establece de dependencia entre dichas variables es de tipo lineal y se conoce como modelo de regresión lineal. Dicho modelo cuenta con unas características y tiene como objetivo, estimar los parámetros de la regresión a partir de la información dada por las observaciones que se tienen.

Con estas consideraciones, la primera parte del capitulo presenta: el modelo clásico de regresión lineal con sus diferentes supuestos, el estimador de mínimos cuadrados ordinarios (MCO), como método de estimación de los parámetros de la regresión, y el modelo clásico de regresión lineal normal con el estimador de máxima verosimilitud, cuya importancia radica en que sus propiedades permiten hacer inferencia.

A continuación se examina la violación de un supuesto importante en el modelo de regresión lineal. Este supuesto hace referencia acerca de la varianza de lo que se ha llamado variable aleatoria no observable o error que es constante, conocido esto como homocedasticidad. Se explican las posibles causas de la presencia de varianza no constante en este término, esto es, causas de la heterocedasticidad; así como los procedimientos para detectarla. Sobre esto último, se realiza una breve exposición de cada una de las pruebas utilizadas para su detección y la estructura dada a la varianza.

Luego, se describen los procedimientos relacionados con la forma de estimar el modelo bajo el escenario de la ausencia de homocedasticidad. Se explica en qué consiste la heterocedasticidad de tipo multiplicativo como modelo general y flexible de la varianza. En esta sección se describe la estimación del modelo clásico de regresión lineal normal con heterocedasticidad multiplicativa.

Para finalizar el capítulo, se analiza la heterocedasticidad en un conjunto de datos, de 100 individuos y su gasto mensual en tarjeta de crédito, tomado de (Greene, 2003).

1.2. Modelo Clásico de regresión lineal

El análisis de regresión es una técnica estadística que pretende identificar la relación entre una variable Y, caracterizada como respuesta, y una o más variables X_i denominadas explicativas. La relación entre las variables consiste en que la variable respuesta Y es afectada por k factores o variables explicativas $X_1, X_2, ..., X_k$, e interesa construir un modelo para explicar el comportamiento de Y en función de las variables X_i.

Los modelos se pueden clasificar como determinísticos o probabilísticos dependiendo si se considera a la variable respuesta como una variable aleatoria, siendo estos últimos quienes hacen esta consideración y son los de interés para el análisis de regresión. Los modelos probabilísticos consideran un error para predecir o pronosticar la variable respuesta en función de las variables explicativas y parten del modelo general $Y = m(X_1, ..., X_k)+\varepsilon$, donde ε es una variable aleatoria que tiene una distribución de probabilidad específica con media cero.

El análisis de regresión no hace supuesto de la distribución marginal de las variables explicativas X_i, pues si bien las variables $X_1, X_2, ..., X_k$ pueden considerarse variables aleatorias, interesa la variable aleatoria Y para valores dados de las variables X_i. Esto hace que el modelo $Y = m(X_1, ..., X_k) + \varepsilon$ este compuesto de una componente determinística y una componente estocástica, que puede entenderse como un experimento que al repetirse en condiciones idénticas da resultados no necesariamente iguales para Y.

Como se supone que la variable aleatoria ε tiene media cero, la función $m(X_1, ..., X_k)$ es la esperanza condicional de Y dados los valores de $X_1, ..., X_k$; esto es, $E(Y|x_1...x_k) = m(X_1, ..., X_k)$. Esta función se denomina "función de regresión".

La regresión paramétrica especifica la forma de $m(X_1, ..., X_k)$ para proponer la relación entre la variable respuesta y las variables explicativas. El modelo más usual es la forma lineal $m(X_1, ..., X_k) = \beta_0 + \beta_1 X_1 + ... + \beta_k X_k$, donde $\boldsymbol{X} = (1, X_1, ..., X_k)$ es el vector de covariables y $\boldsymbol{\beta} = (\beta_0, \beta_1, ..., \beta_k)$ es un vector de parámetros desconocidos. La forma lineal $m(X_1, ..., X_k) = \beta_0 + \beta_1 X_1 + ... + \beta_k X_k$ es una función lineal de las variables explicativas $X_1, ..., X_k$ pero también es función lineal de los parámetros desconocidos $\beta_0, \beta_1, ..., \beta_k$. Al referirse a un modelo lineal para Y, se considera que la función de regresión es una función lineal de los parámetros desconocidos $\beta_0, \beta_1, ..., \beta_k$.

Si se asume que la función de regresión $E(Y|x_1...x_k)$ es una función lineal del vector de parámetros $\boldsymbol{\beta} = (\beta_0, \beta_1, ..., \beta_k)$, esto es, $E(Y|x_1...x_k) = \beta_0 + \beta_1 X_1 + ... + \beta_k X_k$, se trata de estimar los valores no conocidos del vector de parámetros $\boldsymbol{\beta}$ del modelo de regresión lineal

$$Y = \beta_0 + \beta_1 X_1 + ... + \beta_k X_k + \varepsilon. \tag{1.1}$$

Los estimadores del modelo (1.1)que se denotan como $\widehat{\boldsymbol{\beta}} = (\widehat{\beta_0}, \widehat{\beta_1}, ..., \widehat{\beta_k})$, requieren para su cálculo la consideración de una muestra $(y_i, x_{i1}, ..., x_{ik})$ $i = 1, .., n$, donde cada observación se supone que proviene de

$$Y_i = \beta_0 + \beta_1 x_{i1} + ... + \beta_k x_{ik} + \varepsilon_i. \tag{1.2}$$

Las n observaciones de Y, correspondientes a n valores controlados u observados de $X_1, ..., X_k$, que se obtienen de aplicar el modelo (1.2) genera el sistema

$$\begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{bmatrix} = \begin{bmatrix} 1 & x_{11} & x_{12} & \dots & x_{1k} \\ 1 & x_{21} & x_{22} & \dots & x_{2k} \\ \vdots & \vdots & & \vdots & \\ 1 & x_{n1} & x_{n2} & \dots & x_{nk} \end{bmatrix} \begin{bmatrix} \beta_0 \\ \beta_1 \\ \vdots \\ \beta_k \end{bmatrix} + \begin{bmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \vdots \\ \varepsilon_n \end{bmatrix} \tag{1.3}$$

o de forma matricial

$$\boldsymbol{Y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon} \tag{1.4}$$

Donde $\boldsymbol{X}$ ahora representa una matrix $n \times (k+1)$, $\boldsymbol{Y}$ es un vector de variables aleatorias observable de n componentes y $\boldsymbol{\varepsilon}$ es un vector de errores aleatorios no observable de n componentes. Esto se puede entender como una muestra de observaciones o variables aleatorias observadas y_i, $i = 1, ..., n$, seleccionadas aleatoriamente de una distribución con función de densidad $f(y_i|x_{i1}, ..., x_{ik}, \beta)$ con media $\beta_0 + x_{i1}\beta_1 +, ..., +x_{ik}\beta_k$, donde los valores de $x_{i1}, ..., x_{ik}$ han sido preseleccionados y el proceso de muestreo se ha repetido n veces.

Los estimadores $\widehat{\boldsymbol{\beta}}$ se pueden obtener por el método de mínimos cuadrados (MCO) sin especificar la distribución de probabilidad del error aleatorio ε, cuyo resultado es

$$\widehat{\boldsymbol{\beta}} = (X'X)^{-1}X'Y \tag{1.5}$$

1.2.1. Supuestos en el modelo de Regresión Lineal

Los supuestos sobre las variables X_i y el término de perturbación del modelo (1.1), ε_i, son importantes para realizar una interpretación adecuada de los valores estimados de la regresión. Primero, se supone que el modelo (1.1) es lineal en los parámetros pero no necesariamente en las variables, con un valor esperado del termino del error de 0. La presencia del termino del error hace que la relación entre las variables explicativas (exogena) y la explicada (endogena) sea estocástica. Segundo, la varianza del termino del error es constante sin importar el valor de las variables X_i, supuesto conocido como varianza homocedastica, esto es, $Var(\varepsilon_i) = \sigma^2$ con $i = 1, ..., n$. Por último, se supone que las variables X_i y el termino de error son independientes y sin autocorrelación entre los valores del término de perturbación.

1.2.2. Modelo de Regresión Lineal Normal

El método de MCO no hace referencia al tipo de distribución de probabilidad que siga el error aleatorio, trayendo como consecuencia que no se pueda hacer inferencia sobre la función de regresión poblacional de donde se extrajo la muestra de observaciones.

Si se asume que el error aleatorio en el modelo (1.1) es una variable aleatoria con distribución normal de media cero y varianza σ^2, tal que al considerar una muestra de tamaño n, obtenemos una muestra aleatoria $\varepsilon_1, \varepsilon_2, ..., \varepsilon_n$, esto es, ε_i, con $i = 1, ..., n$ independientes e idénticamente distribuidas, entonces las observaciones Y_i son variables aleatorias que tienen una distribución normal con media $X\boldsymbol{\beta}$ y varianza σ^2. A este modelo se le denomina "modelo de regresión normal".

Los estimadores de los parámetros del modelo de regresión normal se pueden obtener mediante el método de máxima verosimilitud y coinciden en construcción con los estimadores que se obtienen por el método de los mínimos cuadrados.

Algunas de las razones para asumir normalidad en el término del error son formuladas por (Gujarati, 2010), de ellas es importante resaltar que con este supuesto, por un lado se derivan las distribuciones de probabilidad de los MCO y por otro permite utilizar las pruebas estadísticas t, F y χ^2 para realizar inferencias respecto de los valores de los parámetros poblacionales, así como construir intervalos de confianza que pueda contener dentro de sus limites el verdadero valor del parámetro poblacional desconocido.

1.2.3. Heterocedasticidad y sus consecuencias en el Modelo Lineal Clásico

El modelo de regresión lineal (1.2) plantea una ecuación de regresión para cada conjunto de valores fijos $(x_{i1}, ..., x_{ik})$ con $i = 1, .., n$, resultado de considerar una muestra de tamaño n para la estimación de parámetros del modelo (1.1). Este modelo se propone bajo el supuesto de homocedasticidad, esto es, que para $i = 1, ..., n$ la variable ε_i tiene media cero y varianza no dependiente de i. Si por algún motivo la varianza cambia con la observación i, es decir $E(\varepsilon_i^2) = \sigma_i^2$, se dice que el modelo (1.2), y por ende (1.1), presentan heterocedasticidad. De forma matricial podemos escribir

$$E[\varepsilon\varepsilon'] = \begin{bmatrix} \sigma_1^2 & 0 & 0 & \dots & 0 \\ 0 & \sigma_2^2 & 0 & \dots & 0 \\ \vdots & \vdots & & \vdots & \\ 0 & 0 & 0 & \dots & \sigma_n^2 \end{bmatrix} \tag{1.6}$$

Si se considera a $\sigma_i^2 = \sigma^2\omega_i$ se tiene que

$$\sigma^2\Omega = \sigma^2 \begin{bmatrix} \omega_1 & 0 & 0 & \dots & 0 \\ 0 & \omega_2 & 0 & \dots & 0 \\ & & \vdots & \vdots & \\ 0 & 0 & 0 & \dots & \omega_n \end{bmatrix} \tag{1.7}$$

De esta forma como menciona (Greene, 2003), ω podría reflejar solamente la variedad en las varianzas de las perturbaciones.

Los estimadores $\widehat{\beta_0}, ..., \widehat{\beta_k}$, obtenidos por el método de mínimos cuadrados, bajo los supuestos mencionados en la sección 1.2.1, resultan ser los mejores estimadores lineales e insesgados, pues son de mínima varianza. Lo anterior se garantiza por el teorema de Gauss-Markov.

El supuesto de homocedasticidad no influye para que estos estimadores sean insesgados o consistentes. Sin embargo como la matriz de varianza covarianza del estimador de MCO es $\sigma^2(X'X)^{-1}$, se puede ver que los errores estándar de los estimadores sí dependen de una varianza constante.

Si hay presencia de heterocedasticidad se podría sobreestimar o subestimar la varianza de $\widehat{\boldsymbol{\beta}}$, y por tanto, la inferencia que se hace, intervalos de confianza, pruebas t y F tradicionales, a partir de ellos, no son confiables.

1.3. Causas de la heterocedasticidad

Existen diversas razones por las que puede producirse la heterocedasticidad de los errores. La más inmediata y frecuente es la de que a valores mayores de alguna o algunas de las variables explicativas del modelo de regresión, puede crecer la dispersión absoluta e incluso la relativa, (Román, 2001).

Una situación muy usual en la cual se presenta heterocedasticidad es analizando gastos en las familias, puesto que se encuentra que hay mayor variación en el gasto de ciertos grupos de productos entre familias de altos ingresos que entre familias de baja renta debido a una mayor elección permitida por los altos ingresos,(Prais, 1971). Esto es, si estamos analizando la relación entre consumo y renta podemos esperar que a medida que aumente la renta aumente σ_i^2. Una familia con mayor renta tiene mayores posibilidades de consumo, no sólo consumir más variedad de productos, sino que aumentará el valor del consumo real.

Otra situación en la cual se puede presentar heterocedasticidad es en los modelos que tengan en cuenta expectativas, donde la expectativa es una medida de lo que un agente espera que ocurra. La formación de esa medida conlleva un proceso de aprendizaje, por tanto, se puede esperar que los agentes aprendan de sus errores y según avance el tiempo se confundan menos, en este caso σ_i^2 se reducirá.

También es frecuente que las empresas con mayores beneficios puedan presentar mayor variabilidad en sus políticas de dividendos. Si las ganancias son muy bajas simplemente no podrán repartir dividendos. Incluso, si se tiene en cuenta las diferencias en los tamaños de las empresas, se espera observar una variación más grande en los beneficios de las empresas grandes que en las pequeñas. La varianza en los beneficios podría depender también de la diversificación del producto, gastos de investigación y desarrollo, y características de la industria y, por tanto, podría también variar entre las empresas de tamaños similares. En el campo de la econometría se dice que en general cuando se trabaja con datos de corte transversal es usual la presencia de heterocedasticidad, (Román, 2001).

Otra causa de heterocedasticidad puede encontrarse en la mala especificación de un modelo. Si en un modelo se ha omitido una variable relevante, su exclusión puede llevar a pensar que existe heterocedasticidad en las perturbaciones del modelo. Por ejemplo, si consideramos la función de demanda de un producto y excluimos los precios de los bienes complementarios a él o de sus competidores, los estimadores MCO serán sesgados y el estudio de los residuos mínimo cuadráticos del modelo puede dar la impresión de que la varianza de la perturbación no es constante. Si incluimos la variable o variables omitidas la impresión puede desaparecer. En este caso la solución al problema pasa por especificar correctamente el modelo.

La desigualdad en las varianzas también surge en ocasiones por la presencia de datos atípicos. La inclusión o exclusión de información de este tipo, en particular para muestras pequeñas, puede ocasionar la alteración en los resultados de la regresión.

En econometría se menciona también como causa de heterocedasticidad la asimetría de la distribución de una o mas de las variables regresoras incluidas en el modelo. Por ejemplo, la distribución del ingreso es desigual pues la mayor parte de él, le corresponde a algunos individuos de las clases sociales más altas, (Gujarati, 2010).

Otra fuente de heterocedasticidad señalada por algunos autores es la que se debe a la incorrecta transformación de los datos o a una forma funcional incorrecta.

Usualmente no se esta seguro que los datos presenten heterocedasticidad y cuál es la forma que ésta toma si la presenta. Entonces, es útil poder probar homocedasticidad y si se hace necesario, modificar el procedimiento de estimación de forma adecuada.

1.4. Pruebas de homocedasticidad

Para analizar la presencia de heterocedasticidad se pueden utilizar dos técnicas complementarias entre sí: el análisis gráfico de los residuos de MCO y las pruebas o test formales: Breusch-Pagan, Golfeld-Quandt, White, Glejser, entre otros.

Cuando no hay información previa acerca de la presencia de heterocedasticidad o de la explicación de la misma, un análisis informal basado en la gráfica de los residuos $\widehat{\varepsilon_i}$, frente a una de las variables explicativas, o en la gráfica de los $\widehat{\varepsilon_i}^2$, frente a los $\widehat{y_i}$, podría ser útil para detectarla. Los métodos gráficos utilizan el hecho que los mínimos cuadrados son estimadores consistentes del término de perturbación. Si se obtienen diferencias entre las varianzas de los residuos para diferentes niveles de la variable explicativa posiblemente hay heterocedasticidad.

Como no se conoce el comportamiento que tiene la heterocedasticidad, se pretende en general que con el método gráfico se pueda identificar algún patrón sistemático. Como lo menciona(Gujarati, 2010), un patrón puede sugerir cuando la varianza del término de perturbación se relaciona de cierta forma con la variable X, esto permitiria saber como transformar los datos tal que la varianza de las perturbaciones sea homocedastica en la regresión sobre los datos transformados. También es posible que no haya ningún patrón sistemático entre las variables lo cual indicaría que posiblemente no hay presencia de ella.

Se han sugerido varios tipos de pruebas formales para detectar la ausencia del supuesto de homocedasticidad; todas ellas tienen en común la hipótesis nula

$$H_0 = \{\sigma_1^2 = ... = \sigma_n^2\}$$

Las pruebas de homocedasticidad con enfoque general se basan en la idea de que la presencia de heterocedasticidad de cualquier tipo en los errores, será discernible del comportamiento de los residuos. Las pruebas que permiten establecer si existe o no heterocedasticidad y la forma concreta de su comportamiento, y por tanto las bases para estimar la matriz Ω, son denominadas constructivas.

Las pruebas basadas en los residuales son robustas en el sentido que detectarán heterocedasticidad de gran variedad de formas. Las pruebas de Golfeld y Quandt, y Breusch-Pagan son algunas. Todas ellas basadas en el comportamiento de los mínimos cuadrados de los residuos. Así pues, la hipótesis alternativa para el enfoque general es que existe algún tipo de heterocedasticidad. Un caso de procedimiento más robusto para la detección de heterocedasticidad es la prueba de White; prueba que no requiere determinar a priori las variables que explican la heterocedasticidad.

Cuando se sospecha cual es la forma funcional concreta de la heterocedasticidad se puede considerar como hipótesis alternativa alguna que dependa del procedimiento de estimación considerado y del modelo explícito de heterocedasticidad que se asuma. De esta forma la prueba va a ser más potente; siendo conveniente encontrar la relación subyacente entre σ_i y las variables explicativas del modelo u otras.

Dentro de las formas de la heterocedasticidad se encuentran los modelos aditivos de la varianza; en los cuales la varianza o desviación estándár de las perturbaciones aumenta o disminuye proporcionalmente a alguna otra variable o conjunto de variables, o la varianza es una combinación lineal de una variable exógena o alguna otra variable ajena al modelo. Otra forma que se presenta es un tipo de heterocedasticidad por grupos, esta se puede interpretar como un caso particular del modelo aditivo donde la variable exógena que se analiza es cualitativa. También se encuentra la de los modelos ARCH donde la varianza de las perturbaciones depende de la varianza del periodo anterior.

Obtener una forma explícita para la heterogeneidad de la varianza en el modelo de regresión permite reducir la cantidad de parámetros desconocidos a estimar.

1.4.1. Prueba de Goldfeld-Quandt

(Goldfeld & Quandt, 1965) supone que el valor de σ_i^2 depende de una variable, generalmente una de las variables explicativas aunque no necesariamente. Esta dependencia se supone de la forma

$$E[\varepsilon_i^2] = \sigma_i^2 = \sigma^2 X_i^2 \tag{1.8}$$

donde σ^2 es una constante y X_i es la variable que su cuadrado se supone proporcional a la varianza de los errores. Para probar esto formalmente (Goldfeld & Quandt, 1965) proponen el siguiente procedimiento.

1. Las observaciones de Y se ordenan de acuerdo con los valores de X_i empezando con el valor más pequeño de X_i
2. Se omiten c observaciones centrales.
3. Se estima dos veces el modelo original, uno con las primeras $(n-c)/2$ observaciones muestrales, y otro con las $(n-c)/2$ observaciones muestrales restantes. Cada grupo de observaciones $(n-c)/2$ debe ser lo suficientemente grande, para que sea posible hacer las estimaciones.
4. Se obtienen las respectivas sumas de cuadrados residuales de las regresiones SCR_1 y SCR_2.

   Bajo el supuesto de homocedasticidad y normalidad para el error aleatorio el estadístico $GQ = \frac{SCR_1/gl}{SCR_2/gl}$ sigue una distribución F con $gl = \frac{n-c}{2} - k$ grados de libertad en el numerador y en el denominador.
5. Si existe heterocedasticidad del tipo (1.8), entonces el estadístico GQ es superior al valor crítico F, dado un nivel de significancia.

(Goldfeld & Quandt, 1965) sugieren que cuantas mas observaciones se omitan, en cualquier caso, menos grados de libertad habrá para la estimación en cada grupo, lo que tendera a disminuir la potencia del contraste, esto es, valores grandes para c incrementan la potencia de la prueba por medio de un incremento en el estadístico F , pero disminuye la potencia por la reducción de los grados de libertad. (Judge, 1988) sugiere por experiencia, que el número de observaciones que se deben omitir es $c = 4$ para $n = 30$, $c = 10$ para $n = 60$; adicionalmente agrega que no es necesario que los dos grupos tengan el mismo tamaño.

1.4.2. Prueba de Park

(Park, 1966) sugiere dos propuestas de heterocedasticidad. La primera, que la varianza del término del error es proporcional al cuadrado de la variable explicativa X_i, esto es, $\sigma_i^2 = \sigma^2 X_i^2$. La segunda una estructura más general,

$$\sigma_i^2 = \sigma^2 X_i^{\gamma} e^{\nu} \tag{1.9}$$

o de forma equivalente

$$\ln[\sigma_i^2] = \ln \sigma^2 + \gamma \ln X_i + \nu \tag{1.10}$$

donde ν es el error aleatorio.

Park Sugiere que se usen los residuales $\widehat{u_i}^2$ de la estimación de mínimos cuadrados ordinarios como estimaciones de σ_i^2, para obtener una estimación de γ mediante una regresión del modelo $\ln \widehat{u_i}^2 = \ln \sigma^2 + \gamma \ln X_i + \nu$. Si γ resulta estadísticamente significativo, entonces se presenta heterocedasticidad de los datos.

1.4.3. Prueba de Rutemiller-Bowers

(Rutemiller & Bowers, 1968) Considera un conjunto de n variables aleatorias $\{Y_i\}$ con distribución normal de parámetros μ_i, σ_i tal que,

$$\begin{aligned} \mu_i &= \beta_0 + \sum_{j=1}^{k} \beta_j X_{ij} \\ \sigma_i &= \gamma_0 + \sum_{j=1}^{k} \gamma_j X_{ij} \end{aligned} \tag{1.11}$$

Esto es, distribución normal con modelado de media y desviación estándar, en el que la propuesta de modelo para la desviación estándar es un modelo lineal. La estimación de los parámetros $\beta_0, ..., \beta_k$ y $\gamma_0, ..., \gamma_k$ se propone por el método de máxima verosimilitud, en el cual se utiliza el "método de scoring"para resolver el sistema de ecuaciones resultante. Para probar el modelo de manera formal, se realiza una prueba de razón de verosimilitud del modelo obtenido frente al modelo homocedastico.

