UNIVERSIDAD DE ORIENTE

NÚCLEO DE SUCRE

ESCUELA DE CIENCIAS

DEPARTAMENTO DE MATEMÁTICAS

DETECCIÓN DE VALORES ATÍPICOS MULTINIVEL MEDIANTE UNA GENERALIZACIÓN DEL ESTADÍSTICO DE ANDREWS Y PREGIBON.

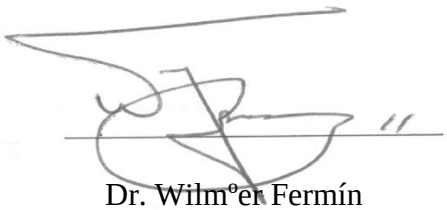
LCDO. ENRIQUE FERNANDO MACHADO RODRÍGUEZ

TRABAJO PRESENTADO COMO REQUISITO PARCIAL ÀRA OPTAR AL TÍTULO DE MAGISTER SCIENTIARUM EN MATEMÁTICAS

CUMANÁ, 2012

DETECCIÓN DE VALORES ATÍPICOS MULTINIVEL MEDIANTE UNA GENERALIZACIÓN DEL ESTADÍSTICO DE ANDREWS Y PREGIBON.

APROBADO POR:

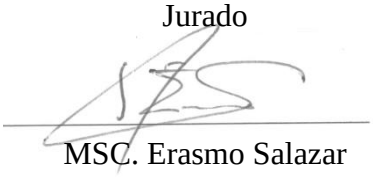


Dr. Wilmer Fermín

Asesor. Académico U.D.O

Dr. José Fermín

Jurado



MSC. Erasmo Salazar

Jurado

ÍNDICE GENERAL

DEDICATORIA i
AGRADECIMIENTO ii
ÍNDICE DE TABLAS iii
ÍNDICE DE FIGURAS iv
RESUMEN v
INTRODUCCIÓN 1
CAPÍTULO I 4
1.1 PLANTEAMIENTO DEL PROBLEMA 4
1.2 OBJETIVOS DE LA INVESTIGACIÓN 5
1.2.1 OBJETIVO GENERAL 5
1.2.2 OBJETIVOS ESPECIFICOS 5
1.3 JUSTIFICACIÓN 6
CAPÍTULO II 7
2.1 ANTECEDENTES 7
2.2 ASPECTOS BÁSICOS DEL ANÁLISIS MULTINIVEL 9
2.2.1 INTRODUCCIÓN 9
2.2.2 ANÁLISIS MULTINIVE 10
2.2.3 MODELO JERÁRQUICO LINEAL 11
2.2.4 MODELO JERÁRQUICO LINEAL A DOS NIVELES 12
2.2.5 MODELO VACÍO 12
2.2.6 MODELO DE INTERCEPTO ALEATORIO 14
2.2.7 MODELO DE INTERCEPTO Y PENDIENTE ALEATORIAS 15
2.2.8 VARIABILIDAD EN EL INTERCEPTO Y LA PENDIENTE 17
2.2.9 MODELO JERÁRQUICO LINEAL GENERAL A DOS NIVELES 18
2.2.10 SUPUESTOS DEL MODELO JERÁRQUICO LINEAL 22
2.2.11 RESIDUOS 23
2.3 CONCEPTOS BASICOS DE VALORES ATIPICOS 24
2.3.1 INTRODUCCIÓN. 24
2.3.2 Valores Atípicos 25
2.3.3 Rechazo de Valores Atípicos 27
2.3.4 Detección de Valores Atípicos en Regresión Lineal 27
2.4 Estadístico de Andrews y Pregibon (A.P) 29
2.4.1 Operador de Supresión 29