1.4.4. Prueba de Glejser

(Glejser, 1969) pone en consideración modelos para la varianza no necesariamente lineales. En estos toma como variable a explicar el valor absoluto de los errores o su cuadrado, que entre otras garantiza que la medida de variabilidad de los errores sea positiva.

La prueba consiste en:

1. Estimar el modelo original por Mínimos cuadrados ordinarios y obtener sus residuos.

2. Utilizar como variable a explicar $|\widehat{\varepsilon_i}|$ o $\widehat{\varepsilon_i}^2$ y estimar una regresión de ellos sobre una potencia (h) de la variable z_i para los distintos valores de h, con $h = \{-1, -1/2, 1/2, 1\}$. esto es,

$$|\widehat{\varepsilon_i}| = \alpha_0 + \alpha_1 z_i^h + v_i$$

donde z_i es una variable que en la mayoría de los casos corresponde a una de las variables explicativas del modelo original. Si α_1 resulta significativo estadísticamente, esto indicaría presencia de heteroscedasticidad.

3. Seleccionar el valor de h, que proporcione la mejor regresión.

4. Después de establecer el valor de h, se dividen las observaciones Y_i del modelo original entre $\alpha_0 + \alpha_1 z_i^h$ para el caso de haber estimado $|\widehat{\varepsilon_i}|$. Si se estimó $\widehat{\varepsilon_i}^2$ se divide por $\sqrt{\alpha_0 + \alpha_1 z_i^h}$. Luego se estima nuevamente por MCO.

5. Escoger el valor de la potencia (h) de z que proporcione el mejor modelo de regresión.

1.4.5. Prueba de White

(White, 1980) propone un estimador de varianza covarianza que puede usarse de forma apropiada para mínimos cuadrados ordinarios en presencia de heterocedasticidad. Este estimador no utiliza una forma específica de heterocedasticidad. Además, lo utiliza para crear una prueba de homocedasticidad al realizar un contraste de este estimador con el estimador de varianza-covarianza habitual.

La matriz de varianza covarianza correcta de (White, 1980) para MCO en presencia de heterocedasticidad es

$$Var[\widehat{\boldsymbol{\beta}}] = \sigma^2[X'X]^{-1}[X'\Omega X][X'X]^{-1} \tag{1.12}$$

Que puede estimarse por

$$\widehat{Var}[\widehat{\boldsymbol{\beta}}] = [X'X]^{-1}[\sum_{i=1}^{n} \widehat{\varepsilon_i}^2 x_i x_i'][X'X]^{-1} \tag{1.13}$$

donde, $\widehat{\varepsilon_i} = y_i - X_i\widehat{\boldsymbol{\beta}}$

Si no existe heterocedasticidad, el estimador convencional $V = s^2[X'X]^{-1}$ será un estimador consistente de $\widehat{Var}[\widehat{\boldsymbol{\beta}}]$ mientras que sí existe no lo será.

De forma operativa la prueba consiste en:

1. Estimar el modelo original por mínimos cuadrados ordinarios

2. Estimar un modelo de regresión de los cuadrados de los errores del modelo original. Para este nuevo modelo se incluye la constante, las variables explicativas del modelo original, sus cuadrados y los productos cruzados de segundo orden.

3. Se toman como hipótesis de la prueba:

$$H_0 : \sigma_t^2 = \sigma^2$$

$$H_1 : \sigma_t^2 \neq \sigma^2$$

4. Se utiliza como estadístico de prueba al producto de una constante T y el coeficiente de determinación de la nueva regresión. T corresponderá al tamaño muestral. Este estadístico sigue una distribución chi-cuadrada con $p-1$ grados de libertad, siendo p el número de variables explicativas del nuevo modelo.

5. Si el valor de $TR^2 > \chi^2_{p-1}$ al nivel de significancia escogido, entonces se rechaza la hipótesis nula de homocedasticidad.

Algunos autores incluido el propio White resaltan de esta prueba la no consideración de una forma particular de la varianza. Sin embargo, (Greene, 2003) manifiesta que esta ventaja trae consecuencias para determinar la potencia de la prueba excepto en el contexto de un problema específico. Además este mismo autor afirma que este contraste no es constructivo en la medida en que en presencia de heterocedasticidad, el contraste no indica qué hacer después. (Cinca, 1996)resalta además que esta prueba requiere que el coeficiente de determinación de la nueva regresión tienda a cero rápidamente.

1.4.6. Prueba de Breusch Pagan

(Breusch & Pagan, 1979) propone una prueba para la heterocedasticidad con las propiedades asintóticas (consistencia, eficiencia asintótica, normalidad asintótica e invarianza), del test de razón de verosimilitud y un enfoque basado en el contraste de multiplicadores de Lagrange (LM). Para esta prueba se supone:

$$\sigma_t^2 = h(z_t'\alpha) = h(\alpha_0 + \alpha_1 z_{1t} + \alpha_2 z_{2t} + ... + \alpha_p z_{pt})$$

h con primera y segunda derivada. Y la varianza del término del error dependiendo de un vector de variables z_t de dimensión p, asociadas al cambio de varianza.

Se toman como hipótesis de la prueba:

$$H_0 : \alpha_1 = \alpha_2 = ... = \alpha_p = 0$$

$$H_1 : \exists \alpha_i \neq 0, \text{ para } j = 1, 2, ..., p$$

De acuerdo al contraste de multiplicadores de Lagrange (conocido también como el test de Score), se estima el modelo restringido y se obtienen sus residuos, luego se estima una regresión de estos contra X, se obtiene el R^2 y se construye el estadístico. De aquí que esta prueba consista en:

- Estimar el modelo original por MCO y obtener los residuos correspondientes.
- Se construye la variable de aproximación de la varianza: $\widehat{e}_t^2 = \frac{\widehat{\varepsilon}}{\widehat{\sigma}^2}$, siendo $\widehat{\sigma}^2$, la estimación de MV de la varianza del termino del error de la regresión original.
- Se estima una regresión auxiliar de $\widehat{e}_t^2 = \alpha_0 + \alpha_1 z_{1t} + \alpha_2 z_{2t} + ... + \alpha_p z_{pt}$

- Se obtiene la suma explicada de la regresión auxiliar.
- Se define a $LM = \frac{1}{2}$ de la suma explicada de $\widehat{e}_t^2$. Éste se distribuye χ^2 con p grados de libertad en condiciones de homocedasticidad y suponiendo que el error se distribuye normal.
- Si $LM > \chi_p^2$ al nivel de significancia escogido, se rechaza la hipótesis nula de homocedasticidad.

Como lo menciona (Breusch & Pagan, 1979), su propuesta se caracteriza por ser constructiva en el sentido en que da una forma específica a la heterocedasticidad en la hipótesis alterna. Acerca de la interpretación de esta prueba,(Cinca, 1996)considera que ésta está en la homocedasticidad de los residuos, pues si es así entonces las variables asociadas al cambio de varianza no explican a los residuos transformados y en tanto la suma explicada de la regresión auxiliar debería ser pequeña.

Debido a la sensibilidad del supuesto de normalidad para (Breusch & Pagan, 1979), (Koenker, 1981) sugiere que se calcule LM con un estimador de la varianza de ε_t^2 más robusto.

$$V = \frac{1}{n}\sum_{i=1}^{n}[e_i^2 - \frac{e'e}{n}]^2$$

De esta forma este estadístico se comporta como el propuesto por (Breusch & Pagan, 1979) pero sin el requisito de la normalidad.

1.5. Estimación del Modelo de regresión lineal en presencia de heterocedasticidad

Para tratar la falta de eficiencia de los estimadores de MCO para un modelo de regresión lineal con presencia de heterocedasticidad se describen a continuación los enfoques para esta situación.

Un primer enfoque consiste en definir una matriz P diagonal cuyo i-ésimo elemento es $\frac{1}{\sqrt{\omega_i}}$. De tal forma que $PP' = \Omega^{-1}$. Multiplicando a esta matriz por el modelo original se tiene

$$PY = P(X\boldsymbol{\beta} + \varepsilon) \tag{1.14}$$

de esta forma las variables del modelo original quedan transformadas así:

$$Y^* = \frac{Y_i}{\sqrt{\omega_i}}, \quad X^* = \frac{X_i}{\sqrt{\omega_i}}, \quad \varepsilon^* = \frac{\varepsilon_i}{\sqrt{\omega_i}} \tag{1.15}$$

Obteniendo el modelo

$$Y^* = X_i^*\boldsymbol{\beta} + \varepsilon^* \tag{1.16}$$

Como ε^* cumple con las condiciones del teorema de Gauss Markov (Verbeek, 2008), entonces el modelo puede ser estimado usando MCO.

Desde este enfoque, si reemplazamos en el estimador $\widehat{\boldsymbol{\beta}}_{MCO}$ a las variables transformadas obtenemos:

$$\widehat{\boldsymbol{\beta}}_{MCP} = (X'^{*}X^{*})^{-1}X'^{*}Y^{*} \tag{1.17}$$

Este se conoce como el estimador de mínimos cuadrados ponderados.

Como $X^{*} = PX$ y $Y^{*} = PY$, reemplazando esto en $\widehat{\boldsymbol{\beta}}_{MCP}$ se obtiene

$$\widehat{\boldsymbol{\beta}}_{MCG} = (X'\Omega^{-1}X)^{-1}X'\Omega^{-1}Y \tag{1.18}$$

conocido como estimador de mínimos cuadrados generalizados.

Este método de estimación aprovecha la información dada por la variabilidad desigual de la variable dependiente Y, dando ponderaciones diferentes a las distintas observaciones muestrales (Cinca, 1996).$\widehat{\boldsymbol{\beta}}$ de MCG es conocido como estimador de Aitken. Aitken presenta una generalización del teorema de Gauss Markov para el modelo de regresión lineal general.

La dificultad para el cálculo de $\widehat{\boldsymbol{\beta}}_{MCG}$ radica en su dependencia de Ω, matriz que es desconocida. Para ello se estima Ω y se sustituye en la expresión del estimador de MCG. Obteniendo

$$\widehat{\boldsymbol{\beta}}_{MCGF} = (X'\widehat{\Omega}^{-1}X)^{-1}X'\widehat{\Omega}^{-1}Y \tag{1.19}$$

Conocido como el $\widehat{\boldsymbol{\beta}}$ de mínimos cuadrados generalizados factibles

Como ya se ha mencionado, si Ω es desconocida, la cantidad de parámetros que se debe estimar obliga a que habitualmente se deba suponer algún tipo de forma funcional que sigue la heterocedasticidad para reducir el número de parámetros a estimar. Suponer que la heterocedasticidad tiene una forma funcional implica que existen una o más variables que afectan a la varianza de las perturbaciones. Se denomina a ese conjunto de variables "variables heteroauxiliares", o también "variables independientes de la ecuación de la varianza".

De acuerdo con (Cinca, 1996), de forma general, la estimación utilizando MCG sigue este procedimiento:

1. Estimar por MCO en presencia o no de heterocedasticidad.
2. Proponer una forma o estructura de la varianza.
3. Realizar una estimación de la forma propuesta de σ_i^2 utilizando los residuos de MCO
4. Dividir a cada observación por la estimación de σ_i
5. Calcular de nuevo MCO con las variables transformadas.

Si el término de error del modelo sigue una distribución normal entonces el estimador de MCG coincide con el de maxima verosimilitud y como el modelo transformado cumple con los supuestos básicos,la forma de hacer inferencia se mantiene.

Se ha mencionado que el supuesto de homocedasticidad no influye para que los estimadores de MCO sean insesgados o consistentes, sin embargo, el problema es la inferencia que se hace a partir de ellos, pues dejan de ser eficientes.

El estimador de mínimos cuadrados generalizados es eficiente siempre y cuando la heterocedasticidad sea de la forma que se ha supuesto al diseñar tal estimador. Sin embargo, (Greene, 2003) menciona que es posible el caso en que la forma de la heterocedastidad resulte desconocida. Un error en el supuesto de la forma de la heterocedasticidad trae también como consecuencia una incorrecta estimación de la matriz de varianza covarianza (Cinca, 1996). Además, la aplicación de métodos apropiados para esta, requiere una especificación bastante buena de Ω.

Aparece otro enfoque para darle solución al problema de la heterocedasticidad, conocido como el estimador de White de la matriz de varianza covarianza para hacer inferencia, en el que no es necesario conocer la forma de σ_i^2.

Como lo resalta (White, 1980), su estimador no requiere un modelo formal de la estructura de la heterocedasticidad ya que usa las variables explicativas y la estimación de mínimos cuadrados de residuales para su cálculo. Dando además condiciones adicionales que permiten al investigador probar la presencia de heterocedasticidad.

La matriz de varianza covarianza habitual bajo el supuesto de homocedasticidad es

$$Var[\widehat{\boldsymbol{\beta}}] = \sigma^2[X'X]^{-1} \tag{1.20}$$

Siendo esta inapropiada ahora. Como lo menciona (Greene, 2003), la matriz que resulta ser apropiada es (1.12), matriz que requiere de Ω para su estimación. De acuerdo con (White, 1980), esta matriz puede estimarse con el estimador de White(1.13).

Para darle uso a este estimador, se puede hacer inferencia empleando por ejemplo, un contraste conocido como el contraste de Wald, especificado como:

$$H_0 : R\boldsymbol{\beta} = r$$
$$H_1 : R\boldsymbol{\beta} \neq r$$

tal que,

$$(R\widehat{\boldsymbol{\beta}} - r)'[R(\widehat{Var}(\widehat{\boldsymbol{\beta}})R']^{-1}(R\widehat{\boldsymbol{\beta}} - r) \sim \chi_j^2 \tag{1.21}$$

donde $\widehat{Var}(\widehat{\boldsymbol{\beta}})$ es el estimador de White y R es una matriz de $j \times k$ constantes conocidas (ceros y unos), cuyo objetivo es seleccionar los parámetros a probar.

En conclusión, ¿Cómo estimar entonces el modelo en presencia de heterocedasticidad?. MCO es insesgado pero no de mínima varianza, incumpliendo el teorema de Gauss Markov. El paso a seguir seria entonces usar MCG, si se conoce la estructura de la varianza, camino que conduce al uso del estimador de MCP. Una forma alternativa es usar el estimador de MCO y el estimador de White de la matriz de varianza covarianza de $\boldsymbol{\beta}$, camino que no requiere dar una forma específica a la varianza y que además permite hacer inferencia. Sin embargo, como lo menciona (Greene, 2003), ¿qué es preferible?, ¿tener ponderaciones inadecuadas o no hacer ponderaciones?. Aparece entonces otro enfoque que es la estimación de máximo verosimilitud con heterocedasticidad, considerando un caso general, por ejemplo, de tipo multiplicativo. Este modelo, desarrollado en la siguiente sección, es un

modelo general y flexible de la varianza que incluye la mayoría de las formulaciones útiles como casos especiales.

En las últimas dos decadas, los econometristas han aprendido a ajustar los errores estándar para que las estadísticas t, F, entre otros, sean validos en presencia de heteroscedasticidad con forma desconocida. Esto es conveniente porque quiere decir que se puede reportar nuevas estadísticas de trabajo, sin importar el tipo de heteroscedasticidad presente en la población.

Varios autores han propuesto estimadores consistentes de la matriz de varianza covarianza bajo heteroscedasticidad con forma desconocida. Las pruebas basadas en una matriz de covarianza consistente heterocedastica (HCCM), incluso son consistentes en presencia de heterocedasticidad de forma desconocida. Muchas aplicaciones que usan HCCM parecen depender de la versión asintótica conocida como HC0. En (Long, 2000) vía simulación de montecarlo se muestra que HC0 en ocasiones resulta en inferencias incorrectas cuando $n \leq 250$, mientras que tres versiones poco conocidas de HCCM para muestras pequeñas, y especialmente la versión conocida como HC3, trabajan bien, incluso para muestras de tamaño inferior a 25.

1.6. Heterocedasticidad Multiplicativa

Dentro de los procedimientos para detectar heteroscedasticidad, como se menciona en la sección 1.4, la varianza del termino del error puede tomar diversas estructuras; Entre ellas, que la varianza es proporcional a una de las variables explicativas del modelo de regresión elevada a cierta potencia. Esta forma, por ejemplo, fue considerada por algunos autores, entre ellos (Park, 1966), (Goldfeld & Quandt, 1965) (ver sección 1.4.1, 1.4.2).

El modelo de regresión lineal normal heteroscedastico desarrollado por (Harvey, 1976), denominado heteroscedasticidad multiplicativa, considera una formulación de la estructura de la varianza más general

$$\sigma_i^2 = \sigma^2 \exp \mathbf{q}'_i \alpha \tag{1.22}$$

donde $\boldsymbol{q}_i$ es un vector $p \times 1$ de observaciones de un conjunto de variables que generalmente son las explicativas del modelo original pero no necesariamente, y $\boldsymbol{\alpha}$ es un vector $p \times 1$ de parámetros.

Una simplificación para la formulación (1.22) es sugerida por Godfrey y presentada por (Greene, 2003), como sigue:

Sea $\boldsymbol{w}'_i = [1, \mathbf{q}_i]$ y $\boldsymbol{\gamma} = [\ln \sigma^2, \alpha]$. Luego el modelo se puede escribir como

$$\sigma_i^2 = \exp \boldsymbol{w}'_i \boldsymbol{\gamma} \tag{1.23}$$

Esta expresión indica que la varianza σ_i^2 es una función exponencial de una combinación lineal de variables conocidas y garantiza que σ_i^2 sea positivo.

1.6.1. Estimación del modelo de regresión lineal normal con heterocedasticidad multiplicativa

El método de los mínimos cuadrados puede usarse para estimar los parámetros de un modelo de regresión lineal sin preocuparse por la distribución de los errores. Si se conoce como se distribuyen los errores, se puede hacer la estimación de parámetros vía "método de maxima verosimilitud".

Considérese el modelo de regresión normal con covariables para el modelado de la heterocedasticidad de la varianza como sigue

$$\begin{aligned} Y_i &= \boldsymbol{\beta}'\boldsymbol{x}_i + \varepsilon_i \\ E[\varepsilon_i^2] &= \sigma_i^2 = \exp(\boldsymbol{\gamma}'\boldsymbol{w}_i) \end{aligned} \tag{1.24}$$

donde $\boldsymbol{w}_i' = (1, w_{i1}, ..., w_{ip})$ es un vector $((p+1)\times 1)$ que contiene valores de las variables no estocásticas que explican la heterocedasticidad del modelo 1.24 y $\boldsymbol{\gamma} = (\gamma_0, \gamma_1, ..., \gamma_p)'$ es un vector $((p+1)\times 1)$ de coeficientes desconocidos del modelo de heteroscedasticidad.

Para la estimación de máxima verosimilitud de los parámetros $\boldsymbol{\delta} = [\boldsymbol{\beta}, \boldsymbol{\gamma}]$ del modelo de regresión lineal normal con heterocedasticidad multiplicativa, sea $(y_i, \boldsymbol{x}_i, \boldsymbol{w}_i)$, $i = 1, ..., n$ una muestra aleatoria que proviene del modelo (1.24).

Como los errores en el modelo (1.24) son independientes e idénticamente distribuidos, la función de verosimilitud es

$$L(\boldsymbol{\delta}) = \prod_{i=1}^{n} \frac{1}{\exp(\boldsymbol{\gamma}'\boldsymbol{w}_i)} \exp\left(-\frac{y_i - \boldsymbol{\beta}'\boldsymbol{x}_i)^2}{2\exp(\boldsymbol{\gamma}'\boldsymbol{w}_i)}\right) \tag{1.25}$$

y la función de log-verosimilitud

$$\log L(\boldsymbol{\delta}) = -\frac{1}{2}\sum_{i=1}^{n} \log \sigma_i^2 - \frac{1}{2}\sum_{i=1}^{n} \frac{(y_i - \boldsymbol{\beta}'\mathbf{x}_i)^2}{\exp(\boldsymbol{\gamma}'\boldsymbol{w}_i)} \tag{1.26}$$

Encontrar la estimación de $\boldsymbol{\delta} = [\boldsymbol{\beta}, \boldsymbol{\gamma}]$ que maximice la función de log-verosimilitud (1.26), plantea un problema de optimización que analíticamente no es posible resolver, por tanto, es necesario usar un método iterativo.

Para el proceso iterativo hacia un punto máximo, se dispone del método de Newton-Raphson, uno de los más conocidos, que consiste en hacer una aproximación cuadrática, mediante la expansión de Taylor de segundo orden, de la función de log-verosimilitud del modelo y encontrar el máximo de esta.

Sea el **score** $\nabla \log L(\boldsymbol{\delta})$ el vector de derivadas parciales de primer orden y la matriz de información observada $\boldsymbol{H}_{\log L(\boldsymbol{\delta})}$ la matriz hessiana de $-\log L(\boldsymbol{\delta})$. La expansión de Taylor de segundo orden alrededor del punto $\boldsymbol{\delta}_t$ es:

$$\log L(\boldsymbol{\delta}) \approx \log L(\boldsymbol{\delta}_t) + (\boldsymbol{\delta} - \boldsymbol{\delta}_t)\nabla \log L(\boldsymbol{\delta}_t) + \frac{1}{2}(\boldsymbol{\delta} - \boldsymbol{\delta}_t)'\boldsymbol{H}_{\log L(\boldsymbol{\delta})}(\boldsymbol{\delta}_t)(\boldsymbol{\delta} - \boldsymbol{\delta}_t) \tag{1.27}$$

Si la aproximación de $\log L(\boldsymbol{\delta})$ es buena, un máximo relativo de $\log L(\boldsymbol{\delta})$ se podría aproximar por un máximo relativo de la expansión de Taylor de segundo orden. Desarrollando el gradiente de la aproximación cuadrática, sustituyendo $\boldsymbol{\delta}$ por $\boldsymbol{\delta}_{t+1}$, igualando a $\mathbf{0}$ y resolviendo para $\boldsymbol{\delta}_{t+1}$ se tiene que

$$\boldsymbol{\delta}_{t+1} = \boldsymbol{\delta}_t - \boldsymbol{H}^{-1}_{\log L(\boldsymbol{\delta})}(\boldsymbol{\delta}_t)(\nabla \log L(\boldsymbol{\delta}_t))' \tag{1.28}$$

La expresión (1.28) se usa como una ecuación de recurrencia para que a partir de un punto inicial, genere una sucesión de puntos que converjan al máximo local de $\log L(\boldsymbol{\delta})$.

Existen dos problemas con el método de Newton. Por un lado, puede ser costoso computacionalmente al calcular la matriz de información observada. Por otro lado, el método de Newton no es un algoritmo ascendente, es decir que no necesariamente $\log L(\boldsymbol{\delta}_{t+1}) > \log L(\boldsymbol{\delta}_t)$. Para obtener un algoritmo ascendente se puede reemplazar la matriz de información por una matriz definida positiva A_t como se observará a continuación.