2.4.2 Operador completo. 30
2.4.3 Definición (estadístico A.P) 31
2.5 Métodos de Detección de Valores Atípicos Multinivel 31
2.6 Generalización del Estadístico de Andrews y Pregibon para datos Multinivel 36
2.6.1 Operador de Supresión multinivel 36
2.6.2 Estadístico de Andrews y Pregibon para datos Multinivel. 38
CAPÍTULO III 39
3.1.- NIVEL DE LA INVESTIGACIÓN 39
3.2.- DISEÑO DE LA INVESTIGACIÓN 39
3.3.- MATERIALES Y MÉTODOS 39
CAPÍTULO IV 41
4.1- PRESENTACIÓN Y ANÁLISIS DE RESULTADOS. 41
4.2.- MATRIZ DE DATOS REFERIDA AL RENDIMIENTO ESTUDIANTIL (RASBASH, ET. AL., 2001) 41
4.3.- DETECCIÓN DE OBSERVACIONES DISCREPANTES MEDIANTE MÉTODOS GRÁFICOS EN LOS DATOS. 42
4.3.1.- AJUSTE DEL MODELO DE LOS DATOS. 42
4.3.2.- Diagnóstico de escuelas discrepantes a través de residuos puros, residuos estandarizados, supresión residual, Leverage y DFITS. 48
4.4 Aplicación del estadístico de Andres y Pregibon multinivel. 52
4.5 Comparación empírica de los métodos gráficos y el Estaditico de (A.P). 54
CONCLUSIONES 56
RECOMENDACIONES 58
REFERENCIAS BIBLIOGRÁFICAS 59
ANEXOS 61
HOJAS DE METADATOS 62

DEDICATORIA

A Dios, el todo poderoso por ser mi guía y mi protector y por darme la sabiduría para ser el hombre que soy.

A mis padre por orientarme en el camino de triunfo, a mis hermanos y primos por brindarme su estimulo y amor incondicional.

A mi tutor por ser la persona que guio este proyecto y por brindarme su apoyo en los momento más indicado.

Y muy especialmente a mis princesas Ángeles Victoria y María Fernanda, por ser las mujeres que me dan la fortaleza para no rendirme.

A todas a1quellas personas que me dieron su apoyo incondicional.

AGRADECIMIENTO

Dedico mis gracias primero a Dios, es placentero expresar mis más sinceras gracias a todas aquellas personas que me han brindado su estímulo y apoyo para alcanzar esta meta.

Al grupo de profesores del departamento de Matemáticas que de alguna manera portaron un granito de arena para la culminación de este trabajo, en especial al profesor Franklin Astudillo por el valioso aporte a este trabajo.

ÍNDICE DE TABLAS

En la tabla 1.1 se presenta el estadístico de Andres y Pregibon multinivel para cada una de las 18 escuelas y para el mayor y el menor estudiante. 53

ÍNDICE DE FIGURAS

Figura 2.1 Ilustración de valores atípicos..... 25
Figura 4.1: Estructura Multinivel de los datos. Salida generada por el 43
Software MLwiN 1.1..... 43
Figura 4.2: Modelo vacio sobre el rendimiento estudiantil en los datos. Salida generada por el Software MLwiN 1.1..... 43
Figura 4.3: Modelo de intercepto y pendiente aleatorios para los datos. Salida generada por el Software MLwiN 1.1..... 45
Figura 4.4: (a) Curva promedio y (b) diagrama de dispersión de las escuelas. Salida generada por el Software MLwiN 1.1 46
Figura 4.5: Representación de los residuos de intercepto y pendiente de las escuelas. Salida generada por el Software MLwiN 1.1..... 47
Figura 4.6: Diagrama de pares de residuos para el intercepto y la pendiente de las escuelas. Salida generada por el Software MLwiN 1.1..... 48
Figura 4.7: gráficas de residuos puros, residuos estandarizados, supresión residual, leverage y DFITS con respecto al intercepto de las escuelas. Salida generada por el Software MLwiN 1.1 50
Figura 4.8: gráficas de residuos puros, residuos estandarizados, supresión residual, leverage y DFITS con respecto a la pendiente de las escuelas. Salida generada por el Software MLwiN 1.1..... 51
Figura 4.9: Residuos Estandarizados del nivel uno vs puntuación normal de zpfinal. Salida generada por el Software MLwiN 1.1..... 52

RESUMEN

En esta investigación se describen los diversos procedimientos para detectar valores discrepantes e influyentes en la modelización multinivel y se propone una nueva alternativa para detectar estos valores. El análisis multinivel es una metodología para analizar datos con estructura jerárquica o multinivel; estos datos se caracterizan por unidades individuales anidadas o agrupadas en niveles superiores. Al igual que en el análisis de regresión clásico donde las estimaciones se ven afectadas por valores atípicos, los datos multinivel pueden presentar uno o más valores atípicos que afectarían los coeficientes estimados del modelo de regresión jerárquico o multinivel ajustado. En análisis de regresión clásico existe una variedad de métodos analíticos y gráficos para detectar valores atípicos (ver Barnett y Lewis, 1978 citado por Montgomery, 1991) aunque éstos son una herramienta valiosa, en la mayoría de los casos es inevitable la subjetividad del investigador, por lo que la verdadera detección puede quedar en tela de juicio. Un estadístico interesante para detectar valores atípicos en regresión lineal es el de Andrews y Pregibon (1978), (ver también Fermín, 1993) el cual se basa en el determinante de la matriz de diseño con el vector de respuesta adjuntado; lo interesante de este estadístico es que detecta tanto los atípicos distantes como los atípicos influyentes en el modelo. Aquí se ha desarrollado un estadístico para la detección de unidades, del nivel dos como del nivel uno, discrepantes y/o influyentes en la estimación de los parámetros de un modelo lineal jerárquico de dos niveles de intercepto y pendiente aleatorios; el estadístico es una generalización del estadístico de Andrews y Pregibon usada en regresión clásica, al caso multinivel.