Con la sustitución de la matriz de observación por una matriz definida positiva en (1.28), $\Delta\boldsymbol{\delta}_t = A_t^{-1}(\nabla \log L(\boldsymbol{\delta}_t))'$ es ahora el incremento en cada iteración, que si es suficientemente contractivo, obliga a incrementar a $\log L(\boldsymbol{\delta})$. Considérese la siguiente expansión de Taylor de primer orden

$$\begin{aligned} \log L(\boldsymbol{\delta}_t + \alpha\Delta\boldsymbol{\delta}_t) - \log L(\boldsymbol{\delta}_t) &= (\nabla \log L(\boldsymbol{\delta}_t))\alpha\Delta\boldsymbol{\delta}_t + o(\alpha) \\ &= \alpha(\nabla \log L(\boldsymbol{\delta}_t))A_n^{-1}(\nabla \log L(\boldsymbol{\delta}_t))' + o(\alpha) \end{aligned} \tag{1.29}$$

Donde la razón de error $\frac{o(\alpha)}{\alpha} \to 0$ cuando la constante de contracción positiva $\alpha \to 0$.

Por propiedades de matriz definida positiva se cumple que la inversa de una matriz definida positiva, también es matriz definida positiva y que para todos los vectores no nulos $\nabla \log L(\delta_t)$, $(\nabla \log L(\boldsymbol{\delta}_t))A_n^{-1}(\nabla \log L(\boldsymbol{\delta}_t))' > 0$. Por tanto, la modificación de la matriz de información observada por una matriz definida positiva conlleva a un algoritmo de ascenso.

Existen muchas formas de aproximar la matriz de información observada. El **algoritmo de Fisher Scoring** reemplaza la matriz de información observada por la matriz de información esperada $E(\boldsymbol{H}_{\log L(\boldsymbol{\delta})}(\boldsymbol{\delta}))$, también denominada matriz de información de fisher esperada. La forma alternativa $E(\boldsymbol{H}_{\log L(\boldsymbol{\delta})}(\boldsymbol{\delta})) = Var[\nabla \log L(\boldsymbol{\delta})]$ muestra que es una matriz no negativa.

De lo anterior la ecuación general de iteraciones para obtener la estimación de $\boldsymbol{\delta}$ por máxima verosimilitud mediante el algoritmo Fisher scoring es:

$$\boldsymbol{\delta}_{t+1} = \boldsymbol{\delta}_t - E(\boldsymbol{H}_{\log L(\boldsymbol{\delta})})^{-1}(\boldsymbol{\delta}_t)(\nabla \log L(\boldsymbol{\delta}_t))' \tag{1.30}$$

A continuación se presenta cada uno de los elementos de la ecuación 1.30 y las ecuaciones de iteraciones resultantes para estimar δ.

El calculo del vector gradiente $\nabla \log L(\boldsymbol{\delta})$ es:

$$\left[\frac{\partial \log L}{\partial \boldsymbol{\beta}}, \frac{\partial \log L}{\partial \boldsymbol{\gamma}}\right] = \left[\sum_{i=1}^{n} \boldsymbol{x}_i \frac{(y_i - \boldsymbol{\beta}'\boldsymbol{x}_i)^2}{\exp(\boldsymbol{\gamma}'\boldsymbol{w}_i)}, \frac{1}{2}\sum_{i=1}^{n} \boldsymbol{w}_i \left(\frac{(y_i - \boldsymbol{\beta}'\boldsymbol{x}_i)^2}{\exp(\boldsymbol{\gamma}'\boldsymbol{w}_i)} - 1\right)\right] \tag{1.31}$$

La matriz de información observada es

$$\begin{aligned}\boldsymbol{H}_{\log L(\boldsymbol{\delta})} &= \begin{bmatrix} \frac{\partial^2 \log L}{\partial \boldsymbol{\beta} \partial \boldsymbol{\beta}'} & \frac{\partial^2 \log L}{\partial \boldsymbol{\beta} \partial \boldsymbol{\gamma}'} \\ \frac{\partial^2 \log L}{\partial \boldsymbol{\gamma} \partial \boldsymbol{\beta}'} & \frac{\partial^2 \log L}{\partial \boldsymbol{\gamma} \partial \boldsymbol{\gamma}'} \end{bmatrix} \\ &= \begin{bmatrix} \sum_{i=1}^n \frac{\boldsymbol{x}_i \boldsymbol{x}_i'}{\exp(\boldsymbol{\gamma}' \boldsymbol{w}_i)} & \sum_{i=1}^n \frac{\varepsilon_i}{\exp(\boldsymbol{\gamma}' \boldsymbol{w}_i)} \boldsymbol{x}_i \boldsymbol{w}_i' \\ \sum_{i=1}^n \frac{\varepsilon_i}{\exp(\boldsymbol{\gamma}' \boldsymbol{w}_i)} \boldsymbol{x}_i \boldsymbol{w}_i' & -\frac{1}{2} \sum_{i=1}^n \frac{\varepsilon_i^2}{\exp(\boldsymbol{\gamma}' \boldsymbol{w}_i)} \boldsymbol{w}_i \boldsymbol{w}_i' \end{bmatrix} \end{aligned} \tag{1.32}$$

Como $E(e_i^2) = \sigma_i^2$ la matriz de información de Fisher esperada que resulta es

$$E(\boldsymbol{H}_{\log L(\boldsymbol{\delta})}) = \begin{bmatrix} \sum_{i=1}^n h_{ij} \frac{\boldsymbol{x}_i \boldsymbol{x}_i'}{\exp(\boldsymbol{\gamma}_j' \boldsymbol{w}_i)} & \mathbf{0} \\ \mathbf{0} & 2 \sum_{i=1}^n \boldsymbol{w}_i \boldsymbol{w}_i' \end{bmatrix} \tag{1.33}$$

Como se puede ver el método de Scoring proporciona una forma sencilla de obtener los estimadores de máxima verosimilitud en este caso (Harvey 1976), además, el algoritmo se reduce a dos procedimientos separados para $\boldsymbol{\beta}$ y $\boldsymbol{\gamma}$ como sigue

$$\hat{\boldsymbol{\beta}}^{t+1} = \hat{\boldsymbol{\beta}}^{(t)} + (\sum_{i=1}^n e^{-\boldsymbol{w}_i' \hat{\boldsymbol{\gamma}}^{(t)}} \boldsymbol{x}_i \boldsymbol{x}_i')^{-1} \sum_{i=1}^n \boldsymbol{x}_i e^{-\boldsymbol{w}_i' \hat{\boldsymbol{\gamma}}^{(t)}} (y_i - \boldsymbol{x}_i' \hat{\boldsymbol{\beta}}^{(t)}) \tag{1.34}$$

y

$$\hat{\boldsymbol{\gamma}}^{t+1} = \hat{\boldsymbol{\gamma}}^{(t)} + (\sum_{i=1}^n \boldsymbol{w}_i \boldsymbol{w}_i')^{-1} \sum_{i=1}^n \boldsymbol{w}_i [e^{-\boldsymbol{w}_i' \hat{\boldsymbol{\gamma}}^{(t)}} (y_i - \boldsymbol{w}_i' \hat{\boldsymbol{\beta}}^{(t)})^2] - 1] \tag{1.35}$$

donde $\hat{\boldsymbol{\gamma}}^{(t)}$ y $\hat{\boldsymbol{\beta}}^{(t)}$ son las estimaciones de $\boldsymbol{\gamma}$ y $\boldsymbol{\beta}$, respectivamente, que se obtienen en la $t - esima$ iteración.

En el trabajo de (Harvey, 1976) se compara el procedimiento de dos pasos con el método de máxima verosimilitud para la estimación de los parámetros del modelo (1.24), concluyendo que son más eficientes los estimadores de los parámetros por el segundo método que por el primero. Se afirma además, que desde el punto de vista de la estimación resulta más atractivo un modelo de heteroscedasticidad multiplicativa que un modelo aditivo.

1.7. Aplicación

Existen múltiples estudios que han enfocado su atención en explicar el gasto en relación con el ingreso. De acuerdo con el desarrollo teórico además del ingreso existen una serie de factores socioeconómicos que determinan el gasto del consumidor. Entre los factores que podrían considerarse están el tamaño de la familia, la estacionalidad, la región, la raza, las características del individuo, entre otros. Como aplicación, se usaron los datos de los gastos de tarjeta de crédito de 100 individuos, seleccionados de una muestra de 13444 personas que presenta (Greene, 2003).

Modelo	P-Valor	AIC
$M1 : E[Y] = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_4$	0.001719	1037.184
$M2 : E[Y] = \beta_0 + \beta_1 X_1 + \beta_2 X_2$	0.0005301	1029.336
$M3 : E[Y] = \beta_0 + \beta_1 X_2$	9.69e-05	1023.337
$M4 : E[Y] = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + \beta_4 X_4$	0.0007952	1042.978
$M5 : E[Y] = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3$	0.0002731	1033.099
$M6 : E[Y] = \beta_0 + \beta_1 X_2 + \beta_2 X_3$	7.833e-05	1025.345
Y = Gasto en tarjeta de crédito		
X_1 = Edad, X_2 = Ingreso, $X_3 = Ingreso^2$,X_4 = propietario de su hogar		

TABLA 1.1. Modelos propuestos para la estimación del gasto con tarjetas de crédito

Dentro de las posibles variables que pueden explicar el gasto en tarjetas de crédito (Greene, 2003), para estimar su modelo, considera la edad, los ingresos promedio del individuo y una variable binaria que toma valor 0 o 1 dependiendo si es o no propietario de su hogar respectivamente. (Greene, 2003) utiliza 72 de las observaciones para las cuales el gasto es positivo. Este gasto en tarjetas de crédito corresponde al gasto mensual promedio durante 12 meses.

Tomando como variables explicativas del gasto las propuestas por (Greene, 2003), se estiman por MCO con las mismas 72 observaciones algunos otros modelos utilizando R (ver tabla 1.1). Al revisar el estadistico t de los coeficientes, la variable que en general presenta significancia individual independientemente del modelo es el ingreso. El valor F indica que en conjunto todas las variables son significativas para la explicación del gasto.

Siguiendo a (Greene, 2003), se usará la estimación del modelo con las cuatro variables regresoras propuestas inicialmente y que habitualmente son factores que explican el gasto; éste es el modelo M4 de la tabla 1.1

Sin embargo, se esperaría observar una variación más grande del gasto a medida que los niveles de ingreso aumenten. Los gráficos de los residuales frente a las variables ingresos e ingresos al cuadrado presentan un patrón característico de una regresión heterocedastica (Figura 1.1). Para valores más altos de los ingresos, los residuos son mayores.

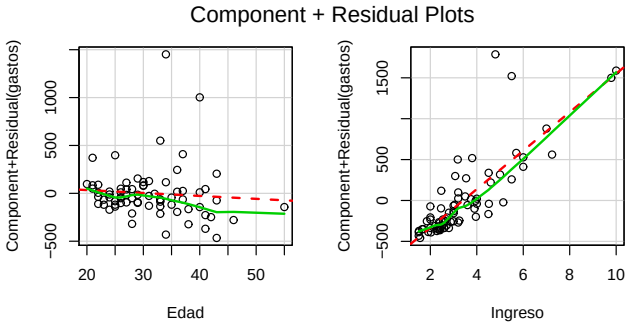
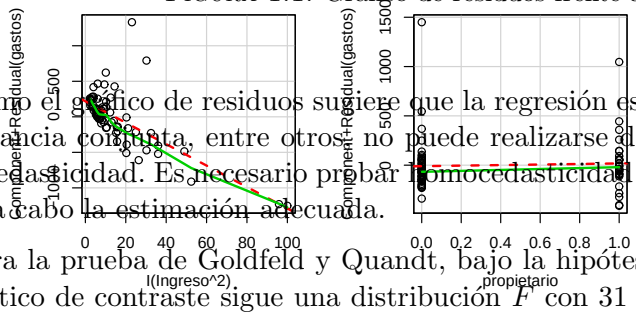


FIGURA 1.1. Gráfico de residuos frente a las variables



Como el gráfico de residuos sugiere que la regresión es heterocedástica, el contraste de significancia conjunta, entre otros, no puede realizarse dado que la prueba F se basa en homocedasticidad. Es necesario probar homocedasticidad en el modelo para luego sí poder llevar a cabo la estimación adecuada.

Para la prueba de Goldfeld y Quandt, bajo la hipótesis nula de homocedasticidad, el estadístico de contraste sigue una distribución F con 31 y 31 grados de libertad, el valor crítico de la tabla F al 5 % es 1.8221, luego el cociente de sumas residuales, $GQ = 85.6806$, entre ambas regresiones construidas para la prueba excede este valor, rechazándose la hipótesis nula.

Para la prueba de Breusch Pagan, se supone:

$$\sigma_t^2 = h(Z_t'\alpha) = h(\alpha_0 + \alpha_1 Z_{1t} + \alpha_2 Z_{2t})$$

Especificando a $Z = [1, \text{ingresos}, \text{Ingresos}^2]$, se obtiene $LM = 41,9203$. Este estadístico se distribuye χ^2 con dos grados de libertad. En este caso 41,9203 excede al valor crítico de χ_2^2 al 95 % que es de 5.99, por lo que también se rechaza la hipótesis nula, que para este contraste es la ausencia de heterocedasticidad.

Para la prueba de White, se estima un modelo de regresión de los cuadrados de los errores del modelo original. Modelo que incluye la constante, las variables explicativas y sus productos cruzados de segundo orden. La regresión de los cuadrados de los residuales de estas 13 variables produce $R^2 = 0.199013$. El estadístico es $72(0.199013) = 14.329$. Al 95 por ciento el valor crítico de chi-cuadrado con 12 grados de libertad es 21,03, por lo que la hipótesis de homocedasticidad no es rechazada por esta prueba. Dado que esta prueba no específica la forma de la estructura de la varianza,esta tiene menor potencia que las anteriores.

Metodología	Valor P
Matriz de White	4.28×10^{-8}
Matriz $HC1$	1.18×10^{-7}
Matriz $HC3$	9.95×10^{-7}
Matriz $HC4$	5.79×10^{-6}

TABLA 1.3. Significancia conjunta de las variables haciendo uso de matriz de White, $HC1$, $HC3$ y $HC4$

Reconociendo la presencia de heterocedasticidad, uno de los caminos para hacer la estimación y realizar inferencias del modelo es usar el estimador de MCO con el estimador de White de la matriz de varianza, sin embargo, como lo menciona (Long, 2000), esta matriz puede llevar a conclusiones erróneas dado que $N \leq 250$, luego se usan además las matrices $HC1$, $HC3$ y $HC4$ que permitan hacer inferencia.

	Constante	Edad	Vivienda propia	Ingresos	Ingresos^2
Coeficiente	-237,15	-3,082	27,941	234,35	-14,997
Error Estándar					
MCO	199,35	5,515	82,922	80,366	7,469
Matriz de White	212,991	3,302	92,187	88,8663	6,944
Matriz HC1	220,795	3,423	95,565	92,1226	7,199
Matriz HC3	229,574	3,604	99,314	95,481	7,476
Matriz HC4	222,569	3,549	96,141	93,046	7,393

TABLA 1.2. Estimación del modelo utilizando MCO y errores estándar bajo matrices de White, $HC1$, $HC3$ y $HC4$

Los resultados de la estimación de la matriz de White y de las matrices $HC1$, $HC3$ y $HC4$ de varianza covarianza, permite observar que los errores estándar de los coeficientes utilizando la matriz habitual de MCO respecto de los obtenidos con la matriz de White o las otras, disminuyen para algunas variables (Ver tabla 1.2). Los errores estándar para el coeficiente de la variable edad, son menores en general, que los estimados bajo la matriz de varianza covarianza habitual.

Utilizando la matriz de White y las otras metodologías, se hace inferencia sobre la significancia conjunta de las variables, sin presentar modificación sobre que todas las regresoras son importantes para explicar el gasto en tarjetas de crédito, (Ver tabla 1.3)

El modelo de heterocedasticidad multiplicativo de Harvey es un modelo que como ya se menciono es flexible y general, siendo su uso más ventajoso que los métodos anteriores. Para este caso la forma funcional de la varianza quedaría determinada de la siguiente forma:

$$\sigma_i^2 = \exp(\gamma_1 + \gamma_2 \text{Ingreso}_i + \gamma_3 \text{Ingreso}_i^2)$$

Son las variables ingresos e ingresos al cuadrado las que explicarían la varianza del término del error. Bajo el contexto de máxima verosimilitud aparecen los test de razón de verosimilitud (LR), el test de Wald y el test de multiplicadores de lagrange (LM) para hacer inferencia, todos ellos distribuidos χ^2, que para el caso tiene 2 grados de libertad

	Constante	Edad	Vivienda Propia	Ingresos	Ingresos^2
Coeficiente	-58.437	-0.37607	33.358	96.823	-3.3008
Error estándar	62.098	0.55000	37.135	31.798	2.6248
Cociente t	-0.941	-0.684	0.898	3.045	-1.448

TABLA 1.4. Modelo de Heterocedasticidad Multiplicativa

y que según los resultados de la tabla 1.4, rechazarían el supuesto de homocedasticidad, reafirmando que las variables que explican la varianza en la forma funcional propuesta son las utilizadas.

Es importante resaltar que los errores estándar de los coeficientes de la regresión, comparados con los obtenidos bajo la matriz de covarianza habitual (ver tabla 1.2 y 1.4), o la de White o cualquier otra apropiada, resultan mayores que los obtenidos bajo el contexto de maxima verosimilitud.

Luego, una relación más precisa del gasto en tarjetas de crédito y las variables edad, los ingresos promedio del individuo, su cuadrado y la variable binaria que toma el valor 0 o 1 dependiendo si es o no propietario de su hogar es:

$$E[Y] = -58.437 - 0.37607X_1 + +96.823X_2 - 3.3008X_3 + 33.358X_4$$

Las estimaciones son consistentes con los resultados anteriores en los que se sugiere que los ingresos y su cuadrado explican significativamente la variación en la varianza de las perturbaciones a través de las observaciones.

CAPÍTULO 2

Modelo de Mixtura Finita de Distribuciones Normales

2.1. Introducción

En este capítulo la atención se centra en el estudio de modelos de mixtura finita, particularmente de distribuciones normales. Estos modelo se han aplicado con éxito en varios campos; por ejemplo, se han empleado para modelar las estrategias que los niños utilizan cuando hacen una actividad académica, donde las componentes de la mixtura corresponden a las diferentes estrategias usadas por los niños o resultados de evaluaciones de calidad de los servicios de instituciones bancarias donde las componentes de la mixtura son los diferentes segmentos del mercado, casos mencionados por (Young, 2007).

El capítulo se divide en dos partes, la primera parte corresponde a mixturas finitas de distribuciones normales no condicionadas y la segunda parte a mixturas finitas de distribuciones normales condicionadas al modelado de la media.

Para el modelo no condicionado inicialmente se hace un recorrido por los trabajos de algunos autores que han tratado el tema, donde por ejemplo, las mixturas finitas surgen de un conjunto de datos con un número a priori de clases conocidas, con proporciones desconocidas. Luego se define el modelo de mixtura, se interpreta y se estima, utilizando para ello el algoritmo EM. Para finalizar esta parte, se ajusta un modelo de mixtura finita de distribución normal a un conjunto de datos correspondientes a 45 puntajes de la prueba ECAES en el 2009 de Economía (Exámenes de Calidad de Educación Superior), ahora conocido como Saber Pro.

En la segunda parte del capítulo se analizan los modelos de mixtura finita condicionados a la media conocidos como "mixturas de regresiones".

En esta parte del capítulo se describen los trabajos de Quandt (1958) y (1972), Quand y Ramsey (1978), Spath (1979), Aitkin y Wilson (1980), Desarbo y Cron (1988) y Lan, Leung y Tse (1988). Todos ellos abordan situaciones que dieron inicio al trabajo de lo que hoy se conoce como mixtura de regresiones.

En la ultima parte del capítulo se ajusta un modelo de mixtura finita de regresiones lineales normales al puntaje total de ECAES de la muestra de 45 estudiantes utilizada en la aplicación de la primera parte del capítulo. La media de la distribución es explicada por

el puntaje obtenido específicamente en la prueba de comprensión lectora. La estimación del modelo se realiza bajo el enfoque clásico vía algoritmo EM utilizando para maximizar mínimos cuadrados ponderados.

2.2. Modelo no condicionado

2.2.1. Introducción

Los modelos de mixtura finita han permitido una aproximación a los modelos estadísticos de gran variedad de fenómenos aleatorios. Estos modelos han continuado recibiendo gran atención y más en los últimos años debido a su utilidad como método flexible de modelado, tanto desde el punto de vista práctico como teórico. De hecho en las últimas décadas el alcance y el potencial de las aplicaciones se ha ampliado considerablemente. Los campos en los cuales los modelos de mixtura han sido aplicados exitosamente incluyen la astronomía, biología, genética, medicina, psiquiatría, economía, ingeniería y mercadeo, entre otros. En estas aplicaciones los modelos de mixtura finita sustentan una variedad de técnicas en áreas importantes de la estadística incluyendo el análisis cluster y de clases latentes, el análisis discriminante, el análisis de imágenes y el análisis de sobrevivencia; además de su papel más directo en el análisis de datos e inferencia, para proporcionar modelos descriptivos para las distribuciones ((McLachlan G, 2000)).

De acuerdo con (Wedel & DeSarbo, 1995), los modelos de mixturas finitas aparecen por primera vez en los trabajos de (Newcomb, 1886) y (Pearson, 1894). En estos se asume que una muestra proviene de un número específico de poblaciones cuyas proporciones son desconocidas.

Las mixturas finitas surgen de forma natural y se interpretan como densidades asociadas a una población estadística que esta constituida por m poblaciones subyacentes, que llamaremos **componentes de la mixtura** f_i, $i = 1, ..., m$, donde cada una de las poblaciones que componen la mixtura es una fracción a_i con $i = 1, ..., m$ y que denominamos **proporciones de la mixtura**. Como lo mencionan (Wedel & DeSarbo, 1995), cada una de las poblaciones subyacente tienen una forma de densidad específica y el propósito de la aproximación a la mixtura finita es descomponer la muestra en sus componentes de mixtura. Dentro de las funciones de densidad de probabilidad que se han utilizado para proponer y estudiar el modelo de mixtura finita se encuentran la normal ((Hasselblad, 1966); (Day, 1969); (Wolfe, 1970)), exponencial ((Teicher, 1961)) y bernoulli (conocidos normalmente como modelos de estructura latente, (Goodman, 1974))(Wedel & DeSarbo, 1995)).

Como se señala en (Redner & Walker, 1984), el problema que plantea la estimación de los parámetros de la función de densidad de probabilidad de la mixtura se ha abordado por diversos autores. La primera investigación publicada relacionada con la estimación de los parámetros, parece ser la de Pearson. Respecto a este trabajo, (Redner & Walker, 1984) indican que el problema que se considera, es la estimación de los parámetros de una mixtura de dos densidades normales univariadas. Para estimar los 5 parámetros independientes que están presentes en esta mixtura de dos densidades normales univariadas, el enfoque propuesto por Pearson es el método de los momentos. Este consiste en determinar las ecuaciones de los primeros 5 momentos muestrales e igualar a sus respectivos momen-

tos poblacionales; después de un desarrollo algebraico se llega a las expresiones para las estimaciones que dependen de una raíz de un polinomio de noveno grado.