INTRODUCCIÓN

En cualquier disciplina es muy frecuente encontrar conjuntos de datos jerárquicos o multinivel; éstos se caracterizan porque las unidades de un nivel inferior están anidadas, agrupadas, en unidades más grandes y, a su vez éstas en otras unidades de un nivel superior, conformándose así una estructura jerárquica o multinivel. Quizás el ejemplo más clásico está en el área de educación; alumnos agrupados o anidados en clases y las clases, a su vez, agrupadas en escuelas, las escuelas en distritos y así sucesivamente. Los datos multinivel exhiben patrones complejos de variabilidad a causa de la estructura anidada; esta complejidad se refleja en una variabilidad en cada nivel de la jerarquía y en una dependencia entre las unidades u observaciones anidadas en las unidades superiores de cada nivel de la jerarquía. Hoy por hoy es común hablar de datos jerárquicos o multinivel en ciencias sociales, salud, biología, agricultura, geografía, economía e incluso en diseño de experimentos (ver Fermín 2004; 2007).

Dada la compleja estructura de variabilidad que presentan los datos multinivel se requieren procedimientos específicos para el análisis de este tipo de datos. Raudenbush (1988) plantea que modelos de regresión tradicional fracasan en el análisis de datos multinivel porque la suposición de observaciones independientes sacadas de una población homogénea se infringe (es insostenible). Por otra parte, el uso del análisis de regresión para el análisis de datos multinivel usando procedimientos de agregación y/o desagregación de las variables también han sido desacreditados (ver Bijleveld y van der Kamp (Eds.), 1998; Lewis, 1997; Snijders y Bosker, 1999). La metodología apropiada para el análisis de datos jerárquicos es el Análisis Multinivel; ésta metodología considera la formulación de una función matemática para modelizar los datos tomando en cuenta la estructura jerárquica mediante la incorporación de variables de todos los niveles de la jerarquía y la determinación de los errores para cada nivel. Así, el análisis multinivel proporciona al investigador un conjunto de herramientas flexibles y poderosas para tratar datos multinivel en cuanto al ajuste de modelos, pruebas de hipótesis y la caracterización de la estructura de covarianza que emerge del agrupamiento en los datos tomando en cuenta que las variables pueden tener efectos diferentes en los diferentes niveles, los efectos pueden variar entre unidades y puede haber efectos cruzados; esto es, interacción entre efectos de diferentes niveles.

Al igual que en el análisis de regresión clásico donde las estimaciones se ven afectadas por valores atípicos, los datos multinivel pueden presentar uno o más valores atípicos que afectarían los coeficientes estimados del modelo de regresión jerárquico o multinivel ajustado. En análisis de regresión clásico existe una variedad de métodos analíticos y gráficos para detectar valores atípicos (ver Barnett y Lewis, 1978 citado por Montgomery, 1991). Un estadístico interesante para detectar valores atípicos en regresión lineal es el de Andrews y Pregibon (1978), (ver también Fermín, 1993) el cual se basa en el determinante de la matriz de diseño con el vector de respuesta adjuntado; lo interesante de este estadístico es que detecta tanto los atípicos distantes como los atípicos influyentes en el modelo.

Los datos multinivel no escapan de la presencia de valores discrepantes y el efecto que tienen éstos en la forma de análisis de los datos multinivel. Los valores discrepantes multinivel son más complejos que en el caso de datos de un único nivel puesto que esos valores pueden estar en cualquier nivel de la estructura jerárquica. Rasbash et. al (2001) presentan datos multinivel referidos a alumnos en escuelas; ellos descubren que hay alumnos discrepantes en una escuela y también que hay una escuela discrepante de las demás escuelas: es decir, los datos tienen valores discrepantes en ambos niveles. Los valores discrepantes en datos multinivel complican el análisis puesto que aumentan la complejidad del modelo multinivel. (Langford y Lewis, 1998).

En un análisis multinivel, hay dos momentos en que los valores extremos pueden ser detectados: primero con técnicas de exploración segundo después del análisis utilizando los residuos. Mediante técnicas gráficas. Las Técnicas de análisis exploratorio y de los modelos multinivel se han discutido en Rasbash et. al., (2001). De las técnicas analíticas o con métodos matemáticos de detección de valores discrepantes multinivel se ha explorado muy poco para la elaboración de los modelos multinivel (ver Barnett y Lewis, 1994).

Esta investigación tiene como propósito presentar una propuesta analítica de detección de valores atípicos multinivel. La propuesta se refiere a una extensión del estadístico de Andrews y Pregibon que permitan detectar valores atípicos, en datos de un único nivel, cuyo análisis se puede realizar mediante la regresión lineal clásica, a la situación que pueda detectarse outlier multinivel presentes en datos multinivel, bajo una estructura jerárquica de dos niveles, cuyo análisis se realiza con la metodología del análisis multinivel ajustando un modelo lineal jerárquico.

La presentación de este trabajo se organizó en cuatro capítulos. En el primer capítulo se presenta el planteamiento del problema, los objetivos y la justificación de la investigación. En el capítulo dos se muestran algunos antecedentes del tema investigado, fundamentos teóricos y métodos de análisis utilizados en la detección de outliers multinivel. El capítulo tres contiene el nivel y diseño de la investigación, técnicas de procesamiento y análisis de datos. El capítulo cuatro presenta los resultados de la aplicación práctica de los métodos existentes y el estadístico de Andrews y Pregibon propuesta para la detección de valores discrepantes en el caso Multinivel. Para finalizar se establecen reflexiones sobre los objetivos propuestos y se evalúa su cumplimiento, se sugieren también líneas de investigación que podrían conducir a dar continuidad a la investigación en curso.

CAPÍTULO I

1.1 PLANTEAMIENTO DEL PROBLEMA

Hasta ahora la manera más común de detectar valores atípicos multinivel es mediante diversos métodos gráficos; sin embargo, aunque éstos son una herramienta valiosa, en la mayoría de los casos es inevitable la subjetividad del investigador, por lo que la verdadera detección puede quedar en tela de juicio. Por otra parte, existe poca información respecto a métodos analíticos basados en procedimientos matemáticos para detectar valores atípicos multinivel. En otras palabras, la detección de valores atípicos multinivel es un campo de investigación que tiene mucho por explorar. En análisis de regresión clásica existen una gran cantidad de procedimientos gráficos y analíticos para detectar valores atípicos; el estadístico de Andrews y Pregibon es un procedimiento analítico que permite detectar valores (distantes e influyentes) atípicos en regresión lineal. (ver Andrews y Pregibon; 1978)

En la modelización multinivel los valores discrepantes también pueden influir indebidamente en la estimación de los parámetros. No obstante con esta metodología la literatura ofrece pocas opciones para la detección de valores atípicos (ver Langford y Lewis, 1998). La detección de valores atípicos se realiza comúnmente a través de gráficos de algunas medidas utilizadas en la regresión clásica y extendidas a la metodología multinivel; entre las medidas están los residuos puros, los residuos estandarizados, la supresión residual, los valores de influencia, y los DFITS; estas medidas se grafican en conglomerados jerárquicos (ver Rasbash, et, Al., 2001).

Indudablemente el análisis de los diferentes gráficos revela la presencia de valores potencialmente discrepantes, mediante valores alejados. No obstante, en estos gráficos la subjetividad está muy presente para considerar un valor distante o no. Es importante adaptar, en el ámbito del análisis multinivel, métodos y/o estadísticos más objetivos para la detección de los valores atípicos y así contar con un mayor número de técnicas de diagnostico para datos multinivel.

Con el propósito de adaptar métodos más objetivos a las estrategias gráficas, para detectar outliers multinivel, surgió la idea de adaptar el estadístico de Andrews y Pregibon usado en regresión lineal múltiple, al caso que se puedan detectar los valores discrepantes multinivel, y así ampliar las posibilidades de detectar este tipo de datos.

1.2 OBJETIVOS DE LA INVESTIGACIÓN

1.2.1 OBJETIVO GENERAL

- Generalizar el método de Andrews y Pregibon para la detección de valores atípicos en análisis multinivel.