Además de Pearson, (Charlie & Wicksell, 1924) y (Quandt & Ramsey, 1978) también utilizan este método. Este se continuó utilizando hasta la década de 1960, pero sólo para problemas de estimación de mixtura de densidades sencillas. En este periodo la mayoría de los esfuerzos se concentraron en las mixtura de densidades normales, especialmente en el caso propuesto por Pearson. La mayoría de estos esfuerzos pretendían simplificar el trabajo realizado por él o en ofrecer estimaciones más accesibles en casos restringidos. (Wedel & DeSarbo, 1995)

Durante la década de 1960 con la aparición de procesadores y métodos numéricos cada vez más potentes y sofisticados, los investigadores empezaron a centrar su atención en el método de máxima verosimilitud como la técnica para abordar el problema de estimación de mixtura de densidades. Una estimación de máxima verosimilitud asociada a una muestra de observaciones es una selección de los parámetros que maximiza la función de densidad de la muestra.

(Redner & Walker, 1984) ofrecen una buena e importante revisión de la literatura dirigida hacia el tema de la estimación de los parámetros por máxima verosimilitud.(Hasselblad, 1966), trata la estimación por máxima verosimilitud para las mixturas de cualquier número de densidades normales univariadas. (Day, 1969) aborda la mixtura de dos densidades normales multivariadas con una matriz común de covarianza desconocida. (Wolfe, 1970) quién muestra el caso general de una mixtura de cualquier número de densidades normales multivariante. (Hathaway, 1985), menciona que el método de máxima verosimilitud conduce a un problema de optimización mal planteado en el caso de una mixtura de distribuciones normales y lo reformula mediante restricciones simples sobre las variables para la estimación en el caso univariado. (Tan & Chang, 1972) compararon el método de momentos y el de máxima verosimilitud para una mixtura de dos densidades normales univariadas. Tambien se destaca el trabajo de (Titterington & Makov, 1985) relacionado con el método de máxima verosimilitud para la estimación de los parámetros de mixturas normales.

La estimación de máxima verosimilitud de los parámetros de mixtura se encuentra al maximizar las funciones de verosimilitud o log verosimilitud resultantes. En (Redner & Walker, 1984) se hace referencia a las dificultades computacionales asociadas con la obtención de dichas estimaciones. Estas dificultades se deben a la compleja dependencia de la función de verosimilitud con los parámetros a estimar. Para dar solución a este problema se recurre a métodos que den una solución aproximada a través de algún tipo de procedimiento iterativo.

De acuerdo a (Everitt, 1984), para la estimación por máxima verosimilitud de mixturas de dos densidades normales, se tienen una variedad de algoritmos que se pueden considerar para este propósito. Entre ellos, menciona: Algoritmo EM, Método de Newton, Algoritmo de Fletcher-Reeves y el Algoritmo Simplex de Nelder y Mead.

El método de Newton es usado por (McHugh, 1956), el algoritmo EM en cambio aparece en los trabajos de (Dempster et al., 1977), y de (DeSarbo, 1988) quienes lo utilizan para estimar su mixtura condicional. (Redner & Walker, 1984) centran su estudio en máxima verosimilitud y el algoritmo EM y (Wedel & DeSarbo, 1995) desarrollan un acercamiento al modelo de mixtura para modelos lineales generalizados donde los parámetros se estiman utilizando máxima verosimilitud y el algoritmo EM (Everitt, 1984).

El método de Newton requiere relativamente pocas iteraciones para converger y proporciona las varianzas asintóticas de las estimaciones de parámetros como un subproducto, sin embargo, la convergencia no se garantiza (Atkinson 1989; McLachlan y Basford 1988). Sobre el uso del algoritmo EM, (Titterington & Makov, 1985) y a (McLachlan & Basford, 1988), consideran que: 1) tiene estabilidad numérica, es decir, en las iteraciones hay un crecimiento constante en la función de verosimilitud de los datos observados; 2) como aspecto negativo tiene que la tasa de convergencia es muy lenta para una vecindad del punto óptimo; 3) el algoritmo puede converger a puntos de silla o máximos locales, y para lograr un máximo global es necesario iniciar con una buena estimación de los parámetros de arranque propuestos.

En (Everitt, 1984) se prueban seis algoritmos para encontrar las estimaciones de máxima verosimilitud de los cinco parámetros de una mixtura normal de dos componentes, utilizando tres conjuntos de datos para hacer la comparación y en cada caso tres conjuntos de valores iniciales. Los resultados indican que los algoritmos más satisfactorios fueron el método de Newton, y el algoritmo EM. De estos dos, el método de Newton tiene la ventaja de que la matriz Hessiana en el máximo se puede utilizar para proporcionar estimaciones de las varianzas y covarianzas de los parámetros.(Louis, 1982), considera que esta ventaja es mínima, ya que en su artículo muestra cómo la matriz de información se puede encontrar al utilizar el algoritmo EM.

2.2.2. El modelo

Dada una variable aleatoria Y, los modelos de mixtura finita descomponen una función de densidad de probabilidad $f(y)$ en la suma de m funciones de densidad de probabilidad. Si $f_j(y)$ es la j-ésima función de densidad de probabilidad que compone la mixtura finita con m componentes, entonces el modelo de mixtura finita se define como:

$$f(y) = \sum_{j=1}^{m} p_j f_j(y; \boldsymbol{\theta}_j) \tag{2.1}$$

donde p_j es la proporción de la mezcla o peso de la j-ésima componente en la mixtura con la restricción de que $0 \leq p_j \leq 1$ y $\sum_{j=1}^{m} p_j = 1$. para $j = 1, ..., m$. La proporción p_j se puede interpretar como la probabilidad a priori de observar una muestra de la componente j.

Un modelo de mixtura finita frecuentemente usado para datos univariados $y_1, ..., y_n$, como ya se mencionó, es asumir que las observaciones son realizaciones independientes e idénticamente distribuidas de una variable Y, que sigue una mixtura de m distribuciones normales univariadas. La función de densidad de esta distribución esta dada por

$$f(y) = \sum_{j=1}^{m} p_j f_j(y; \mu_j, \sigma_j^2) \tag{2.2}$$

donde $f_j(y; \mu_j, \sigma_j^2)$ es la función de densidad de una distribución normal univariada.

Contrario a la intuición, de todos los modelos de mixtura finita, los modelos de mixtura de distribuciones normales, tienen propiedades poco deseables. su función de verosimilitud no es acotada, a no ser que las varianzas de las componentes de la mixtura sean iguales

o por lo menos restringidas, la información de fisher puede ser infinita y la condición de identificabilidad estricta no se satisface.

Se puede considerar 3 clases de problemas de inferencia estadística en los modelos de mixtura finita. Primero, el modelo de mixtura requiere que se especifique m, esto es, el número de componentes de la mixtura. El problema de inferencia estadística de una mixtura con un número de componentes desconocido es un tema que requiere bastante atención, sin embargo este no será abordado aquí y se supondrá que el número de componentes es conocido. Segundo, los parámetros de cada componente $\theta_1, ..., \theta_m$ y los pesos de cada componente $p_1, ..., p_m$ posiblemente son desconocidos y se deben estimar a partir de los datos. En lo que sigue denotamos con Θ al conjunto de todos los parámetros que aparecen en el modelo de mixtura 2.1. Con el fin de obtener un vector de parámetros distintos, uno de los pesos $p_1, ..., p_m$ se debe omitir ya que cada p_j, $j = 1..m$ se puede determinar a partir de los restantes pesos. El último problema es la asignación de cada observación y_i a una determinada componente.

En esta caso se asume que en el modelo 2.1 la cantidad m de componentes y la familia paramétrica de distribuciones de donde provienen las densidades de cada componente se conoce, pero los parámetros de las densidades de cada componente, los pesos y la asignación de cada observación a qué componente pertenece se desconocen. En este caso la estimación de los parámetros, incluso en las mixturas compuestas por normales univariadas, no es sencillo y se requiere el uso de algún método numérico.

2.2.3. Estimación del modelo vía algoritmo EM

Las primeras referencias en la literatura del algoritmo EM se encuentran en (Newcomb, 1886), quien consideró la estimación de parámetro de una mixtura de dos normales univariadas. (M'Kendrick, 1925) propuso un método iterativo, que en esencia tienen el espíritu del algoritmo EM.

Como se mencionó anteriormente, en (Dempster et al., 1977) se trabajo el algoritmo EM para el cálculo de los estimadores de máxima verosimilitud, de las medias de una mixtura de proporciones conocidas, con un número finito de poblaciones normales univariadas y varianzas conocidas. Con base en esto se generó un esquema iterativo, en un contexto general por medio del algoritmo EM, cuyas propiedades de convergencia de solución de máxima verosimilitud fueron establecidas bajo una base teórica. El papel de (Dempster et al., 1977) fue el de un catalizador oportuno de las investigaciones realizadas en su época, sobre las aplicaciones para calcular los estimadores de máxima verosimilitud (EMV) en los modelos de mixturas.

Así, desarrollado y puesto en práctica el algoritmo EM varios estadísticos lo han utilizado para estimar parámetros en modelos de mixturas, ya que estos tienen la particularidad de depender de varios parámetros y por ende resulta bastante complicado calcular una estimación. Para el uso del algoritmo EM se pueden consultar por ejemplo (Ganesalingam & Mclachlan, 1978), O´Neill (1978) y (Aitkin & Tunnicliffe, 1980), entre muchos otros. Recientemente K.S. Sultan et al. (2007) estimaron los parámetros en una mixtura de dos distribuciones Weibull inversa via el algoritmo EM, realizando demostraciones numéricas y cálculos a través de simulaciones Monte Carlo.

2.2.4. Estimación de máxima verosimilitud y datos completos

El algoritmo EM fue desarrollado, asumiendo que las observaciones del modelo de mixtura finita estan incompletas, lo cual se hace pertinente debido a la falta de identificación de cada una de las observaciones respecto a la componente de la mixtura a la cual pertenece. A continuación se presenta el problema de máxima verosimilitud con datos incompletos.

Suponga que se dispone de un conjunto de datos $\boldsymbol{y} = (y_1, ..., y_n)$, independientes e idénticamente distribuidos de una variable aleatoria que sigue una distribución de mixtura finita. Sea $\boldsymbol{y}$ un conjunto de datos incompletos y suponga que existen datos no observados que son los que indican de qué componente de la mixtura fue generado cada y_i. Para lo anterior, suponga que un vector aleatorio no observable Z_i de tamaño $m \times 1$, asociado a la observación y_i y además, sea $\boldsymbol{z}_i$ la realización de Z_i. El vector $\boldsymbol{z}_i$ contiene m variables indicadoras z_{ij}, que toman el valor de uno o cero dependiendo de si y_i proviene o no de la j-ésima componente de la mixtura. Como cada y_i proviene de exactamente una componente, entonces $\sum_{j=1}^{m} z_{ij} = 1$. Adicionalmente la probabilidad de que y_i provenga de la j-ésima función de densidad de la mixtura es p_j, por tanto se puede asumir que el vector Z_i se distribuye de acuerdo a una distribución multinomial con un ensayo y m resultados con probabilidades $p_1, ..., p_m$. Así, la función de masa de probabilidad es

$$P(Z_i = \boldsymbol{z}_i) = \prod_{j=1}^{m} p_j^{z_{ij}} \tag{2.3}$$

donde z_{ij} denota el j-ésima elemento del vector z_i.

Así pues, nos referimos a las mediciones $\boldsymbol{y}$, como los datos incompletos o datos observados, $\boldsymbol{Z}$ como los datos no observados y $(\boldsymbol{y}, \boldsymbol{Z})$ como los datos completos.

Las funciones de verosimilitud y de log-verosimilitud correspondientes a los datos observados son:

$$\begin{aligned} L_o(\boldsymbol{\Theta}; \boldsymbol{Y}) &= \prod_{i=1}^{n} \sum_{j=1}^{m} p_j f_j(y_i; \boldsymbol{\theta}_j) \\ l_o(\boldsymbol{\Theta}) &= \ln \left[\prod_{i=1}^{n} \sum_{j=1}^{m} p_j f_j(y_i; \boldsymbol{\theta}_j) \right] = \sum_{i=1}^{n} \ln \left[\sum_{j=1}^{m} p_j f_j(y_i; \boldsymbol{\theta}_j) \right] \end{aligned} \tag{2.4}$$

Si consideramos la distribución asociada a los datos completos, esta puede expresarse como:

$$P(Y, \boldsymbol{Z}|\Theta) = P(Y|\boldsymbol{Z}, \Theta) P(\boldsymbol{Z}|\Theta) \tag{2.5}$$

Y la función de verosimilitud de los datos datos completos es:

$$
\begin{aligned}
L_c(\boldsymbol{\Theta}) &= \prod_{i=1}^{n} f_{Y,\boldsymbol{Z}}(y_i, z_{ij}) \\
&= \prod_{i=1}^{n} f_{Y|\boldsymbol{Z}}(y_i|z_{ij}) P_{\boldsymbol{Z}}(z_{ij}) \\
&= \prod_{i=1}^{n}\prod_{j=1}^{m} [p_j f_j(y_i)]^{z_{ij}} \qquad (2.6) \\
l_c(\boldsymbol{\Theta}) &= \ln \prod_{i=1}^{n}\prod_{j=1}^{m} [p_j f_j(y_i)]^{z_{ij}} \\
l_c(\boldsymbol{\Theta}) &= \sum_{i=1}^{n}\sum_{j=1}^{m} z_{ij} \ln[p_j f_j(y_i;\boldsymbol{\theta}_j)]
\end{aligned}
$$

$$
l_c(\boldsymbol{\Theta}) = \sum_{i=1}^{n}\sum_{j=1}^{m} z_{ij}\{\ln p_j + \ln f_j(y_i;\boldsymbol{\theta}_j)\} \qquad (2.7)
$$

Dado que es necesaria la estimación de los parámetros Θ, se debe calcular la función de verosimilitud $L(\Theta|Y, \boldsymbol{Z})$, de lo que surge la pregunta, ¿como calcular esto si no se conoce a $\boldsymbol{Z}$. Para responder al anterior interrogante se supone a $\boldsymbol{Z}$ como matriz aleatoria y se calcula el valor esperado de $L(\Theta|Y, \boldsymbol{Z})$. Como la función de log-verosimilitud de los datos completos, $l_c(\boldsymbol{\Theta})$, es lineal para los datos no observables z_{ij}, simplemente se requiere el cálculo de la esperanza condicional de $\boldsymbol{Z}_{ij}$ dados los datos observados $\boldsymbol{y}$, donde $\boldsymbol{Z}_{ij}$ es la variable aleatoria correspondiente a z_{ij}, por tanto,

$$
E_{\Theta}[\boldsymbol{Z}_{ij}|Y_i = y_i) = Pr_{\boldsymbol{\Theta}}\{\boldsymbol{Z}_{ij} = 1|Y_i = y_i\} \qquad (2.8)
$$

por el teorema de Bayes y el teorema de probabilidad total se tiene que

$$
Pr_{\boldsymbol{\Theta}}\{\boldsymbol{Z}_{ij} = 1|Y_i = y_i\} = \frac{Pr(y_i|z_{ij})Pr(z_{ij})}{P(y_i)} = \frac{p_j f_j(y_i;\boldsymbol{\theta}_j)}{\sum_{j=1}^{m} p_j f_j(y_i;\boldsymbol{\theta}_j)} \qquad (2.9)
$$

Usando 2.9 tenemos que el valor esperado condicional de (2.7) dado $\boldsymbol{y}$ es

$$
Q(\boldsymbol{\Theta};\boldsymbol{\Theta}^{(t)}) = \sum_{i=1}^{n}\sum_{j=1}^{m} E(\boldsymbol{Z}_{ij})[\ln p_j + \ln f_j(y_i;\boldsymbol{\theta}_j)] \qquad (2.10)
$$

$$
Q(\boldsymbol{\Theta};\boldsymbol{\Theta}^{(t)}) = E_{\boldsymbol{\Theta}^{(t)}}\{\ln L_c(\boldsymbol{\Theta})|\boldsymbol{Y} = \boldsymbol{y}\} \qquad (2.11)
$$

$$
Q(\boldsymbol{\Theta};\boldsymbol{\Theta}^{(t)}) = \sum_{i=1}^{n}\sum_{j=1}^{m} \frac{p_j^{(t)} f_j(y_i;\boldsymbol{\theta}_j^{(t)})}{\sum_{h=1}^{m} p_h^{(t)} f_h(y_i;\boldsymbol{\theta}_h^{(t)})}\{\ln p_j + \ln f_j(y_i;\boldsymbol{\theta}_j)\} \qquad (2.12)
$$

denotemos la ecuación 2.9 con $h_{ij}^{(t)}$, que se puede interpretar como la probabilidad a posteriori de que la i-ésima observación y_i pertenezca a la j-ésima componente de la mixtura, por tanto, la ecuación 2.12 se puede resumir como

$$Q(\boldsymbol{\Theta};\boldsymbol{\Theta}^{(t)}) = \sum_{i=1}^{n}\sum_{j=1}^{m} h_{ij}^{(t)}\{\ln p_j + \ln f_j(y_i;\boldsymbol{\theta}_j)\} \tag{2.13}$$

Lo anterior se puede definir como el paso E en el algoritmo, que no es otra cosa que calcular la esperanza de la verosimilitud con respecto a la información conocida y a unos parámetros propuestos $\boldsymbol{\Theta}^{(t)}$.

El paso M del algoritmo EM en la $t+1$-ésima iteración requiere la maximización de $Q(\Theta;\Theta^t)$ con respecto a Θ sobre el espacio paramétrico Ω para dar la actualización de la estimación Θ^{t+1}. El algoritmo EM continua hasta que la diferencia $L(\Theta^{t+1}) - L(\Theta^t)$ alcance un determinado valor de tolerancia.

Para el modelo de mixtura finita, la estimación actualizada $p_j^{(t+1)}$ de la proporción de mezcla p_j se calcula independientemente de la actualización estimada del vector que contiene los parámetros desconocidos de las densidades que componen la mixtura, y se determinan maximizando $Q(\Theta;\Theta^t)$, sujeto a la restricción $\sum_{j=1}^{m} p_j = 1$. Al maximizar $Q(\Theta;\Theta^t)$ haciendo uso de multiplicadores de lagrange obtenemos que la estimación actualizada de p_j es

$$p_j^{(t+1)} = \frac{1}{n}\sum_{i=1}^{n} \frac{p_j^{(t)} f_j(y_i;\boldsymbol{\theta}_j^{(t)})}{\sum_{h=1}^{m} p_h^{(t)} f_h(y_i;\boldsymbol{\theta}_h^{(t)})} \quad (j = 1, ..., k) \tag{2.14}$$

por tanto, en la formación de la estimación de p_j en la $t+1$-ésima iteración, hay una contribución de cada observación y_i igual a su probabilidad a posteriori de membresía de la j-ésima componente de la mixtura.

Los parámetros restantes $\boldsymbol{\theta}_1, ..., \boldsymbol{\theta}_j$ se deben determinar respecto a su densidad subyacente. Para esto se actualiza $\{\boldsymbol{\theta}_j^{(t+1)}\}_{j=1}^{m}$ resolviendo la ecuación

$$\begin{aligned} \sum_{i=1}^{n}\sum_{j=1}^{m} \frac{p_j^{(t)} f_j(y_i;\boldsymbol{\theta}_j^{(t)})}{\sum_{h=1}^{m} p_h^{(t)} f_h(y_i;\boldsymbol{\theta}_h^{(t)})} \left\{\frac{\partial \ln f_j(y_i;\boldsymbol{\theta}_j)}{\partial \boldsymbol{\theta}_j}\right\} = 0 \\ \sum_{i=1}^{n}\sum_{j=1}^{m} h_{ij}^{(t)} \left\{\frac{\partial \ln f_j(y_i;\boldsymbol{\theta}_j)}{\partial \boldsymbol{\theta}_j}\right\} = 0 \end{aligned} \tag{2.15}$$

2.2.5. EL algoritmo EM para modelos de mixtura finita con componentes normales

Suponga que se dispone de un conjunto de datos $y_1, ..., y_n$, independientes e idénticamente distribuidos de una variables aleatoria que sigue una distribución de mixtura finita de distribuciones normales

La aplicación del algoritmo EM a un modelo de mixtura de distribuciones normales es sencilla. Sea $f(y;\mu_j,\sigma_j)$ la función de densidad normal con media μ_j y desviación estándar σ_j. Como se mencionó anteriormente en la formulación general del algoritmo

EM, la actualización de las probabilidades a priori p_j son independientes de la forma de las densidades de las componentes de la mixtura. Luego la actulización de p_j esta dada por la ecuación 2.14. Para la estimación de μ_j σ_j de $f(y;\mu_j,\sigma_j)$ en la t-ésima iteración del algoritmo EM, debemos resolver la ecuación 2.15. Note que el logaritmo de $f(y;\mu_j,\sigma_j)$ esta dado por

$$\log f(y;\mu_j,\sigma_j) = -\log(\sqrt{2\pi}\sigma_j) - \frac{(y-\mu_j)}{2\sigma_j^2} \tag{2.16}$$

Primero consideremos las actualizaciones para la media μ_j de cada componente. La derivada de $\log f(y;\mu_j,\sigma_j)$ con respecto μ_j es:

$$\frac{\partial \log f(y;\mu_j,\sigma_j)}{\partial \mu_j} = -\frac{y-\mu_j}{\sigma_j^2} \tag{2.17}$$

y sustituyendo este resultado en la ecuación 2.15 se obtiene

$$\sum_{i=1}^{n}\sum_{j=1}^{m} \frac{p_j^{(t)} f_j(y_i;\theta_j^{(t)})}{\sum_{h=1}^{m} p_h^{(t)} f_h(y_i;\theta_h^{(t)})} \frac{y-\mu_j}{\sigma_j^2} = 0 \tag{2.18}$$

Simplificando lo anterior y resolviendo para $\hat{\mu}_j^{t+1}$, se obtiene la actualización

$$\hat{\mu}_j^{(t+1)} = \frac{\sum_{i=1}^{n} z_{ij}^{(t)} y_i}{\sum_{i=1}^{n} z_{ij}^{(t)}} \tag{2.19}$$

Para la actualización de la desviación estándar σ_j. La derivada de $\log f(y;\mu_j,\sigma_j)$ con respecto σ_j es:

$$\frac{\partial \log f(y;\mu_j,\sigma_j)}{\partial \sigma_j} = -\frac{1}{\sigma_j} + \frac{(y-\mu_j)^2}{\sigma_j^3} \tag{2.20}$$

Y sustituyendo esta ecuación en la ecuación 2.15 se obtiene

$$\hat{\sigma}_j^{(t+1)} = \sum_{i=1}^{n} z_{ij}^{(t)} \left(\frac{(y-\mu_j)^2}{\sigma_j^3} - \frac{1}{\sigma_j}\right) \tag{2.21}$$

Simplificando la ecuación anterior y resolviendo para σ_j^2, se obtiene la actualización

$$\hat{\sigma}_j^{(t+1)} = \sqrt{\frac{\sum_{i=1}^{n} z_{ij}^{(t)} (y_i - \hat{\mu}_j^{(t+1)})^2}{\sum_{i=1}^{n} z_{ij}^{(t)}}} \tag{2.22}$$

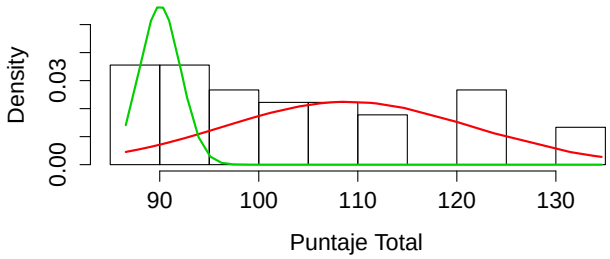


FIGURA 2.1. Histograma de 45 puntajes totales de la prueba saber pro de economía del año 2009

Modelo	Componentes	AIC	CAIC	BIC	AIC3
Gamma	1	366.35	371.47	369.47	368.35
Lognormal	1	363.25	**368.36**	366.36	365.25
Logística	1	369	374.13	372.13	371.01
Normal	1	367.03	372.14	370.14	369.03
Mixtura Normal	2	**356.22**	368.99	**363.99**	**361.22**
	3	356.5	376.95	368.95	364.51
	4	359.57	387.68	376.68	370.57
	5	362.67	398.44	384.44	376.67

TABLA 2.1. Valores de los criterios para la determinación del número de componentes de la mixtura de los puntajes de la prueba ECAES 2009 en economía

2.2.6. Aplicación

Para esta sección la aplicación consiste en realizar la estimación de un modelo de mixtura de distribuciones normales. Se considera una muestra de 45 puntajes de la prueba ECAES (ahora saber Pro) obtenidos por los estudiantes de Economía durante el 2009.