1.2.2 OBJETIVOS ESPECIFICOS

- Enunciar los efectos que produce los valores atípicos en la estimación de los coeficientes de un modelo jerárquico de dos niveles.

- Adaptar la matriz ampliada de datos de un único nivel al caso de datos jerárquicos de dos niveles

- Aplicar el operador complemento a la matriz ampliada de datos jerárquicos de dos niveles

- Definir el estadístico, bajo la perspectiva de Andrews y Pregibon para detectar valores discrepantes multinivel.

- Aplicar el estadístico obtenido en (4) a un conjunto de datos reales para detectar valores atípicos.

1.3 JUSTIFICACIÓN

En muchas situaciones experimentales queremos investigar y modelar las relaciones cuantitativas entre variables, un método útil para proceder con esta situación es la regresión clásica, una vez obtenida la ecuación lineal que se ajusta a los datos podemos encontrar que los coeficientes estimado están indebidamente influenciado por una o más observaciones (valores atípicos), por lo que la ecuación estimada no es idónea para describir la verdadera realidad del proceso investigado, actualmente en regresión lineal clásica existen muchos procedimientos para detectar y tratar valores atípico, mientras que en el análisis multinivel hasta ahora la manera más común de detectar valores atípicos multinivel es mediante diversos métodos gráficos, aunque éstos son una herramienta valiosa, en la mayoría de los casos es inevitable la subjetividad del investigador, por lo que la verdadera detección puede quedar en tela de juicio. Crear un modelo analítico para detectar y tratar valores atípico multinivel es un valor teórico de relevancia para la comunidad científica que modélica datos jerárquicos, de los cuales sabemos que son propensos a tener cualquier valor discrepante que puedan influenciar indebidamente la estimación de la ecuación multinivel. Con estas razones o planteamiento queda plenamente justificada y pertinente la investigación planteada.

CAPÍTULO II

2.1 ANTECEDENTES

En esta sección se presenta un resumen de la literatura consultada para la realización de este trabajo y cuyos contenidos se relacionan de una u otra forma con el mismo, tal como se muestra a continuación:

David F. Andrews y Pregibon (1978), plantean un método para examinar, en un modelo lineal, la presencia de uno a o más valores extraños importantes. Ellos presentan su estadístico A.P y discuten las propiedades de distribución del mismo. Este estadístico es capaz de detectar uno o más valores extraños con atención directa con los que pueden influir, en los parámetros estimados. Para la definición del estadístico A.P, ellos previamente definen un operador $D_{ij}^{(k)}$ ... (.) que suprime k filas de una matriz con n filas, k<n.

Cook R.D. y Wang P.C (1983), proponen un método de diagnostico para determinar el grado en el cual casos individuales y grupos de casos influencian el parámetro estimado de la transformación de *Box-Cox* para variables respuesta en el modelo de regresión lineal. Definen una distancia de verosimilitud que es una comparación de máxima verosimilitud, $\hat{\lambda}$, de λ de la transformación de Box-Cox con $\hat{\lambda}_{(i)}$, la estimación de máxima verosimilitud de λ en el modelo de supresión, reflejando así la inseguridad del estimador $\hat{\lambda}$.

Goldstein, H. (1995), en su obra "Multilevel Statistical Models", presenta descripciones de la modelación estadística y análisis de datos con estructura jerárquica, aplicados en áreas tales como educación, epidemiología, geografía, ciencia social, entre otros. El autor se refiere al término de jerarquía como unidades agrupadas en diferentes niveles, donde no se puede ignorar la importancia de los efectos de los grupos anidados en otros, ya que las técnicas de análisis estadísticos no serían confiables. El texto está enfocado primordialmente en la estimación de modelos generales de dos niveles incluyendo coeficientes aleatorios, además, expone ilustraciones gráficas para la detección de valores atípicos, el software MLwin nos facilita una estimación V_{0k} y V_{1k} de estos residuos y de sus diagnósticos de errores estándar (Goldstein, 1995, Apéndice 2.2) el cual denotamos mediante de (V_{ok}) y se (V_{1k}).

Langford y Lewis, (1998), proponen enfocarse en los residuos estudentizados, los valores de influencia o en una medida de la influencia llamada DFITS**.** Los residuos estudentizados, y en particular los residuos de eliminación, consistentes con el modelo en el cual se basan, pueden ser considerados como cantidades aproximadas normales N(0,1). Entonces, para propósitos de diagnostico, un valor fuera del rango – 3 a +3 indica que el punto de datos no es consistente con el modelo. Este punto es etiquetado como "especial".