Estos datos se han seleccionado de una base de datos suministrada por el ICFES que contiene un código de registro por estudiante, universidad de donde procede, puntaje total de las pruebas que componen el examen, entre otras variables. Las áreas consideradas conjuntamente por los departamentos de economía en Colombia de conocimiento común en la formación del economista son: macroeconomía, microeconomía, econometría, historia, pensamiento económico, comprensión lectora e inglés.

Haciendo un análisis exploratorio con el puntaje total de esta prueba, aparece un posible modelo de mixtura de normales. Este conjunto de datos representado en la figura 2.1 permite observar un histograma que refleja la presencia de dos grupos centrados en valores distintos, con dos máximos diferentes, característico de una distribución bimodal. Es posible entonces, que un modelo de mixtura de dos componentes sea aplicable a este conjunto de datos.

Para evaluar el número de componentes de la mixtura se usan los criterios de información AIC, BIC, CAIC y AIC3 dados en la tabla 2.1. Se elige el número de componentes que corresponda al menor de los valores dado para cada criterio. De esta forma y de acuerdo con los criterios AIC, BIC y AIC3 el mejor ajuste corresponde a una mixtura normal de dos componentes. Bajo el CAIC el mejor modelo que se ajusta es el lognormal. Para las

Parametro	Componente 1	Componente 2
p_j	0.295	0.705
μ_j	90	109
σ_j	2.08	12.53

TABLA 2.2. Estimación de parámetros de la mixtura de dos componentes normales vía algoritmo EM para los puntajes de la prueba ECAES 2009 de economía.

componentes mayores que 3 se seleccionaron valores iniciales de los parámetros tal que ninguna de las varianzas de las componentes de la mixtura tienda a cero.

Esta situación corresponde a una mixtura de normales univariada, utilizando R y vía algoritmo EM se realiza la estimación del modelo 2.2 con $m = 2$, los parámetros de cada componente θ_j y los pesos de cada componente p, aparecen en la tabla 2.2.

2.3. Modelo condicionado a la media

2.3.1. introducción

Los modelos de mixtura finita condicionados a la media, han sido estudiados ampliamente en la literatura econométrica. A diferencia del análisis de regresión lineal convencional que supone que la función de regresión en la muestra surge de una única distribución normal, los modelos de mixtura de regresiones lineales permiten funciones de regresión heterogéneas mediante el modelado de una mixtura de distintas distribuciones normales, cada una correspondiente a una clase latente. De esta forma, el análisis de regresión de mixturas relaja el supuesto de una única población presente, permitiendo diferencias de parámetros a través de subpoblaciones no observadas. El uso de clases latentes implica que los individuos se distribuyen entorno a las diferentes funciones de regresión.(Ding, 2006)

Los modelos de mixtura de regresiones se presentan cuando parece incorrecto suponer que una sola regresión explica adecuadamente la relación con las variables disponibles, la detección de valores atípicos y la estimación de regresión robusta.

A continuación se describe brevemente el trabajo de algunos autores que dieron origen a lo que se conoce como mixtura de regresiones.

Quandt, 1958

En (Quandt, 1958) se considera un conjunto de puntos que dependen del tiempo y pretende determinar el instante de tiempo donde hay un cambio de comportamiento de los datos. Esto es, si se tienen T datos, se quiere encontrar dos conjuntos, el primero con t datos y el segundo con $T - t$ datos, donde t representa el instante de tiempo del cambio. Para esto (Quandt, 1958) propone el siguiente sistema de dos regresiones lineales que obedece a dos regímenes[1].

[1]Desde el punto de vista económico, un régimen esta caracterizado por un conjunto de reglas e instituciones que representan la economía y generan su conducta dinámica cualitativa. Por lo tanto, un cambio de régimen se asocia con un cambio en ese conjunto de reglas e instituciones. Desde el punto de vista matemático, hay un cambio de régimen cuando cambia la naturaleza de una ecuación.

$$\begin{aligned} y &= a_1 + b_1 + u_1 \\ y &= a_2 + b_2 + u_2 \end{aligned} \tag{2.23}$$

donde u_1, u_2 son los términos de error que son independientes y se distribuyen normal con media cero y desviación estándar σ_1 y σ_2.

Así pues, se muestra un método de estimación del punto de cambio para el sistema de regresiones lineales que obedece a dos regímenes. El procedimiento de estimación que utiliza es máxima verosimilitud. Sugiere una prueba asintótica y una para muestras pequeñas para probar la hipótesis de que no hay un cambio frente a la hipótesis alterna de que se dio un cambio.

Quandt, 1972

A diferencia de (Quandt, 1958), cuyos regímenes están divididos por un único punto de cambio, (Quandt, 1972) introduce el problema de los cambios discontinuos en los regímenes de regresión en puntos desconocidos de la serie de datos, suponiendo que se selecciona alguno de los regímenes con probabilidad λ y $1-\lambda$. Asumiendo que el error en los dos regímenes se distribuyen independiente y normal, la densidad condicional del i-ésimo valor de la variable dependiente, y_i, condicionada a los valores de k variables independientes, $x_{1i}, ..., x_{ki}$ es

$$\begin{aligned} h(y_i|x_{1i}, ..., x_{ki}) = &\frac{\lambda}{\sqrt{2\pi}\sigma_1} \exp\left\{ -\frac{1}{2\sigma_1^2} \left(y_i - \sum_{j=1}^{k} \beta_1 xji \right)^2 \right\} \\ &+ \frac{1-\lambda}{\sqrt{2\pi}\sigma_2} \exp\left\{ -\frac{1}{2\sigma_2^2} \left(y_i - \sum_{j=1}^{k} \beta_2 xji \right)^2 \right\} \end{aligned}$$

donde β_{1j} y β_{2j} denotan el j-esímo coeficiente de regresión de los dos regímenes y x_{ji} es la i-ésima observación en la j-ésima variable.

La función de log-verosimilitud que se obtiene del modelo plantea un problema de maximización no lineal. (Quandt, 1972) maximiza esta función mediante el uso del algoritmo del gradiente conjugado de Powell. La desventaja notable del método es que no permite identificar a qué regimen pertenece cada una de las observaciones individuales.

El trabajo de (Quandt, 1972) se considera el origen del modelo de mixtura de regresiones. Esto debido a que el modelo se puede ver como una clase de mixtura de distribuciones normales, cuya extensión consiste en proponer una modelo de regresión a la media de cada componente de la mixtura.

Quandt y Ramsey, 1978

(Quandt & Ramsey, 1978) dedujeron un procedimiento denominado método de función generatriz de momento para aplicarlo al problema de estimación de parámetros en un mixtura de dos componentes con densidad normal. Este método también fue aplicado al modelo de regresión cambiante:

$$\begin{aligned} y_i &= x_i'\beta_1 + u_{1i} \quad \text{con probabilidad} \quad \lambda \\ y_i &= x_i'\beta_2 + u_{2i} \quad \text{con probabilidad} \quad 1-\lambda \end{aligned}$$

donde $u_{1i} \sim N(0, \sigma_1^2)$, $u_{2i} \sim N(0, \sigma_2^2)$, y λ,β_1, β_2 y σ_1^2, σ_2^2 son desconocidos.

Spath, 1979

(Späth, 1979) aborda el problema combinatorio de aproximación clusterwise discreta lineal, que se define como la búsqueda de un número determinado de grupos de observaciones de tal manera que la suma total de la suma de errores al cuadrado dentro de estos grupos sea mínima.

Específicamente el problema se describe de la siguiente manera:

la fórmula usual de mínimos cuadrados de la regresión lineal se puede escribir como: m observaciones (y_i, a_{ik}), $(i = 1, ..., m, k = 1, .., l)$ con $m > l$; determinar $(x_1, ..., x_l)$ tal que

$$\text{mín} \ \sum_{i=1}^{m} (y_i - \sum_{k=1}^{l} a_{ik} x_k)^2$$

Pero si las observaciones deben ser asignadas a grupos diferentes que son desconocidos, entonces para el caso $m > l$, la formulación más adecuada del problema es encontrar un partición $C_1, ..., C_n$ de determinada longitud de n observaciones; descrito de la siguiente forma:

$$C_j \subset M = 1, ...m, |C_j| > 0, C_j \cap C_k = \emptyset$$

para $j \neq k$, $\bigcup_{j=1}^{n} C_j = M$,

y n vectores $(x_{j1}, ..., x_{jl})(j = 1, ...n)$ tal que

$$D = (C_1, ..., C_n) = \sum_{j=1}^{n} E(C_j)$$

es mínima, donde

$$E(C_j) \approx \sum_{i \in C_j} (y_1 - \sum_{k=1}^{l} a_{ik} x_{jk}^2)$$

Esta formulación del problema esta relacionado con el criterio de mínima varianza del análisis cluster. Para (Lau et al., 1999) este modelo de regresión clusterwise, puede ser expresado de esta forma:

$$\begin{aligned} Y_i &= a_1 + a_2 X_i + \varepsilon_{1i} \quad \text{si} \quad Z_{1i} = 1 \\ Y_i &= b_1 + b_2 X_i + \varepsilon_{2i} \quad \text{si;} \quad Z_{2i} = 1 \end{aligned}$$

$$\begin{aligned} \text{mín} \quad & \sum_{i=1}^{n} Z_{1i}(\varepsilon_{1i})^2 + Z_{2i}(\varepsilon_{2i})^2 \\ \text{Sujeto a} \quad & Y_i = a_1 + a_2 X_i + \varepsilon_{1i} \\ & Y_i = b_1 + b_2 X_i + \varepsilon_{2i} \\ & Z_{1i} + Z_{2i} = 1, \quad Z_{1i}, Z_{2i} \geq 0 \end{aligned}$$

A Späth (1979, 1981, 1982, 1985) se le reconoce como el primero en considerar la regresión clusterwise. (Späth, 1981) presenta una corrección al algoritmo presentado en su trabajo de 1979, y argumenta que a pesar del error en el algoritmo presentado inicialmente, este también daría resultados aceptables. (Späth, 1982), da una aplicación al algoritmo propuesto en sus trabajos anteriores, mostrando rendimiento del mismo.

Aitkin y Wilson,1980

(Aitkin & Tunnicliffe, 1980) Proponen identificar valores atípicos en una muestra, o en un modelos de regresión, mediante modelos de mixtura de normales. El método de estimación de los modelos es máxima verosimilitud vía algoritmo EM. Respecto a este algoritmo (al igual que otros autores), resaltan la facilidad para ser programado, su rápida convergencia y que además proporciona la estimación de máxima verosimilitud de los parámetros, su matriz de covarianza asintótica y el logaritmo de verosimilitud maximizada de la función, con una indicación del número de componentes de la mixtura.

(Aitkin & Tunnicliffe, 1980), se refiere a que en un modelo teórico común, cuando hay presencia de valores atípicos en una muestra, es una mixtura de dos componentes normales, en la que una de las componentes, con una mayor probabilidad apriori representa las "buenas observacionesz la otra con una menor probabilidad a priori representa las malas observaciones.

Para una muestra los modelos probabilísticos considerados son los siguientes:

- Una muestra, dos componentes y suponiendo igual varianza
- Dos componentes, igual media y suponiendo diferencias en las varianzas
- Dos componentes, diferente media y varianza.

Para el caso 1 el modelo probabilístico para los datos $y_1, ..., y_n$, observaciones idénticamente distribuidas de una población con una función de densidad

$$f(y) = pf_1(y) + (1-p)f_2(y) \tag{2.24}$$

donde

$$f_j(y) = \frac{1}{\sqrt{2\pi}\sigma} \exp \frac{-(y-\mu_j)^2}{2\sigma^2} \quad j = 1, 2$$

.

Para el caso 2, el modelo:

$$f(y) = pf_1(y) + (1-p)f_2(y) \tag{2.25}$$

donde

$$f_1(y) = \frac{1}{\sqrt{2\pi}\sigma} \exp \frac{-(y-\mu)^2}{2\sigma^2}$$

.

$$f_2(y) = \frac{1}{\sqrt{2\pi\kappa}\sigma} \exp \frac{-(y-\mu)^2}{2\kappa\sigma^2} \quad \kappa > 0$$

Y para el caso 3, el modelo:

$$f(y) = pf_1(y) + (1-p)f_2(y) \tag{2.26}$$

donde

$$f_j(y) = \frac{1}{\sqrt{2\pi}\sigma_j} \exp \frac{-(y-\mu_j)^2}{2\sigma_j^2} \quad j = 1, 2$$

Presentando además una modelo para el caso de más de dos componentes.

Después de tratado el tema de una muestra, desarrollan el análisis para los modelos de regresión, considerando un modelo de dos componentes en el que se aplica un modelo de regresión en una de las componentes:

$$f(y_i) = pf_1(y_i) + (1-p)f_2(y_i) \tag{2.27}$$

donde

$$f_1(y_i) = \frac{1}{\sqrt{2\pi}\sigma} \exp \frac{-\left(y_i - \sum_{i=0}^{s} \beta_r x_{ri}\right)^2}{2\sigma^2}$$

$$f_2(y_i) = \frac{1}{\sqrt{2\pi}\sigma} \exp \frac{-(y_i-\mu)^2}{2\sigma^2} \quad i = 1, ..., n$$

Aitkin y Wilson ilustran, además, el ajuste de los modelos presentados a un conjunto de datos particulares, encontrando que:

- Para algunas observaciones se producen grandes residuos positivos o negativos; observaciones que resultan ser valores atípicos.

- Se realizan estimaciones de los parámetros vía algoritmo EM y se muestra la probabilidad de pertenencia a las componentes del modelo.
- El número de iteraciones para que se dé la convergencia del algoritmo, está relacionada con la asignación inicial de una observación atípica en cierta componente.

Desarbo y Cron, 1988

(DeSarbo, 1988) presentan una mixtura finita condicional, empleando la metodología de máxima verosimilitud para la regresión lineal clusterwise. Este enfoque estima simultáneamente las funciones de regresión por separado y la pertenencia de las observaciones al grupo respectivo. Revisan además, algunos de los procedimientos relacionados con el tema. El algoritmo EM es usado para la estimación de los parámetros. Para examinar el rendimiento de este procedimiento se utiliza un análisis Monte Carlo a través de un diseño factorial fraccionado.

El modelo de regresión considerado en este trabajo es:

$$\begin{aligned} y_i &\sim \sum_{k=1}^{K} \lambda_k f_{ik}(y_i|X_{ij}, \sigma_k^2, b_{jk}) \\ &= \sum_{k=1}^{K} \lambda_k (2\pi\sigma_k^2)^{-1/2} \exp\left[\frac{-(y_i - \boldsymbol{x}_i \mathbf{b}_k)^2}{2\sigma_k^2}\right] \end{aligned}$$

Donde, $k = 1...k$ clusters, b_{jk}=valor del j-ésimo coeficiente de la regresión del k-ésimo cluster y σ_k^2 = varianza del k-ésimo cluster

con $\boldsymbol{x}_i = ((X_j))_i$ y $\boldsymbol{b}_k = ((b_j))_k$

A través de una aplicación al Marketing, muestran los resultados obtenidos del ajuste de un modelo de mixtura condicional a dos grupos diferentes de directores que utilizan diferentes criterios para evaluar sus gastos promocionales en presentaciones comerciales. Mencionan otras posibles aplicaciones de esta metodología, sobre todo en el campo de las ciencias sociales. Por ejemplo, en psicología para identificar a los grupos de encuestados que contestan particularmente bien o mal, ítems específicos de un test.

Lau, Leung y Tse, 1988

(Lau et al., 1999) proponen un modelo generalizado clusterwise, incluyendo a los modelos ya existentes como casos especiales, formulado como un problema de programación matemática no lineal, con restricciones lineales, para resolver simultáneamente el problema combinatorio y para estimar la pertenencia al cluster. En esta investigación se integra el análisis cluster y el análisis discriminante, desarrollando un modelo clusterwise discriminante para incorporar el parámetro de heterogeneidad en el análisis discrimiante tradicional. Demuestran que la estimación del modelo de regresión clusterwise es equivalente a la solución de un problema de programación no lineal entera mixta (modelo NMIP) adicionalmente transforman este modelo a uno de programación lineal simple con restricciones lineales.

Como uno de sus intereses en particular se centra en el uso de la programación matemática para la solución de problemas estadísticos, (Lau et al., 1999) destaca entre algunos: en regresión los trabajos de Arthanari y Dodge (1981); en análisis discriminante a

Casetti (1964) y Casetti y Semple (1968), Hand (1981) y en análisis cluster a Aronson y klein (1989), Jensen (1969), Rao (1971),por mencionar algunos.

(Lau et al., 1999) propone un modelo de programación no lineal para solucionar el problema de la regresión en el que los coeficientes de regresión del clusterwise son idénticos para los miembros dentro del cluster, pero diferentes para los de otros cluster, considerando que la mejor formulación del problema es integrar el análisis cluster en la regresión, de modo que los parámetros de cada segmento puedan ser estimados de forma simultánea. Menciona que su propuesta es más general que la de (Späth, 1979) y con un costo computacional menor que la de (DeSarbo, 1988). (Späth, 1979) menciona en su trabajo que el rendimiento de su algoritmo depende del la partición inicial y además su sensibilidad para valores atípicos.

Kiefer, 1978

(Kiefer, 1978) realiza una formulación considerada más sencilla del modelo de regresión "switching", la cual está constituida por n observaciones de alguna variable dependiente y y p variables independientes $x' = (x_1, ..., x_p)$ con la estructura

$$\begin{aligned} y_i &= x_i'\beta_1 + u_{1i} \quad \text{con probabilidad} \quad \lambda \\ y_i &= x_i'\beta_2 + u_{2i} \quad \text{con probabilidad} \quad 1-\lambda \end{aligned}$$

Los errores se asumen independientes y con distribución normal con media cero y varianza σ_1^2, σ_2^2. Bajo estos supuestos la función de densidad de probabilidad de y_i es:

$$f_i = f(y_i \mid \lambda, \beta_1, \beta_2, \sigma_1, \sigma_2) = \frac{\lambda}{\sqrt{2\pi}\sigma_1} \exp \frac{-1}{2\sigma_1^2}(y_i - x_i'\beta_1)^2 + \frac{1-\lambda}{\sqrt{2\pi}\sigma_2} \exp \frac{-1}{2\sigma_2^2}(y_i - x_i'\beta_2)^2$$

Y la función de verosimilitud de los parámetros desconocidos es:

$$L = \Pi_{i=1}^n f_i$$

Si $\hat{\beta}_1$ se selecciona tal que y_i es igual a a $x_i'\hat{\beta}_1$ para algún i entonces si $\hat{\sigma_1}$ tiende a cero, f_i crece indefinidamente. Como en el segundo término en f algunas de las otras observaciones están lejos de cero, L no es acotada.

Estas características de la función de verosimilitud ha llevado a desarrollar otras técnicas distintas a la máxima verosimilitud para problemas en los cuales restricciones a la varianza no sean impuestas a priori. Una técnica habitual es el método de momentos pero la extensión de este método a modelo de regresión no es inmediato. Un segundo método, es el método de las funciones generatrices de momentos. El método de momentos no es eficiente para este problema, mientras que el método de máxima verosimilitud descrito en (Kiefer, 1978), proporciona estimaciones de los parámetros eficientes y errores estándar asintóticos.

Junto con los avances en la verosimilitud y en los algoritmos bayesianos, además del desarrollo computacional, los estudios de regresión de mixturas se han extendido.

Otros autores que también trabajan sobre mixturas de regresiones son: (Wedel & DeSarbo, 1995), quienes revisan la literatura inicial relacionada hasta el momento con

regresión de mixturas junto con las aplicaciones que utilizan la terminología de regresión de clases latentes. DeVeaunx (1989) desarrolla un punto inicial consistente para el algoritmo EM . (Viele & Tong, 2002) y (Hurn et al., 2003) analizan la regresión de mixturas desde la perspectiva bayesiana.

2.3.2. El modelo

Cuando una variable aleatoria con distribución de mixtura finita depende de algunas covariables, se tiene un modelo de mixtura finita de regresiones.

Para el modelo de mixtura finita no condicionado 2.16 considérese los modelos de regresión lineal normal para cada μ_j, $j = 1, ..., m$ como sigue

$$\mu_j = \beta_{0j} + \beta_{1j}x_1 + ... + \beta_{kj}x_k\mu_j = \boldsymbol{\beta}'_j\boldsymbol{x} \tag{2.28}$$

donde $\boldsymbol{x}' = (x_1, ..., x_k)$ es un vector de k variables independientes explicativas para la media de la j-ésima componente de la mixtura, y $\boldsymbol{\beta}'_j = (\beta_{0j}, \beta_{1j}, ..., \beta_{kj})$ son los coeficientes del modelo de regresión de la media de la j-ésima componente de la mixtura. Tenemos que la estructura de la mixtura de la función de densidad condicional $y|\boldsymbol{x}$ es

$$f(y, \boldsymbol{x}, \psi) = \sum_{j=1}^{m} p_j(2\pi\sigma_j^2)^{-\frac{1}{2}} \exp(-\frac{1}{2\sigma_j^2}(y - \boldsymbol{\beta}'_j\boldsymbol{x})^2) \tag{2.29}$$

donde $\boldsymbol{\psi} = (\boldsymbol{\beta}_1, ..., \boldsymbol{\beta}_m, \sigma_1, ..., \sigma_m, p_1, ..., p_{m-1})$ es el vector de parámetros de la mixtura de regresiones.

Este modelo fue introducido por Goldfeld and Quandt (1976) y ha sido estudiado principalmente desde el punto de vista de la verosimilitud

2.3.3. Estimación del modelo vía algoritmo EM

La función de log-verosimilitud de los datos observados es:

$$L(\psi) = \sum_{i=1}^{n} \log \sum_{j=1}^{m} \pi_j(2\pi\sigma_j^2)^{-\frac{1}{2}} \exp(-\frac{1}{2\sigma_j^2}(y_i - \boldsymbol{x}'_i\boldsymbol{\beta}_j)^2) \tag{2.30}$$

Nuevamente agregamos a los datos observados $(y_i, \boldsymbol{x}'_i)'$, las variables indicadoras.

$$z_{i,j} = I\{\text{si la observación } i \text{ pertenece a la componente } j\}$$

$1 \leq j \leq m$. La función de log-verosimilitud considerando ahora datos completos es:

$$L(\psi) = \sum_{i=1}^{n} \sum_{j=1}^{m} z_{i,j} \log(\pi_j(2\pi\sigma_j^2)^{-\frac{1}{2}} \exp(-\frac{1}{2\sigma_j^2}(y_i - \boldsymbol{x}'_i\boldsymbol{\beta}_j))) \tag{2.31}$$

El paso E dice que para la iteración t, $t = 0, 1, ...$ se calcula el valor esperado de la función log-verosimil de los datos completos.

$$Q(\psi, \psi^{(t)}) = \sum_{i=1}^{n}\sum_{j=1}^{k} h_{i,j}^{t} \log(\pi_j (2\pi\sigma_j^2)^{-\frac{1}{2}} \exp(-\frac{1}{2\sigma_j^2}(y_i - \boldsymbol{x}_i'\boldsymbol{\beta}_j)^2)) \tag{2.32}$$

Donde

$$\begin{aligned} h_{i,j}^{(t)} &= \frac{\pi_j^{(t)} (2\pi\sigma_j^{2(t)})^{-\frac{1}{2}} \exp(-\frac{1}{2\sigma_j^{2(t)}}(y_i - \boldsymbol{x}_i'\boldsymbol{\beta}_j^{(t)})^2)}{\sum_{l=1}^{k} \pi_l^{(t)} (2\pi\sigma_l^2)^{-\frac{1}{2}} \exp(-\frac{1}{2\sigma_l^{2(t)}}(y_i - \boldsymbol{x}_i'\boldsymbol{\beta}_l^{(t)})^2)} \\ &= (1 + \sum_{l\neq j}^{k} \frac{\pi_l^{(t)}}{\pi_j^{(t)}} \frac{\sigma_j^{(t)}}{\sigma_l^{(t)}} \exp(\frac{1}{2}(\frac{1}{\sigma_j^{2(t)}}(y_i - \boldsymbol{x}'\boldsymbol{\beta}_j^{(t)})^2 - \frac{1}{\sigma_l^{2(t)}}(y_i - \boldsymbol{x}_i'\boldsymbol{\beta}_l^{(t)})^2)))^{-1} \end{aligned} \tag{2.33}$$

El paso M maximiza $Q(\psi, \psi^{(t)})$ con respecto a ψ, para obtener la actualización $\psi^{(t+1)}$

Sea $W_j^{(t)} = diag(z_{1,j}^{(t)}, ..., z_{n,j}^{(t)})$, $\psi^{(t+1)}$ se determina como

$$\pi_j^{(t+1)} = \frac{1}{n}\sum_{i=1}^{n} z_{i,j}^{(t)}$$

$$\beta_j^{(t+1)} = (\boldsymbol{x}'\boldsymbol{w}_j^{(t)})^{-1}\boldsymbol{x}'\boldsymbol{w}_j^{(t)}\mathbf{y}$$

y

$$\sigma_j^{2(t+1)} = \frac{||\boldsymbol{w}_j^{\frac{1}{2}(t)}(\boldsymbol{y} - \boldsymbol{x}'\boldsymbol{\beta}_j^{(t+1)})||^2}{tr(\boldsymbol{w}_j^{(t)})}$$

Donde $||A||^2 = A^T A$ y $tr(A)$ es la traza de la matriz A.

Es bien conocido que la función de log-verosimilitud no es acotada y tiende a infinito si una observación se encuentra exactamente sobre la linea de una de las componentes y la correspondiente varianza de la componente tiende a cero. Se ha hecho considerable investigación que tratan el tema de la verosimilitud no acotada.

2.3.4. Aplicación

En la sección 2.2.5 se estimó un modelo de mixturas de distribuciones normales de dos componentes a 45 de los puntajes totales del ECAES (Ahora Saber Pro) 2009 en Economía. En la búsqueda de encontrar una variable que explique la media de las distribuciones se observó a partir del gráfico de dispersión del puntaje total de ECAES con el puntaje en la prueba de comprensión lectora, que este impacta positivamente al puntaje total, (Ver Figura 2.2.). El puntaje de la prueba en comprensión lectora puede explicar el puntaje obtenido en el ECAES en la medida en que esta es un área básica y común en la formación de cualquier profesional, aportando con su transversalidad al puntaje total del ECAES, ya que genera habilidades que permiten desarrollar estrategias para dar solución a situaciones en contexto en otras áreas de conocimiento específicas de la carrera

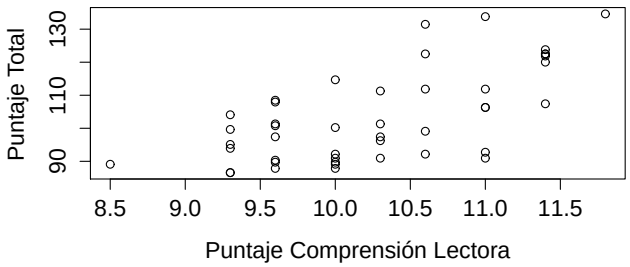


FIGURA 2.2. Gráfico de dispersión de una muestra de los puntajes de comprensión lectora y puntaje total en el ECAES 2009 de economía

de Economía como Macroeconomía, Microeconomía, Econometría, entre otras. Además, el puntaje de la prueba de comprensión lectora de forma independiente aporta directamente al puntaje total del ECAES, al ser esta un área que compone el examen.

Se propone entonces, un modelo de mixtura de regresiones descrito en la sección 2.3.2 para el puntaje total de ECAES, cuya variable explicativa de la media sea el puntaje en comprensión lectora. Los datos se analizan desde una perspectiva clásica y su estimación se realiza de acuerdo a los presentado en la sección 2.3.3.

Al obtener los puntajes de los criterios de información AIC, BIC, CAIC y AIC3 para diferentes cantidades de componentes resulta que los valores más bajos, corresponden a una mixtura de dos componentes en cualquiera de los criterios, (ver tabla 2.3). Dado este resultado, se estima un modelo de dos componentes de mixtura de distribuciones normales y los valores de la estimación para los parámetros de cada componente se presentan en la tabla 2.4.

Criterio	Valor de Criterio	Ganador
AIC	330.014	2
BIC	340.9	2
CAIC	347.9	2
AIC3	337.014	2

TABLA 2.3. Criterios de Información para evaluar la cantidad de componentes del modelo de mixtura de regresiones de los puntajes totales de la prueba ECAES 2009 de economía

La figura 2.3 muestra las rectas de regresión estimadas para las dos componentes de la mixtura. En esta se observa que la componente con menor proporción de la muestra, correspondiente al 25.2 % de los estudiantes, agrupa puntajes bajos en el ECAES y tiene una varianza homogénea a lo largo de la recta de regresión. De las dos componentes, esta es la que presenta menor varianza.

La pendiente de la recta de regresión estimada para esta componente refleja la poca influencia de la prueba de comprensión lectora sobre el puntaje total del ECAES. Es importante resaltar que con solo el efecto directo del puntaje de la prueba de comprensión lectora no se logra buen puntaje en el examen. Para esta componente se observa que puntajes altos en comprensión lectora no implican puntajes altos en el ECAES y esto

es efecto de la no transversalidad de esta área, (ver figura 2.3). La comprensión lectora posibilita el desarrollo de habilidades que ayudan a tener mejores desempeños en las demás áreas evaluadas en el examen y estas no se observan en este grupo.

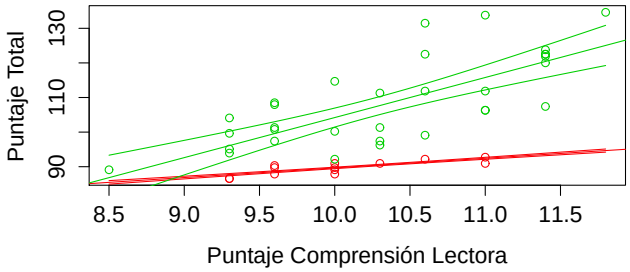


FIGURA 2.3. Rectas de regresión ajustadas al modelo de regresión de mixturas del puntaje total obtenido por una muestra de 45 estudiantes en la prueba ECAES de Economía de 2009

Parámetro	Estimación EM
p_1	0.2522
β_1	61.81 2.78
σ_1	1.1621
p_2	0.7478
β_2	-11.6 11.58
σ_2	8.69

TABLA 2.4. Estimación de parámetros de la mixtura de regresiones de dos componentes normales vía algoritmo EM de los puntajes totales de la prueba ECAES 2009 de economía

Para la otra componente, que corresponde al 74.7 % de los estudiantes de la muestra, la figura 2.3 refleja que a puntajes más altos en comprensión lectora, puntajes más altos en el ECAES con mayor varianza. Su recta de regresión estimada tiene una pendiente positiva.

En esta componente se resalta, además del aporte directo del puntaje de comprensión lectora al ECAES, la influencia de esta sobre las demás áreas evaluadas y que en conjunto con los desempeños en estas, tienen un mayor o menor efecto sobre el puntaje total del examen.

CAPÍTULO 3

Mixtura Finita de distribuciones normales con modelado de media y varianza

3.1. Introducción

Como se mostró en el capítulo 2, cuando una variable aleatoria con un modelo de mixtura finita depende de algunas variables explicativas se obtiene un modelo de mixtura de regresiones, que se ha denominado modelo de mixtura finita condicionado al modelado de la media. Si se asume la presencia de heterocedasticidad en cada regresión del modelo de mixtura, el objetivo es modelar este cambio de variabilidad.

En este capítulo se desarrolla el tema de mixtura finita de distribuciones normales con modelado conjunto de media y varianza. Para ello se presenta el modelo de mixtura finita para m componentes de distribuciones normales con una regresión lineal para la media y una estructura de heterocedasticidad multiplicativa para la varianza.

El enfoque con el que se analiza la estimación de este modelo sigue siendo el clásico, proponiendo un algoritmo de estimación de máxima verosimilitud vía algoritmo EM con la implementación del algoritmo de Fisher Scoring. El uso de esta metodología como se observará, permite hacer la estimación conjunta de los parámetros tanto para la media como para la varianza de cada componente de la mixtura del modelo.

Como propuestas anteriores se mencionan los trabajos de (Garrido et al., 2011) y (Garrido & Cepeda, 2012), quienes muestran resultados obtenidos en la aplicación y simulación de este tipo de modelos de mixtura.

Para finalizar, se aplica un modelo de mixtura finita de distribuciones normales condicionado al modelado de la media y la varianza a 45 de los puntajes totales de la prueba ECAES 2009 de Economía; esta variable respuesta ya ha sido utilizada en las aplicaciones del capítulo 2. La variable explicativa relacionada con la media y la varianza de las distribuciones de cada componente de la mixtura del modelo es el puntaje obtenido en la prueba de comprensión lectora. Adicionalmete se simulo el algoritmo con el objetivo de estudiar su desempeño.

3.2. Propuestas Anteriores

En el trabajo de (Garrido et al., 2011), se presenta un análisis desde la perspectiva Bayesiana de una mixtura de regresiones de distribuciones de la familia exponencial con modelado de varianza. Se considera como un caso especial la mixtura de distribuciones exponencial y normal, en cuyo modelo se tiene dos variables explicativas, una de las cuales está relacionada con la media de ambas distribuciones y la otra con la varianza de la distribución normal. Se utiliza MCMC para dar algunos resultados. Los resultados de simulación en este estudio mostraron buenas estimaciones bayesianas para los parámetros de los modelos propuestos y resaltan particularmente, los obtenidos para el modelo de la varianza de distribución normal.

En este trabajo se suponen por ejemplo un caso en el que los pesos están dados por enlaces logit y una de las variables explicativa esta también relacionada con estos. Una de las aplicaciones presentadas corresponde a un experimento dosis respuesta con un grupo de insectos, que se ajusta a un modelo de regresión de mixtura normal exponencial con modelado de varianza. Este ejemplo está basado en los datos utilizados en Achcar y Pereira (1999). Otro ejemplo está relacionado con la contaminación de pescado por DDT, basado en datos de Mendenhall y Terry (1997).

En (Garrido & Cepeda, 2012), se propone una mixtura de distribuciones biparamétricas de la familia exponencial, con el modelamiento conjunto de la media y la varianza; se asume un mixtura de normal gamma. Se utiliza la metodología bayesiana para ajustar el modelo. Presentan además tres simulaciones, todas incluyen modelado de media y varianza. Como aplicación, se estima un modelo de mixtura de regresiones de distribuciones normal y gamma a un conjunto de registros de hogares de un país de América Central, donde la media de la distribución gamma es no lineal. El método en general tuvo buen comportamiento para el caso práctico. También se concluye sobre el buen comportamiento del algoritmos para el caso de dos componentes de mixtura y el caso de mas de tres componentes bien distantes y tamaño de muestra grande. Para este último caso, con un tamaño de muestra pequeño no todos los parámetros se estiman con precisión.

3.3. El modelo

Los modelos de mixtura finita de distribuciones normales con modelado de media, como el presentado en la sección 2.29, combinan las características y el uso de los modelos de regresión y de mixtura finita, para construir modelos que describan fenómenos en los que se presenten grupos, no identificados a priori, a los cuales es posible explicarles parte de su variabilidad por medio de un modelo para la media con un determinado conjunto de variables.

Los modelos de mixtura de regresiones lineales normales presentados en la sección 2.3.1 y en la sección 2.3.2 comparten que los modelos de regresión presentes en la mixtura asumen homocedasticidad. Sin embargo, es posible contemplar situaciones en las cuales cada componente de regresión de la mixtura presente adicionalmente heteroscedasticidad. Esta nueva consideración plantea un problema de estimación de un modelo que incluya el modelado de las varianzas de cada regresión de la mixtura. Denominese a este modelo **Mixtura finita con modelado de media y varianza**.

Sea Y una variable respuesta de interés y $\boldsymbol{x} = (x_1, ..., x_k)$ un vector de covariables las cuales se cree tienen efecto en Y y se consideran como fijas. Se dice que Y sigue un modelo de mixtura finita de distribuciones normales con modelado de media y varianza si sigue una distribución condicionada como sigue:

$$\begin{aligned} f(y|\mathbf{x}, \Psi) &= \sum_{j=1}^{m} \pi_j \frac{1}{\sqrt{2\pi\sigma_j}} \exp\left[-\frac{1}{2\sigma_j^2}(y - \boldsymbol{\beta}_j'\boldsymbol{x})^2\right] \\ \sigma_j^2 &= \exp(\boldsymbol{\gamma}_j'\mathbf{w}_i) \end{aligned} \tag{3.1}$$

Donde $\Psi = (\boldsymbol{\beta}_1, ..., \boldsymbol{\beta}_m, \boldsymbol{\gamma}_1, ..., \boldsymbol{\gamma}_m, \boldsymbol{\pi})$, $\boldsymbol{\beta}_j = (\beta_{0j}, ..., \beta_{kj})$, $\boldsymbol{\gamma}_j = (\gamma_{0j}, ..., \gamma_{pj})$ $\boldsymbol{\pi} = (\pi_1, ..., \pi_{m-1})$ tal que $\pi_j > 0$ y $\sum_{j=1}^{m} = 1\pi_j = 1$. Las componentes de la mixtura son funciones de densidad normal univariadas con $\mu_j = \boldsymbol{\beta}_j'\boldsymbol{x}$ y varianza $\sigma_j^2 = \exp(\boldsymbol{\gamma}_j'\boldsymbol{w}_i)$ donde $\boldsymbol{w} = (1, w_1, ..., w_p)$ es un vector de variables que explica la heteroscedasticidad del modelo y $\boldsymbol{\gamma}_j = (\gamma_0 j, ..., \gamma_p j)'$ es un vector de coeficientes desconocidos del modelo de heteroscedasticidad de la componente j-ésima.

3.4. Estimación del modelo vía algoritmos EM y Fisher Scoring

Sea $(y_1, \boldsymbol{x}_1, \boldsymbol{w}_1), ..., (y_n, \boldsymbol{x}_n, \boldsymbol{w}_n)$ una muestra de observaciones del modelo (3.1). Nuevamente considérese a $\mathbf{Z}$ como matriz indicadora de tamaño $n \times m$ donde la ij-ésima componente z_{ij} es igual a 1 cuando y_i proviene de la j-ésima componente de la mixtura y cero en otro caso, las cuales son variables no observadas. Luego para el conjunto de datos dados $(y_i; \boldsymbol{x}_i; \boldsymbol{w}_i) : i = 1, ..., n$ la función de log-verosimilitud de los datos completos del modelo (3.1) es:

$$L(\Psi) = \sum_{i=1}^{n}\sum_{j=1}^{m} z_{ij} \log(\pi_j (2\pi(\sigma_j^2)_i)^{-\frac{1}{2}} \exp\left[-\frac{(y_i - \boldsymbol{\beta}_j'\boldsymbol{x}_i)^2}{2(\sigma_j^2)_i}\right] \tag{3.2}$$

Para estimar los parámetros de una mixtura de m regresiones normales heteroscedasticas, nuevamente se considera el algoritmo EM por ser un contexto de mixtura y porque proporciona un enfoque conveniente al problema de optimización de datos incompletos que plantea la estimación vía máxima verosimilitud.

Sea $\Psi^{(t)}$ la estimación de los parámetros de la t-ésima iteración. En el **paso E** se calcula la esperanza condicional de la función (3.2) con respecto a z_{ij} dados los datos $(y_i, \boldsymbol{x}_i)$ y asume que la estimación actual $\Psi^{(t)}$ son los verdaderos parámetros del modelo.

Se obtiene que la esperanza condicional de (3.2) es:

$$Q(\Psi, \Psi^{(t)}) = \sum_{i=1}^{n}\sum_{j=1}^{m} h_{ij}\log(\pi_j) + \sum_{i=1}^{n}\sum_{j=1}^{m} h_{ij}\log(f_j(y_i|x_i, \boldsymbol{\beta}_j, \boldsymbol{\gamma}_j)) \tag{3.3}$$

$$\begin{aligned} &= \sum_{i=1}^{n}\sum_{j=1}^{m}\log(\pi_j)h_{ij} + \sum_{i=1}^{n}\sum_{j=1}^{m} h_{ij}\log\left((2\pi(\sigma_j^2)_i\right)^{-\frac{1}{2}}\exp\left[\frac{-(y_i-\boldsymbol{\beta}_j'\boldsymbol{x}_i)^2}{2(\sigma_j^2)_i}\right] \\ &= \sum_{i=1}^{n}\sum_{j=1}^{m}\log(\pi_j)h_{ij} + \sum_{i=1}^{n}\sum_{j=1}^{m}\left(h_{ij}\log\left((2\pi(\sigma_j^2)_i)^{-\frac{1}{2}}\right) - h_{ij}\frac{(y_i-\boldsymbol{\beta}_j'\boldsymbol{x}_i)^2}{2(\sigma_j^2)_i}\right) \end{aligned} \tag{3.4}$$

donde

$$h_{ij} = \frac{\pi_j^{(t)}\left(2\pi(\sigma_j^2)_i^{(t)}\right)^{-\frac{1}{2}}\exp\left[-\frac{(y_i-\boldsymbol{\beta}_j'^{(t)}\boldsymbol{x}_i)^2}{2(\sigma_j^2)_i^{(t)}}\right]}{\sum_{l=1}^{m}\pi_l^{(t)}\left(2\pi(\sigma_l^2)_i^{(t)}\right)^{-\frac{1}{2}}\exp\left[\frac{-(y_i-\boldsymbol{\beta}_l'^{(t)}\boldsymbol{x}_i)^2}{2(\sigma_l^2)_i^{(t)}}\right]}$$

y

$$(\sigma_j^2)_i = \exp(\boldsymbol{\gamma}_j'\boldsymbol{w}_i)$$

En el **paso M**, en la iteración $t+1$, se maximiza la función $Q(\Psi, \Psi^{(t)})$ respecto a Ψ. La estimación actualizada de π_j^{t+1} se calcula de forma independiente a los parámetros de las componentes de la mixtura, por tanto de la ecuación (2.14), se tiene que

$$\begin{aligned} \pi_j^{(t+1)} &= \frac{1}{n}\sum_{i=1}^{n} h_{ij}^{(t)} \\ &= \frac{1}{n}\sum_{i=1}^{n}\frac{\pi_j^{(t)}\left(2\pi(\sigma_j^2)_i^{(t)}\right)^{-\frac{1}{2}}\exp\left[-\frac{(y_i-\boldsymbol{\beta}_j'^{(t)}\boldsymbol{x}_i)^2}{2(\sigma_j^2)_i^{(t)}}\right]}{\sum_{l=1}^{m}\pi_l^{(t)}\left(2\pi(\sigma_l^2)_i^{(t)}\right)^{-\frac{1}{2}}\exp\left[\frac{-(y_i-\boldsymbol{\beta}_l'^{(t)}\boldsymbol{x}_i)^2}{2(\sigma_l^2)_i^{(t)}}\right]} \end{aligned} \tag{3.5}$$

Lo anterior muestra que se puede considerar a los π_j, $j = 1, ..., m$ como constantes en $Q(\Psi, \Psi^{(t)})$, y por tanto,

$$\begin{aligned} Q(\Psi, \Psi^{(t)}) = &\underbrace{\sum_{i=1}^{n}\sum_{j=1}^{m}\log(\pi_j)h_{ij} - \frac{1}{2}\sum\sum(h_{ij}\log(2\pi))}_{C} \\ &- \frac{1}{2}\sum_{i=1}^{n}\sum_{j=1}^{m} h_{ij}\log((\sigma_j^2)_i)) - \frac{1}{2}\sum_{i=1}^{n}\sum_{j=1}^{m}\frac{(y_i-\boldsymbol{\beta}_j'\boldsymbol{x})^2}{(\sigma_j^2)}h_{ij} \end{aligned} \tag{3.6}$$

$$Q(\Psi, \Psi^{(t)}) = C - \frac{1}{2}\sum_{i=1}^{n}\sum_{j=1}^{m} h_{ij}\log((\sigma_j^2)_i)) - \frac{1}{2}\sum_{i=1}^{n}\sum_{j=1}^{m}\frac{(y_i - \boldsymbol{\beta}_j'\boldsymbol{x})^2}{(\sigma_j^2)}h_{ij} \tag{3.7}$$

Luego es posible ahora maximizar $Q(\Psi, \Psi^{(t)})$ respecto a los demás parámetros en Ψ.