John F. Bell y Eva Malacova (2004), en la publicación titulada "Outliers y modelos Multinivel" consideran algunas dificultades que causan los valores atípicos multinivel en los modelos jerárquicos. Analizaron dos conjuntos de datos: los primeros datos fueron referentes al progreso de la educación secundaria y los otros al proceso de admisión de estudiantes en la educación superior. En ambos conjuntos de datos se demostró que había procesos contaminantes que generaban valores atípicos y, en consecuencia los modelos multinivel ajustados eran innecesariamente complejos. Fue necesario considerar cuidadosamente la naturaleza de los datos y el proceso modelado, con esto se logró eliminar los outliers y en consecuencia obtener un modelo multinivel confiable.

Shi Lei y Chen Gemai **(2007),** en su trabajo "Detección de Outliers en Modelos Multinivel" estudian la influencia local de observaciones discrepantes en los modelos multinivel. Para tal fin ellos realizan cambios simultáneamente en las varianzas, la respuesta y la matriz de diseño. El cambio local causado por las perturbaciones se mide mediante el estadístico de Cook a través de los parámetros fijos y aleatorios estimados. Se aplicó la propuesta en datos simulados mostrando buena eficiencia en la detección de influencia de valores atípicos multinivel.

2.2 ASPECTOS BÁSICOS DEL ANÁLISIS MULTINIVEL

2.2.1 INTRODUCCIÓN

El modelo lineal jerárquico es un modelo de coeficientes aleatorios para datos multinivel o datos estructurados jerárquicamente y es, por ahora, la principal herramienta del análisis multinivel. Este modelo representa relaciones dentro de grupos y entre grupos en un único análisis; se concibe la variación no explicada dentro de los grupos y la variación no explicada entre los grupos como aleatorias. Por ejemplo, para un estudio de alumnos dentro de escuelas, esto significa que no solamente hay variación no explicada entre los alumnos, sino que también hay variación no explicada entre escuelas y es considerada como variabilidad aleatoria.

Para el estudio del modelo lineal jerárquico o multinivel se considerará una estructura jerárquica de dos niveles. Las unidades del nivel uno son referidas como unidades individuales, y las unidades del nivel dos son referidas como ”grupos”. Es evidente que se pueden encontrar otros nombres para las unidades dependiendo del área de aplicación. Por ejemplo, si la aplicación es en estudios longitudinales las unidades del nivel uno son las ocasiones de medición y las unidades del nivel dos son los sujetos. Se denotará por el número de grupos en los datos y, n_j representa el número de unidades individuales en el j-ésimo grupo, j=1,…N. El número total de individuos será $M = \sum_j^N n_j$.

El modelo lineal jerárquico es un tipo de **modelo de regresión** particularmente adecuado para datos multinivel. Éste difiere del modelo de regresión múltiple usual en el hecho de que la ecuación que define al modelo lineal jerárquico contiene más de un término de error: uno (o más) para cada nivel. Como en todos los modelos de regresión, hay una distinción entre variables dependientes y variables explicativas o regresoras. Se considerara una sola variable dependiente y el objetivo es construir un modelo que exprese cómo la variable dependiente depende de, o es explicada por, las variables explicativas (regresoras, predictoras). La variable dependiente debe ser medida al nivel uno: el modelo lineal jerárquico es un modelo para explicar lo que está pasando en el livel más bajo, en el nivel más detallado de la jerarquía.

A lo largo de la descripción del modelo lineal jerárquico se asumirá que están disponibles variables explicativas en ambos niveles de la estructura. En la notación que se seguirá se distinguen los siguientes tipos de índices y variables:

- j es el índice para indicar los grupos o unidades del nivel dos (j=1, 2, …, N),
- i es el índice para indicar los individuos dentro de los grupos (i=1, 2, …,n_j),
- Para el individuo i en el grupo j se tienen las siguientes variables:
    - Y_{ij} es la variable dependiente o respuesta. Se considera que está variable es continua.
    - X_{ijs} es la variable explicativa al nivel 1. s=1, 2, …, p
- Para el j-ésimo grupo se tienen las siguientes variables:
    - Z_{jl} es la variable explicativa al nivel 2. s=1, 2, …, q.

Para entender la notación es importante darse cuenta que los índices "i" y "j" indican justamente de lo que las variables dependen. Por ejemplo, Y_{ij} indica que el valor de la variable Y depende del individuo i dentro del j-ésimo grupo.

2.2.2 ANÁLISIS MULTINIVE

El análisis multinivel es básicamente una metodología para el análisis de datos jerárquicos o multinivel con complejos patrones de variación, con atención en las fuentes de variabilidad anidadas. En este tipo de análisis la variación entre las unidades de los distintos niveles es el elemento de principal atención. La metodología considera la formulación de una función matemática para modelizar los datos tomando en cuenta esa estructura jerárquica y hacer posible la incorporación de variables de todos los niveles de la jerarquía.

Los coeficientes de regresión varían de nivel a nivel, debido a que el efecto de las variables puede variar en los distintos grupos. Si no se considera la estructura jerárquica en el análisis, las estimaciones de los parámetros serían sesgadas; el análisis multinivel modela explícitamente estas relaciones jerárquicas, produciendo estimaciones no sesgadas (Fermín, 2004)

2.2.3 MODELO JERÁRQUICO LINEAL

El principal modelo multinivel es el Modelo Jerárquico Lineal, el cual es una extensión del modelo de regresión lineal múltiple que incluye coeficientes aleatorios de regresión anidados. El modelo jerárquico lineal difiere del modelo de regresión múltiple, en el sentido que contiene más de un término de error, uno o más para cada nivel de la jerarquía. Este modelo puede definirse en una estructura jerárquica de dos, tres e incluso cuatro niveles o jerarquías, permitiendo la determinación de las relaciones dentro y entre las unidades de cada nivel de las jerarquías; mediante la modelización de la respuesta principal, generalmente medida al nivel individual o nivel uno (nivel más bajo de la jerarquía) y explicar los cambios de ésta causada por las variables explicativas consideradas en los distintos niveles. El más común es el modelo jerárquico lineal de dos niveles; el cual se desarrolla en esta sección desde el modelo vacío hasta el modelo general.

Cuando se está en presencia de un modelo jerárquico lineal cada uno de los niveles de la estructura jerárquica está representado por su propio submodelo; estos submodelos a su vez expresan las relaciones entre las variables dentro de un nivel dado y cuando se integran (forman el modelo jerárquico lineal), especifican cómo las variables de un nivel influencian las relaciones que ocurren en otro (efectos de niveles cruzados). Además con este modelo se facilita la estimación de las componentes de varianzas y covarianzas con datos anidados desbalanceados. El modelo jerárquico lineal, como muchos otros modelos de regresión, hace una distinción entre la variable respuesta (dependiente) y variables explicativas (independientes). La variable respuesta debe ser una variable del nivel uno, ya que el modelo jerárquico lineal es un modelo para explicar lo que sucede en el nivel más bajo de la jerarquía; las variables explicativas, como se mencionó anteriormente, pueden ser de cualquier nivel de la jerarquía. El fin, es construir un modelo que exprese cómo la variable respuesta depende de, o es explicada por, las variables independientes

2.2.4 MODELO JERÁRQUICO LINEAL A DOS NIVELES

El modelo jerárquico lineal más común en la literatura es el definido bajo una estructura de dos niveles (por ejemplo, alumnos en escuelas). A continuación se desarrolla dicho modelo pasando por el más simple, el modelo vacío, hasta el más complejo, su generalización; se presenta también supuestos del modelo.

2.2.5 MODELO VACÍO

El modelo jerárquico lineal de dos niveles más simple es el "Modelo Vacío", un tipo de modelo donde no existe ninguna variable explicativa. Su objetivo es tratar de detectar diferencias significativas entre los grupos.

El modelo dentro de grupos contiene dos términos, expresados de la siguiente manera:

$$Y_{ij} = \beta_{0j} + \varepsilon_{ij} \quad \begin{cases} i = 1,2,\ldots, n_j \\ j = 1,2,\ldots, m \end{cases} \qquad \text{Submodelo del nivel 1} \qquad (2.1)$$

donde, β_{0j} es un coeficiente aleatorio que varía de grupo a grupo, ε_{ij} es el término de error, n_j representa el número de unidades en el j-ésimo grupo y m representa el número de grupos.

La modelización simple para el coeficiente β_{0j}, consiste en un parámetro fijo y una componente aleatoria que representa la fluctuación del grupo respecto al parámetro fijo. Denotado como:

$$\beta_{0j} = \beta_0 + U_{0j} \; con \; j = 1,2,\ldots, m \qquad \text{Submodelo del nivel 2} \qquad (2.2)$$

Sustituyendo (2.2) en (2.1), se obtiene Modelo Jerárquico Lineal Vacío, cuya expresión es:

$$Y_{ij} = \beta_0 + U_{0j} + \varepsilon_{ij} \quad \begin{cases} \mathrm{i} = 1,2,\ldots, n_j \\ \mathrm{j} = 1,2,\ldots \mathrm{m} \end{cases} \qquad (2.3)$$

Este modelo como todos los modelos jerárquicos lineales puede ser expresado como la suma de:

Una parte fija, la cual es β_0 y representa la media global de todos los grupos.