Sea $\boldsymbol{\delta}_j = [\boldsymbol{\beta}_j, \boldsymbol{\gamma}_j]$, al plantear el sistema de ecuaciones (2.15) se nota que resultan m ecuaciones independientes de la forma

$$\frac{\partial Q(\Psi, \Psi^{(t)})}{\partial \boldsymbol{\delta}_j} = \frac{\partial}{\partial \boldsymbol{\delta}_j}\left[-\frac{1}{2}\sum_{i=1}^{n} h_{ij}\log((\sigma_j^2)_i) - \frac{1}{2}\sum_{i=1}^{n}\frac{(y_i - \boldsymbol{\beta}_j'\boldsymbol{x})^2}{(\sigma_j^2)}h_{ij}\right] = \mathbf{0} \tag{3.8}$$

Para resolver las anteriores m ecuaciones se hace uso del algoritmo Fisher Scoring como en la sección 1.6.1.

A continuación se presenta cada uno de los elementos de la ecuación 1.30 y las ecuaciones de iteraciones resultantes para estimar $\boldsymbol{\delta}_j$, $j = 1..., m$.

El calculo del vector gradiente $\nabla Q(\Psi, \Psi^{(t)})$ es:

$$\left[\frac{\partial Q(\Psi, \Psi^{(t)})}{\partial \boldsymbol{\beta}_j}, \frac{\partial Q(\Psi, \Psi^{(t)})}{\partial \boldsymbol{\gamma}_j}\right] = \left[\sum_{i=1}^{n} h_{ij}\boldsymbol{x}_i\frac{(y_i - \boldsymbol{\beta}_j'\boldsymbol{x}_i)^2}{\exp(\boldsymbol{\gamma}_j'\boldsymbol{w}_i)}, \frac{1}{2}\sum_{i=1}^{n} h_{ij}\boldsymbol{w}_i\left(\frac{(y_i - \boldsymbol{\beta}_j'\boldsymbol{x}_i)^2}{\exp(\boldsymbol{\gamma}_j'\boldsymbol{w}_i)} - 1\right)\right] \tag{3.9}$$

La matriz de información observada es

$$\begin{aligned}\mathbf{H}_{Q(\Psi,\Psi^{(t)})} &= \begin{bmatrix} \frac{\partial^2 Q(\Psi,\Psi^{(t)})}{\partial\boldsymbol{\beta}_j\partial\boldsymbol{\beta}_j'} & \frac{\partial^2 Q(\Psi,\Psi^{(t)})}{\partial\boldsymbol{\beta}_j\partial\boldsymbol{\gamma}_j'} \\ \frac{\partial^2 Q(\Psi,\Psi^{(t)})}{\partial\boldsymbol{\gamma}_j\partial\boldsymbol{\beta}_j'} & \frac{\partial^2 Q(\Psi,\Psi^{(t)})}{\partial\boldsymbol{\gamma}_j\partial\boldsymbol{\gamma}_j'} \end{bmatrix} \\ &= \begin{bmatrix} \sum_{i=1}^{n} h_{ij}\frac{\boldsymbol{x}_i\boldsymbol{x}_i'}{\exp(\boldsymbol{\gamma}_j'\boldsymbol{w}_i)} & \sum_{i=1}^{n} h_{ij}\frac{(y_i - \boldsymbol{\beta}_j'\boldsymbol{x}_i)}{\exp(\boldsymbol{\gamma}_j'\boldsymbol{w}_i)}\boldsymbol{x}_i\boldsymbol{w}_i' \\ \sum_{i=1}^{n} h_{ij}\frac{(y_i - \boldsymbol{\beta}_j'\boldsymbol{x}_i)^2}{\exp(\boldsymbol{\gamma}_j'\boldsymbol{w}_i)}\boldsymbol{x}_i\boldsymbol{w}_i' & -\frac{1}{2}\sum_{i=1}^{n} h_{ij}\frac{(y_i - \boldsymbol{\beta}_j'\boldsymbol{x}_i)^2}{\exp(\boldsymbol{\gamma}_j'\boldsymbol{w}_i)}\boldsymbol{w}_i\boldsymbol{w}_i' \end{bmatrix}\end{aligned} \tag{3.10}$$

Como el valor esperado de $\frac{\partial^2 Q(\psi,\psi(t))}{\partial\boldsymbol{\beta}_j\partial\boldsymbol{\gamma}_j'}$ es 0 dado que $E((\varepsilon_j)_i) = 0$ y El valor esperado de la fracción de $\frac{\partial^2 Q(\Psi,\Psi^{(t)})}{\partial\boldsymbol{\gamma}_j\partial\boldsymbol{\gamma}_j'}$ es $E\left(\frac{(\varepsilon_j)_i^2}{(\sigma_j)_i^2}\right) = 1$ la matriz de información de Fisher esperada que resulta es:

$$-E\left[\frac{\partial^2 Q(\Psi, \Psi^{(t)})}{\partial\boldsymbol{\delta}_j\partial\boldsymbol{\delta}_j'}\right] = \begin{bmatrix} \sum_{i=1}^{n} h_{ij}\frac{\boldsymbol{x}_i\boldsymbol{x}_i'}{\exp(\boldsymbol{\gamma}_j'\boldsymbol{w}_i)} & \mathbf{0} \\ \mathbf{0} & 2\sum_{i=1}^{n} h_{ij}\boldsymbol{w}_i\boldsymbol{w}_i' \end{bmatrix} \tag{3.11}$$

Dado que la matriz de información esperada es diagonal por bloques, las iteraciones pueden escribirse como ecuaciones separadas:

$$\boldsymbol{\beta}_j^{(t+1)} = \boldsymbol{\beta}_j^{(t)} + \left(\sum_{i=1}^{n} h_{ij} e^{-\boldsymbol{w}_i' \boldsymbol{\gamma}_j^{(t)}} \boldsymbol{x}_i \boldsymbol{x}_i' \right)^{-1} \sum_{i=1}^{n} h_{ij} \boldsymbol{x}_i e^{-\boldsymbol{w}_i' \boldsymbol{\gamma}_j^{(t)}} (y_i - \boldsymbol{x}_i' \boldsymbol{\beta}_j^{(t)}) \tag{3.12}$$

y

$$\boldsymbol{\gamma}_j^{(t+1)} = \boldsymbol{\gamma}^{(t)} + \left(\sum_{i=1}^{n} \boldsymbol{w}_i h_{ij} \boldsymbol{w}_i' \right)^{-1} \sum_{i=1}^{n} h_{ij} \boldsymbol{w}_i [e^{-\boldsymbol{w}_i' \boldsymbol{\gamma}_j^{(t)}} (y_i - \boldsymbol{w}_i' \boldsymbol{\beta}_j^{(t)})^2] - 1] \tag{3.13}$$

donde $\boldsymbol{\gamma}_j^{(t)}$ y $\boldsymbol{\beta}_j^{(t)}$ son las estimaciones de $\boldsymbol{\gamma}_j$ y $\boldsymbol{\beta}_j$, respectivamente, que se obtienen en la t-ésima iteración para la j-ésima componente de la mixtura (3.1).

Iniciando con valores iniciales $\Psi^{(0)}$ se itera entre el paso E y el Paso M hasta que algún criterio de convergencia se satisfaga.

3.5. Valores Iniciales

La elección de los valores iniciales para el algoritmo EM en un contexto de mixturas finitas, es un tema que ha preocupado a algunos autores, entre ellos (**?**). En ese trabajo se mencionan algunas de las propuestas para elegir los valores iniciales y se comparan varias de las estrategias utilizando simulación, afirmando que esta elección puede influir en la velocidad de convergencia del algoritmo y su capacidad para localizar el máximo global. Estas formas de elegir los valores iniciales se presentan a continuación.

Laird (1978), propone una búsqueda en grilla para establecer los valores iniciales. Leroux (1992), sugirió uso de información complementaria con el fin de formar grupos cuyas medias fueron utilizadas como valores iniciales. McLachlan (1998), propuso el uso del análisis de componentes principales para la selección de valores iniciales en el caso de mixtura multivariada. Finch (1989), sugiere que dada la proporción de la mixtura p, la muestra se separa en dos partes, una que contiene la primera $[np]$ observaciones asumiendo que pertenecen a la primera componente de la mixtura y la otra, que contiene las observaciones restantes, suponiendo que pertenecen a la segunda componente. La media de las observaciones de la primera parte de la muestra se usa como un valor inicial para la media de la primera componente de la mixtura, mientras que la de la segunda parte se usa como un valor inicial para la media de la segunda componente. En el método de Finch con proporciones de mixturas iguales, un valor inicial para la varianza está dado por $\sigma_0^2 = s^2 - \widehat{\sigma^2}(\mu)$. Böhning (1994), propone comenzar con valores iniciales bien separados, que en su experiencia, ayuda a que el algoritmo converja más rápido. De este mismo autor, pero en su trabajo de 1999, propone una partición inicial de los datos mediante la maximización con el criterio de suma de cuadrados. Otra opción es comenzar con las estimaciones obtenidas por el método de momentos, esta estrategia es usada para mixturas normales por Furman y Lindsay (1994a, b) y Lindsay y Basak (1993).

Otras estrategias que también se proponen en (**?**) son: Puntos de partida al azar, que consiste en generar los valores iniciales de las medias a partir de distribuciones uniformes sobre el rango de datos, las proporciones de mezcla se generan a partir de una distribución de Dirichlet y la varianza se genera a partir de una distribución uniforme que va desde 0 al valor de la varianza de la muestra; El mejor de los diez puntos de partida diferentes al azar, que consiste en que para cada conjunto de valores iniciales se calcule la logverosimilitud y

el conjunto con la mayor probabilidad se considera como el mejor y se utiliza como punto de partida. Otro método consiste en obtener los valores por el método de comparación de momentos, que utiliza proporciones iguales de mixtura y medias de las componentes dadas por $\overline{x} \pm j/2s$ para $j = 1, 2, ..., k/2$ incluso para k y $j = 0, ..., (k-1)/2$ para k impar. La varianza se determina como en el método de Finch. Otra estrategia expuesta por (**?**), asigna los valores iniciales al vector latente ω_i, donde $\omega_{ij} = 1$ si la $i-$ésima observación pertenece la $j-$ésima componente y 0 en otro caso. En este método inicialmente los percentiles que dividen al conjunto de datos en partes iguales son considerados como valores iniciales para las medias y cada observación se asigna a la componente con la media inicial que está más cerca de la observación en distancia absoluta.

En (**?**), los resultados de simulación mostraron que el algoritmo EM puede ejecutar una gran cantidad de iteraciones atrapado en áreas lejos del máximo local y que es dependiente de la elección de los valores iniciales. Se recomienda partir de varios valores iniciales, hacer un pequeño número de iteraciones sin examinar convergencia y luego ejecutar hasta la convergencia desde el punto con la mayor probabilidad después de estas iteraciones iniciales, utilizando un criterio estricto.

En este trabajo se propone una metodología para una mixtura con modelado de media y varianza de dos componentes como se presenta a continuación.

1. Dado el conjunto de observaciones $(y_i, \boldsymbol{x}_i, \boldsymbol{w}_i)$, $i = 1, ..., n$ que provienen del modelo 3.1. se divide inicialmente las observaciones en dos grupos mediante un análisis cluster con la técnica k-means para dos grupos.

2. Una vez clasificada cada observación i, $i = 1, ..., n$ en cada grupo se calculan los estimadores de mínimos cuadrados ordinarios con las observaciones $(y_i, \boldsymbol{x}_i)$ que quedaron en el grupo 1 y con las que quedaron en el grupo 2 respectivamente, obteniendo dos modelos de regresión lineal estimados.

3. Luego se calculan los residuos cuadráticos de cada observación en cada modelo de regresión estimado. Si la observación que esta clasificada en el grupo 1 tiene un error cuadrático menor en la regresión 1 que en la regresión 2 esta observación queda en el grupo 1, de lo contrario se cambia al grupo 2. Si la observación que esta clasificada en el grupo 2 tiene un error cuadrático menor en la regresión 2 que en la regresión 1 esta observación queda en el grupo 2, de lo contrario se cambia al grupo 1. Con esto quedan reclasificadas las observaciones $(y_i, \boldsymbol{x}_i, \boldsymbol{w}_i)$, $i = 1, ..., n$ en los dos grupos, lo que permite obtener valores iniciales para el peso de cada componente.

4. Con la actual clasificación de las observaciones $(y_i, \boldsymbol{x}_i, \boldsymbol{w}_i)$, $i = 1, ..., n$ nuevamente calculamos los estimadores de mínimos cuadrados ordinarios para cada grupo de observaciones. De esta manera obtenemos valores iniciales para los coeficientes de regresión del modelado de la media de cada componente de la mixtura 3.1 de dos componentes.

5. Finalmente se estima un modelo de regresión auxiliar con el logaritmo natural de los cuadrados de los residuos de cada modelo de regresión que se obtuvo para la media de cada componente y las variables explicativas de la varianza. De esta forma se obtienen valores iniciales para los coeficientes del modelo para la varianza de cada componente de la mixtura 3.1 de dos componentes.

	M1	M2	M3	M4	M5
n	100	300	400	200	100
$\boldsymbol{\beta}_1$	$(3, 2)$	$(3, 2)$	$(40, 0)$	$(18, 3)$	$(18, 3)$
$\boldsymbol{\beta}_2$	$(-2, -3)$	$(-2, -3)$	$(-3, 8)$	$(1, 3)$	$(1, 3)$
$\boldsymbol{\gamma}_0$	$(-6, 0.52, 0.25)$	$(-6, 0.52, 0.25)$	$(3.4, -0.2)$	$(0.69, 0.83, -0.08)$	$(0.69, 0.83, -0.08)$
$\boldsymbol{\gamma}_1$	$(-5, 0.45, 0.15)$	$(-5, 0.45, 0.15)$	$(1.1, 0.3)$	$(1, 0, 0)$	$(1, 0, 0)$
$\boldsymbol{\pi}$	(0.4,0.6)	(0.4,0.6)	(0.1,0.9)	(0.7,0.3)	(0.7,0.3)
Replicas	1200	1200	1200	1200	1200

TABLA 3.1. Modelos de Mixtura para la Simulación

3.6. Simulación

Se desarrolló un estudio de simulación para el algoritmo propuesto en la sección 3.4, considerando modelos de mixtura finita de dos componentes con modelo de regresión lineal normal para las medias y modelo de heterocedasticidad multiplicativa para la varianzas de cada componente. Lo anterior con el fin de evaluar el desempeño del algoritmo para estos modelos.

En la tabla 3.1 se proponen cinco modelos de mixtura de la forma 3.1 para simulación: **M1**, **M2**, **M3**, **M4** y **M5**, cuyas características generales se tratan más adelante en el diseño de la simulación.

La tabla 3.1 contiene la configuración de cada modelo en la que se consideró: el tamaño de muestra (n) de cada replica de la simulación del respectivo modelo, el vector de parámetros del modelo para la media de la primera componente $\boldsymbol{\beta}_1$, el vector de parámetros del modelo para la media de la segunda componente $\boldsymbol{\beta}_2$, el vector de parámetros del modelo para la varianza de la primera componente $\boldsymbol{\gamma}_1$, el vector de parámetros del modelo para la varianza de la segunda componente $\boldsymbol{\gamma}_2$, el vector de los pesos correspondientes a cada componente de la mixtura y el número de réplicas. De esta forma, por ejemplo, el modelo M1 presenta la siguiente configuración, (Ver tabla 3.1):

- Tamaño de muestra $n = 100$
- Vector de parámetros del modelo para la media de la primera componente $\boldsymbol{\beta}_1 = (3, 2)$
- Vector de parámetros del modelo para la media de la segunda componente $\boldsymbol{\beta}_2 = (-2, -3)$
- Vector de parámetros del modelo para la varianza de la primera componente $\boldsymbol{\gamma}_1 = (-6, 0.52, 0.25)$
- Vector de parámetros del modelo para la varianza de la segunda componente $\boldsymbol{\gamma}_2 = (-5, 0.45, 0.15)$
- Pesos correspondientes a cada componente $\pi_1 = 0.4$ y $\pi_2 = 0.6$,
- 1200 réplicas.

3.6.1. Diseño del Estudio

A continuación se describen las características consideradas en el desarrollo del estudio.

Dadas

1. **Tamaño de la muestra y cantidad de réplicas:** cada conjunto de datos consta de 1100 réplicas, cada una con un tamaño de muestra n, donde n puede tomar $100, 200, 300$ y 400 dependiendo del modelo.

2. **Conjunto de datos:** La $i-$ésima observación, que se representa por el vector $(y_i, \boldsymbol{x}_i, \boldsymbol{w}_i)$, se genera a partir del siguiente esquema:

   (a) Inicialmente se genera un número aleatorio de una distribución de probabilidad discreta cuyos valores de la variable aleatoria corresponden a la posible $j-$ésima componente de la mixtura de la cual proviene la $i-$ésima observación, y la función de masa de probabilidad corresponde a los pesos de cada componente. Así pues el número generado permite seleccionar una componente de regresión en particular de la cual se generará la observación y_i.

   (b) Después se genera valores aleatorios de una distribución uniforme en el intervalo $[a, b]$ para cada una de las variables explicativas de la media x_r, $r = 1, ..., k$, y las variables explicativas del modelo heterocedastico de la varianza w_s $s = 1, ..., p$, de la componente seleccionada, donde el intervalo $[a, b]$ corresponde al dominio de cada una de las variables x_r y w_s.

   (c) Una vez identificada la componente de la cual proviene la observación y los valores de las variables x_r y w_s se evalúa la media $\mu_{y_i|\boldsymbol{x}_i}$ y la varianza $\sigma_{y_i|\boldsymbol{w}_i}$ con los parámetros poblacionales correspondientes

   (d) Finalmente se genera una valor aleatorio de una distribución normal con media $\mu_{y_i|\boldsymbol{x}_i}$ y varianza $\sigma_{y_i|\boldsymbol{w}_i}$, para obtener el valor y_i.

   De esta forma queda conformada la $i-$ésima observación.

3. **Criterio de parada:** las iteraciones se detienen cuando la diferencia entre la estimación t y la $t + 1$ sea menor de 10^{-4}.

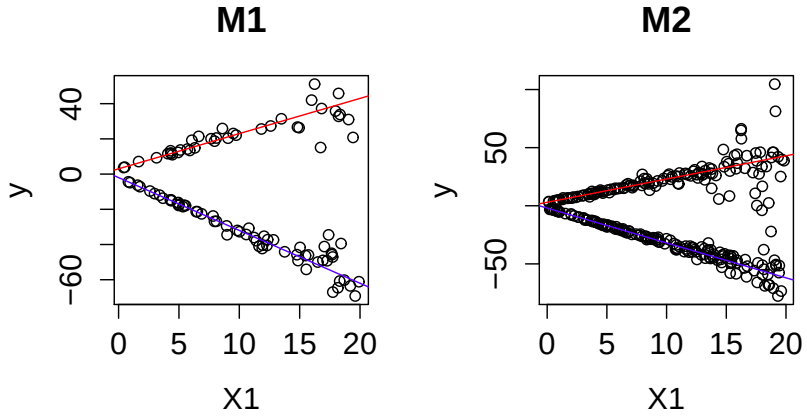


FIGURA 3.1. Modelo 1 y 2 de estudio de simulación

4. **Modelos de mixtura para la simulación:** Se proponen 5 modelos de mixtura de dos componentes: **M1**, **M2**, **M3**, **M4** y **M5** tal que las diferencias entre ellos se hallan, por ejemplo, en la selección del tamaño de muestra, los pesos de cada componente de la mixtura, los valores dados a los parámetros del modelo de la

media y modelo de la varianza, especificándose para el respectivo caso el número de variables explicativas para la media y la varianza. Para analizar el ajuste del modelo, se definió las rectas de regresión poblacional, así como el modelo de la varianza. A continuación se describen de forma general los modelos y en las figuras 3.1, 3.2, 3.3 se muestran ejemplos del conjunto de datos generados para cada simulación así como sus rectas de regresión poblacional. La tabla 3.1 resume la configuración dada a los modelos.

El modelo **M1** es un modelo con tamaño de muestra $n = 100$ y que se ha configurado para dos componentes de mixtura, con rectas de regresión concurrentes, heterocedasticidad baja para los primeros valores del dominio de la variable explicativa de cada componente y alta al final del mismo. Se usa una variable explicativa para la media x_1, cuyo dominio es $[0, 20]$ y para el modelo de la varianza dos variables explicativas w_1 y w_2 donde una de ellas corresponde a la usada para la media, esto es, $w_1 = x_1$. La variable w_2 también tiene como dominio $[0, 20]$. Para el modelo **M2**, se tienen las mismas características de **M1**, excepto por el tamaño de la muestra, que es $n = 300$ (ver figura 3.1).

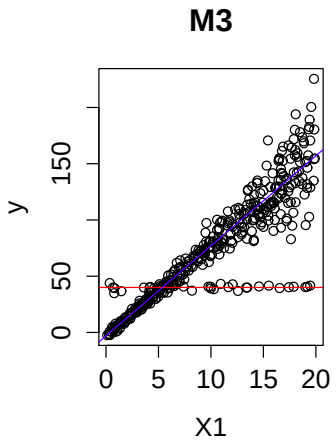


FIGURA 3.2. Modelo 3 de estudio de simulación

En el modelo **M3** el tamaño de muestra es $n = 400$. La varianza se modela con una variable explicativa y esta corresponde a la misma variable explicativa de la media, esto es, $w_1 = x_1$. El dominio de esta variable es $[0, 20]$. Una de las componentes tiene mayor proporción que la otra. La heterocedasticidad en la componente de menor proporción es alta en el inicio del dominio de la variable explicativa y baja al final, mientras que en la otra componente es baja al inicio y alta al final. Una de las rectas de regresión es paralela al eje X, (ver figura 3.2).

El modelo **M4** tiene tamaño de muestra $n = 200$, las rectas de regresión poblacional son paralelas donde una de ellas es homocedastica. Una sola variable explicativa para la media. Esta variable explicativa se utilizó en el modelo de la varianza como se muestra a continuación: $\sigma_i^2 = \exp(\gamma_{0j} + \gamma_{1j} w_1 + \gamma_{2j} w_1^2)$, con $j = 1, 2$ El dominio de w_1 es $[0, 20]$.

Para el modelo **M5**, se propuso las mismas características que para el modelo **M4** con la diferencia en el tamaño de muestra, que para este caso es $n = 100$.(ver figura 3.3).

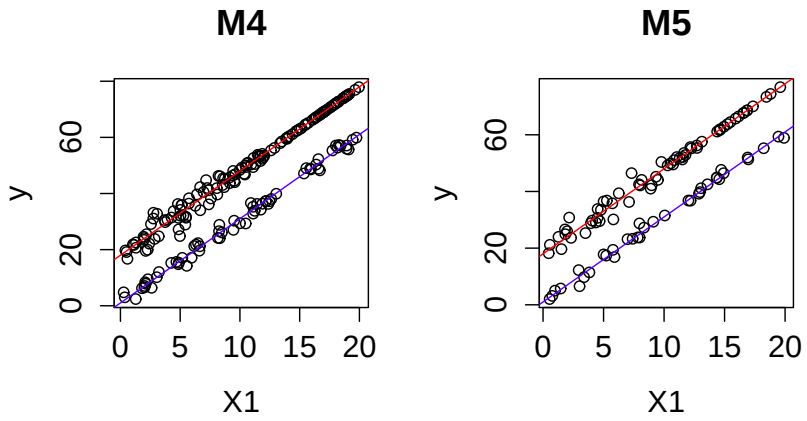


FIGURA 3.3. Modelo 4 y 5 de estudio de simulación

5. **Medidas del rendimiento del algoritmo:** para determinar el rendimiento del algoritmo en los diferentes modelos propuestos se tienen en cuenta las siguientes medidas:

   - Número promedio de iteraciones de algoritmo EM y del algoritmo de Fisher Scoring.
   - Sesgo de las estimaciones de los parámetros.
   - Error cuadrático de las estimaciones de los parámetros.

6. **Pasos de la simulación:** la simulación se desarrolla en los siguientes pasos:

   (a) Se genera el conjunto de datos de tamaño n.

   (b) Se ajusta el modelo. Esto corresponde hacer las estimaciones de los parámetros de acuerdo al modelo propuesto.

   (c) Se repite el paso anterior, tantas veces como réplicas se hayan dispuesto en el diseño del estudio.

   (d) Se calculan las medidas de rendimiento.

3.6.2. Resultados de la Simulación

Los resultados encontrados en la simulación para una mixtura de regresiones de dos componentes con modelado de varianza y de acuerdo con el diseño de simulación expuesto anteriormente se describen a continuación.

La tabla 3.2 muestra los valores promedio de iteraciones requeridas para la convergencia de los algoritmos EM y Fisher Scoring en cada uno de los modelos propuestos.

Para los modelos **M1** y **M2**, cuya diferencia está en el tamaño de la muestra los valores promedio de iteraciones requeridas para la convergencia de los dos algoritmos es menor en el modelo de mayor tamaño de muestra. Para los modelos M4 y M5, que también se diferencian en el tamaño de muestra, se necesitaron en promedio menor número de iteraciones para el algoritmo que tiene menor tamaño de muestra, en contraste con lo

	M1	**M2**	**M3**	**M4**	**M5**
EM	4.96	4.70	12.56	8.24	6.97
FS	23.22	18.30	57.23	50.00	45.74

TABLA 3.2. Resultados de desempeño del algoritmo EM-Fisher Scoring en las iteraciones de la estimación de parámetros

ocurrido entre **M1** y **M2**. Esto puede indicar que es recomendable hacer mayor número de iteraciones en los modelos con menor tamaño de muestra.

El modelo que necesito mayor número de iteraciones promedio para tener convergencia de los dos algoritmos fue el modelo **M3**. Los modelos **M4** y **M5** necesitaron mayor número de iteraciones en promedio para la convergencia tanto del algoritmo EM como el de Fisher Scoring, en comparación con los modelos **M1** y **M2**.

Estas diferencias en el número promedio de iteraciones como modelos que se relacionan tales como **M1** y **M2** o **M4** y **M5**, también tiene que ver con los valores iniciales. Los problemas que se refieren a la convergencia, como se mencionó anteriormente, han sido tratados en la literatura para el caso del algoritmo EM, (ver (**?**)), en donde se ha manifestado, por ejemplo, la necesidad de generar metodologías para la elección de valores iniciales con el fin de alcanzar el máximo global en menos iteraciones.

	π_1	π_2	β_{11}	β_{12}	β_{21}	β_{22}	γ_{11}	γ_{12}	γ_{13}	γ_{21}	γ_{22}	γ_{23}
Desvío	0.0057	-0.0057	-0.0031	0.0015	-0.0059	0.0005	-0.3614	0.0106	0.0111	-0.1883	0.0047	0.0044
ECM	0.0021	0.0021	0.0150	0.0014	0.0094	0.0008	0.6397	0.0023	0.0024	0.3799	0.0014	0.0015
Desvío	0.0015	-0.0015	-0.0012	0.0005	0.0011	-0.0004	-0.1040	0.0031	0.0027	-0.0542	0.0014	0.0012
ECM	0.0008	0.0008	0.0028	0.0003	0.0023	0.0002	0.1471	0.0006	0.0006	0.0875	0.0004	0.0004
Desvío	0.0111	-0.0111	-0.0486	0.0033	-0.0313	0.0077	-0.0883	0.0001	-0.0351	0.0003	NA	NA
ECM	0.0003	0.0003	0.9693	0.0037	0.1028	0.0050	0.2780	0.0018	0.0258	0.0002	NA	NA
Desvío	0.0014	-0.0014	0.0001	0.0000	0.0151	-0.0007	-0.0692	0.0178	-0.0012	-0.2004	0.0346	-0.0018
ECM	0.0010	0.0010	0.0003	0.0000	0.2060	0.0015	0.1507	0.0081	0.0000	0.4895	0.0252	0.0001
Desvío	-0.0080	0.0080	0.0033	-0.0003	0.0497	-0.0046	-0.2148	0.0471	-0.0026	-0.3425	0.0660	-0.0037
ECM	0.0019	0.0019	0.0039	0.0000	0.7023	0.0045	0.4093	0.0268	0.0001	1.1949	0.0626	0.0002

TABLA 3.3. Resultados de desempeño del algoritmo EM-Fisher Scoring en la estimación de parámetros

La tabla 3.3 muestra los valores del sesgo y del error cuadrático medio de las estimaciones de los parámetros en cada modelo propuesto. Las estimaciones obtenidas tienen sesgo que tiende a cero, con la excepción de los γ_{11}, y el error cuadrático medio tiende a disminuir a medida que el tamaño de muestra aumenta. Las estimaciones de los parámetros fueron buenas para todos los modelos, esto es indicador de un buen rendimiento del algoritmo.

Los valores de sesgo más considerables en general para todos los parámetros, fueron los presentados para los modelos **M3** y **M5**.

3.7. Aplicación

En el ejemplo de la sección 2.3.4 se estimó un modelo de mixtura de regresiones a una muestra de 45 puntajes del ECAES (hoy Saber Pro) 2009 en Economía, utilizando como variable explicativa de la media de las distribuciones el puntaje obtenido en la prueba de comprensión lectora. En esta aplicación se estimó un modelo de dos componentes, en donde la componente con mayor proporción de muestra presenta la mayor varianza. Se puede observar en esta componente una variación más grande del puntaje del ECAES a

Modelo	Componente	π	β_0	β_1	γ_0	γ_1	γ_2
Propuesta 1	1	0.572	−11.57	11.84	−5.728	0.939	$N.A$
	2	0.428	25.95	6.53	−30.19	3.14	$N.A$
Propuesta 2	1	0.405	14.04	9.81	−461.065	84.84	−3.86
	2	0.595	97.54	−0.40	−60.9628	13.919	−0.74989

TABLA 3.4. Estimación de parámetros de la mixtura de dos distribuciones normales con modelado de media y varianza vía algoritmo EM-Fisher Scoring de los puntajes totales de la prueba ECAES 2009 de economía

medida que aumenta el puntaje de la comprensión lectora, indicando con ello la presencia de heterocedasticidad en la recta de regresión, (ver figura 2.3).

Como la figura 2.3) sugiere que la regresión para esa componente es heterocedástica, se propone la estimación de un modelo de mixtura de regresiones con modelado de varianza, modelo que ha sido presentado en la sección 3.3. El puntaje en comprensión lectora correspondería a la variable que explica a la varianza.

En la búsqueda del modelo más adecuado, y teniendo en cuenta la forma en como están dispersos los datos respecto a las rectas de regresión ajustada en la aplicación de las sección 2.3.4, se proponen las siguientes alternativas dentro de la estructura para la varianza bajo el modelo general de heterocedasticidad de tipo multiplicativo:

Propuesta 1: $\sigma_i^2 = \exp(\gamma_0 + \gamma_1 w_1)$

Propuesta2: $\sigma_i^2 = \exp(\gamma_0 + \gamma_1 w_1 + \gamma_2 w_1^2)$

La tabla 3.4 contiene los parámetros estimados para el modelo con cada una de las propuestas en la estructura de la varianza. Este modelo se estimó con el algoritmo propuesto en la sección 3.4, implementado en el software R. Este procedimiento realiza la estimación vía máxima verosimilitud del modelo usando los algoritmos EM y Fisher Scoring.

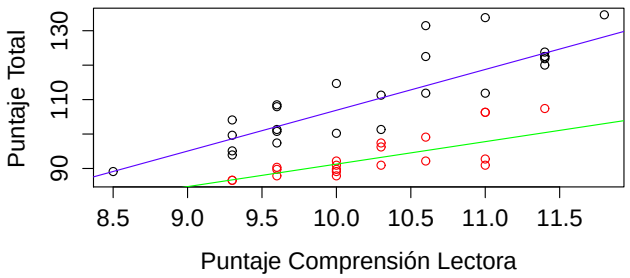


FIGURA 3.4. Rectas de regresión ajustadas al modelo de regresión de mixturas del puntaje total obtenido por una muestra de 45 estudiantes en la prueba ECAES de Economía 2009, con modelado conjunto de media y varianza. Con la propuesta 1 para la varianza

Con la propuesta 1 en la estructura de la varianza, el modelo estimado de dos componentes de mixtura agrupa al 42.7 % de la muestra en una de las componentes. En la figura 3.4 se observa que para esta componente menos variabilidad en él puntaje del ECAES a medida que aumenta el puntaje en la comprensión lectora. Clasifica las observaciones

que en general tienen los mejores puntajes en el ECAES y donde tanto el efecto directo como el transversal de la comprensión lectora hacen la diferencia en el puntaje total del Examen.

La otra componente que agrupa al $57,27\,\%$ de la muestra, agrupa a los puntajes bajos en el ECAES, mostrando mayor variabilidad a medida que aumenta el puntaje de comprensión lectora. Esta componente identifica observaciones donde el único efecto sobre el puntaje total del examen lo da el efecto directo de la comprensión lectora sin ningún efecto transversal de esta área.

Con la propuesta 2 en la estructura de la varianza, este modelo estimado agrupa en una de las componentes al $59,4\,\%$ de la muestra. Se observa al igual que en el modelo de la aplicación 2.3.4, para la componente con mayor proporción de muestra, una variación más grande del puntaje del ECAES a medida que aumenta el puntaje de la comprensión lectora, (ver figura 3.5). La recta de regresión es positiva, a mayor puntaje en comprensión lectora mayor puntaje en el ECAES.

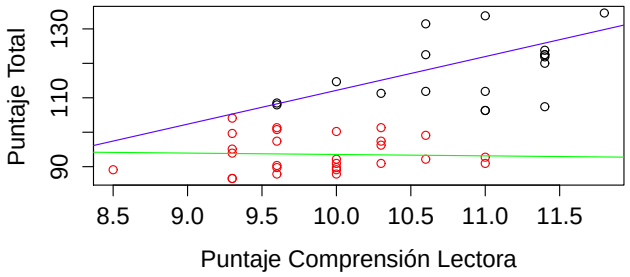


FIGURA 3.5. Rectas de regresión ajustadas al modelo de regresión de mixturas del puntaje total obtenido por una muestra de 45 estudiantes en la prueba ECAES de Economía 2009, con modelado conjunto de media y varianza. Con la propuesta 2 para la varianza

La otra componente que tiene un $40,5\,\%$ de las observaciones identifica puntajes bajos en el ECAES, con puntajes altos en comprensión lectora. La pendiente de la recta de regresión estimada para esta componente es negativa, implicando que a mayor puntaje en la prueba de comprensión lectora, menor puntaje en el ECAES, siendo esta conclusión poco acertada para la situación. El puntaje de la prueba de comprensión lectora tiene un efecto directo sobre el puntaje del ECAES pues ella independientemente de las otras áreas que se evalúan aporta al puntaje total del Examen. Al analizar si el efecto es positivo o negativo sobre el puntaje total de examen, la respuesta más apropiada sería positivo, pues su aporte directo por pequeño que sea, implicaría siempre un aumento en el puntaje total del examen.

De acuerdo con los criterios de información, los mejores resultados los obtuvo el modelo con la propuesta 1 para la varianza, (Ver tabla 3.5).

	AIC	CAIC	BIC	AIC3
Propuesta 1	347.288	370.286	361.28	356.28
Propuesta 2	359.44	387.54	376.54	370.44

TABLA 3.5. Criterios de Información para evaluar los modelos de mixtura con modelado de media y varianza propuestos para el puntaje total de la prueba ECAES 2009 de economía

Es importante resaltar que el conjunto de datos se ajusta bien a un modelo de mixtura de regresiones de dos componentes con modelado de varianza, pero no es mejor que el ajuste realizado en la sección 2.3.4.

Conclusiones

En este trabajo se presentó un modelo de mixtura de regresiones para m componentes de distribuciones normales con modelado de la varianza de las distribuciones. Por ser un modelo de mixturas, su aplicación se puede extender a una gran variedad de fenómenos aleatorios en donde la explicación de la varianza permitiría en general lograr un mejor ajuste del modelo y decir más del fenómeno al investigador.

El procedimiento propuesto para la estimación del modelo, permite identificar cluster de regresiones que presentan heterocedasticidad, ampliando la técnica para él análisis cluster en grupos de regresiones. Este procedimiento arrojó buenos resultados tanto en la parte práctica como en la simulación del algoritmo.

Al revisar las propuestas de modelado de heterocedasticidad en el modelo de regresión lineal normal, el algoritmo de Fisher Scoring por las características del problema, es una buena herramienta junto al algoritmo EM para la estimación de máxima verosimilitud de los parámetros del modelo de mixtura presentado en este trabajo.

En el desarrollo de este trabajo se identificó un grupo de datos al cual se les aplico un modelo de mixtura de regresiones con modelado de varianza. La muestra de 45 de los puntajes del ECAES (hoy Saber Pro) en Economía en 2009, se ajustó bien a una mixtura de dos componentes de distribuciones normales y luego a un modelo de mixturas de regresiones, explicando la media de la distribución a través del puntaje obtenido en la comprensión lectora. Esto mostró la heterocedasticidad presente en una componente. Este conjunto de datos se ajustó nuevamente bien a un modelo de regresión de mixturas con modelado de varianza pero se observó que éste no era mejor que el ajuste hecho con el modelo de regresión de mixturas. Una recomendación en este sentido está en considerar tamaños de muestra más grandes en las aplicaciones, que permitan evidenciar mejor el comportamiento del conjunto de datos.

En la simulación se destaca la forma positiva como el algoritmo reacciona en casos donde los datos tienen componentes con presencia de alta y baja heterocedasticidad. Se confirmó además, la importancia que tiene la elección de los valores iniciales en la convergencia del algoritmo. Este trabajo presenta una estrategia para la elección de estos valores en una mixtura de dos componentes, sin embargo es necesario buscar metodologías más generales que contribuyan al rendimiento del mismo.

Bibliografía

Aitkin, M. & Tunnicliffe, G. (1980). Mixture Models, Outliers, and the EM Algorithm, *Journal of the American Statistical Association* **22**(3): 325–331.

Böhning, D., Seidel, W., Alfó, M., Garel, B., Patilea, V. & Walther, G. (2007). Editorial: Advances in Mixture Models, *Comput. Stat. Data Anal.* **51**(11): 5205–5210.

Breusch, T. & Pagan, A. (1979). A Simple Test for Heteroscedasticity and Random Coefficient Variation, *Econometrica* **47**(5): 1287–94.

Charlie, C. & Wicksell, D. (1924). Arkiv for Matematik, *Astronomi och Fysik* **18**(6).

Cinca, A. (1996). *Estadística y Econometría*, McGraw-Hill.

Day, N. (1969). Estimating the components of a mixture of normal distributions, *Biometrika* **56**(3).

Dempster, A., Laird, N. & Rubin, D. (1977). Maximum likelihood from Incomplete Data Via the EM Algorithm, *Journal of the Royal Statistical Society* **39**(1): 1–38.

DeSarbo, W.and Cron, W. (1988). A Maximum Likelihood Methodology for Clusterwise Linear Regression, *Journal of Classification* (5): 249–282.

Ding, C. (2006). Using Regression Mixture Analysis in Education Research, *A Peer-Reviewed Electronic* **11**(11).

Everitt, B. (1984). Maximum Likelihood Estimation of the Parameters in a Mixture of Two Univariate Normal Distributions; A Comparison of Different Algorithms, *Journal of the Royal Statistical Society* **33**(2): 205–215.

Ganesalingam, S. & Mclachlan, G. (1978). The efficiency of a linear discriminant function based on unclassified initial samples, *Biometrika* **65**(3).

Garrido, L. & Cepeda, E. (2012). Mixture of Distributions in the Biparametric Exponential Family: A Bayesian Approach, *Communications in Statistics - Simulation and Computation* **41**(3): 355–375.

Garrido, L., Cepeda, E. & Achcar, J. (2011). Heteroscedastic Normal-Exponential Mixture Models: Bayesian and classical Approaches, *Applied Mathematics and Computation* **218**(7): 3635–3648.

Glejser, H. (1969). A New Test for Heteroskedasticity, *Journal of the American Statistical Association* **64**(325): 316–323.

Godfrey, L. (2006). Tests for Regression Models with Heteroskedasticity of Unknown Form, *Computational Statistics and Data Analysis* **50**(10): 2715 – 2733.

Goldfeld, S. & Quandt, E. (1965). Some Tests for Homoscedasticity, *Journal of the American Statistical Association* **60**(310): 539–547.

Goodman, L. (1974). Exploratory latent structure analysis using both identifiable and unidentifiable models, *Biometrika* **61**.

Greene, W. (2003). *Econometric Analysis*, Pearson Education.

Gujarati, D. (2010). *Econometria*, McGraw-Hill.

Harvey, A. (1976). Estimating Regression Models with Multiplicative Heteroscedasticity, *Econometrica* **44**(3): 461–465.

Hasselblad, V. (1966). Estimation of Parameters for a Mixture of Normal Distributions, *Technometrics* **8**(3).

Hathaway, R. (1985). A Constrained Formulation of Maximum-Likelihood Estimation for Normal Mixture Distributions, *The Annals of Statistics* **13**(2).

Hennig, C. (1996). Identifiability of Finite Linear Regression Mixtures, *Technical report*.

Hennig, C. (1999). Models And Methods For Clusterwise Linear Regression, Springer.

Hurn, M., Justel, A. & Rober, C. P. (2003). Estimating mixtures of regressions, *Journal of Computational and Graphical Statistics* **12**(1): 55–79.

Judge, G. (1988). *Introduction to the Theory and Practice of Econometrics*, Wiley.

Karlis, D. & Xekalaki, E. (2003). Choosing initial values for the em algorithm for finite mixtures, *Computational Statistics and Data Analysis* (3-4): 577 – 590.

Kiefer, N. (1978). Discrete Parameter Variation: Efficient Estimation of a Switching Regression Model, *Econometrica* **46**(2): 427–34.

Koenker, R. (1981). A Note on Studentizing a Test for Heteroscedasticity, *Journal of Econometrics* **17**(1): 107 – 112.

Lau, K., Leung, P. & Tse, K. (1999). A Mathematical Programming Approach to Clusterwise Regression Model and Its Extensions, *European Journal of Operational Research* **116**(3): 640 – 652.

Lee, T. (1973). Nonlinear methods in econometrics : S.M. Goldfeld and R.E. Quandt, *Journal of Econometrics* **1**(4): 399 – 401.

Long, J.and Ervin, L. (2000). Using Heteroscedasticity Consistent Standard Errors in the Linear Regression Model, *The American Statistician* **54**(3): 217–224.

Louis, A. (1982). Finding the Observed Information Matrix when Using the EM Algorithm, *Journal of the Royal Statistical Society* **44**(2).

McHugh, R. (1956). Efficient estimation and local identification in latent class analysis, *Psychometrika* **21**(4).

McLachlan, G. & Basford, K. (1988). Mixture models: inference and applications to clustering, *Marcel Dekker Inc* .

McLachlan G, P. D. (2000). *Finite Mixture Models*, John Wiley and Sons Inc.

M'Kendrick, A. (1925). Applications of Mathematics to Medical Problems, *Proceedings of the Edinburgh Mathematical Society* **44**: 98–130.

Newcomb, S. (1886). A Generalized Theory of the Combination of Observations so as to Obtain the Best Result, *American Journal of Mathematics* **8**.

Park, R. (1966). Estimation with Heteroscedastic Error Terms, *Econometrica* **34**(4): 888.

Pearson, K. (1894). Contributions to the Mathematical Theory of Evolution, *Philosophical Transactions of the Royal Society of London. A* **185**.

Prais, S & Houthakker, H. (1971). The analysis of family budgets, *Cup Archive* **4**.

Quandt, R. (1958). The Estimation of the Parameters of a Linear Regression System Obeying Two Separate Regimes, *Journal of the American Statistical Association* **53**(1): 873–880.

Quandt, R. (1972). A New Approach to Estimating Switching Regressions, *Journal of the American Statistical Association* **67**(338): 306–310.

Quandt, R. & Ramsey, J. (1978). Estimating Mixtures of Normal Distributions and Switching Regressions, *Journal of the American Statistical Association* **73**(1): 730–738.

Quandt, R. & Ramsey, J. (1979). Estimating Mixtures of Normal Distributions and Switching Regressions:Rejoinder, *Journal of the American Statistical Association* **74**(365): 56.

Redner, R. A. & Walker, H. F. (1984). Mixture Densities, Maximum Likelihood and the Em Algorithm, *SIAM Review* **26**(2): 195–239.

Román, A.and García, J. (2001). *Modelos Econometricos*, Pirámide.

Rutemiller, C. & Bowers, A. (1968). Estimation in a Heteroscedastic Regression Model, *Journal of the American Statistical Association* **63**(322): 552–557.

Ruud, P. (2000). *An Introduction to Classical Econometric Theory*, Oxford University Press.

Shalabh, Heumann, C., Grün, B. & Leisch, F. (2008). Finite Mixtures of Generalized Linear Regression Models.

Späth, H. (1979). Algorithm 39 Clusterwise linear regression, *Computing* **22**: 367–373.

Späth, H. (1981). Correction to Algorithm 39;Clusterwise Linear Regression, *Computing* **26**: 275.

Späth, H. (1982). Algorithm 48: A Fast Algorithm for Clusterwise Linear Regression, *Computing* **29**: 175–181.

Späth, H. (1985). Cluster Dissection and Analysis, *Wiley* **28**: 182.

Tan, W. & Chang, W. (1972). Some Comparisons of the Method of Moments and the Method of Maximum Likelihood in Estimating Parameters of a Mixture of Two Normal Densities, *Journal of the American Statistical Association* **67**(339).

Teicher, H. (1961). Maximum Likelihood Characterization of Distributions, *The Annals of Mathematical Statistics* **32**.

Titterington, D, S. A. & Makov, U. (1985). Statistical analysis of finite mixture distributions, *Wiley series in probability and mathematical statistics. Applied probability and statistics* .

Verbeek, M. (2008). *A Guide to Modern Econometrics*, Wiley.

Viele, K. & Tong, B. (2002). Modeling with Mixtures of Linear Regressions, *Statistics and Computing* **12**: 315–330.

Wedel, M. & DeSarbo, W. (1995). A mixture likelihood approach for generalized linear models, *Journal of Classification* **12**(1): 21–55.

White, H. (1980). A Heteroskedasticity-Consistent Covariance Matrix Estimator and a Direct Test for Heteroskedasticity, *Econometrica* **48**(4): 817–838.

Wolfe, J. (1970). Pattern clustering by multivariate mixture analysis, *Multivariate Behavioral Research* **5**(3).

Young, D. (2007). *A Study of Mixtures of Regressions*, The Pennsylvania State University.