Una parte aleatoria la cual es $U_{0j} + \varepsilon_{ij}$; donde el término U_{0j} representa el efecto aleatorio del j-ésimo grupo y el término ε_{ij} representa el error o residuo de la respuesta del i-ésimo sujeto en el j-ésimo grupo. El término U_{0j} se supone que tiene distribución independiente $N(0, \sigma_U^2)$ y ε_{ij} también se supone que tiene distribución independiente $N(0, \sigma_e^2)$.

El modelo vacío sólo contiene grupos aleatorios y variación aleatoria dentro de los grupos, y la variación total esta claramente particionada entre los dos niveles; la varianza al nivel de grupos σ_U^2, y la varianza al nivel uno σ_e^2, ésta expresada de la siguiente forma:

$$V(Y_{ij}) = \sigma_U^2 + \sigma_e^2 \tag{2.4}$$

Además la covarianza es igual a la varianza a nivel de los grupos, es decir,

$$Cov(Y_{ij}, Y_{i'j}) = \sigma_U^2 \tag{2.5}$$

y el coeficiente de correlación intra-grupos ρ, se define tal que:

$$\rho(Y_{ij}, Y_{i'j}) = \frac{\sigma_U^2}{\sigma_U^2 + \sigma_e^2} \tag{2.6}$$

Este coeficiente se puede interpretar de dos formas:

1. La proporción de la varianza debida al nivel grupal o nivel dos.
2. La correlación entre dos unidades del nivel uno seleccionadas aleatoriamente desde una unidad (grupo) al nivel dos seleccionada aleatoriamente.

2.2.6 MODELO DE INTERCEPTO ALEATORIO

Otro caso sencillo en el modelo jerárquico lineal, es el "Modelo de Intercepto Aleatorio" este modelo incluye al menos una variable explicativa del nivel uno con coeficientes fijos y el único coeficiente aleatorio es el intercepto. Con una variable explicativa, X al nivel uno, el modelo viene dado por:

$$Y_{ij} = \beta_{0j} + \beta_1 X_{ij} + \varepsilon_{ij} \quad \begin{cases} \text{i} = 1,2,..., \text{n}_\text{j} \\ \text{j} = 1,2,...\text{m} \end{cases} \tag{2.7}$$

donde, β_{0j} depende de las unidades o grupos del nivel dos y β_1 es una constante para todos los grupos.

Al sustituir la modelización del coeficiente β_{0j} denotado en (2.5) en la ecuación (2.7), se obtiene Modelo Jerárquico Lineal de dos niveles de Intercepto Aleatorio, el cual viene dado por:

$$Y_{ij} = \beta_0 + \beta_1 X_{ij} + U_{0j} + \varepsilon_{ij} \tag{2.8}$$

De igual manera este modelo se expresa con:

- Una parte fija, la cual es $(\beta_0 + \beta_1 X_{ij})$ justamente la regresión de la respuesta sobre la variable explicativa y representa la curva promedio de todas las unidades del nivel dos.
- Una parte aleatoria, la cual es los errores al nivel uno, ε_{ij} y los errores al nivel dos U_{0j}. La varianza total, la covarianza y el coeficiente de correlación intra-grupo son iguales a las del modelo vacío solo que, ahora, están controlados por valores de la variable explicativas X. Las suposiciones para estos errores son idénticas como en el modelo vacío.

A este modelo (2.8) se le conoce también como Modelo de Componente de Varianza. La generalización del modelo de intercepto aleatorio es inmediata al añadir más variables explicativas del nivel uno. Es decir,

$$Y_{ij} = \beta_0 + \beta_1 X_{1ij} + \beta_2 X_{2ij} + ... + \beta_h X_{hij} + U_{0j} + \varepsilon_{ij} \quad (2.9)$$

reagrupando términos se obtiene,

$$Y_{ij} = \beta_0 + \sum_{h=1}^{p} \beta_h X_{hij} + U_{0j} + \varepsilon_{ij} \quad (2.10)$$

donde $X_1, X_2, ..., X_p$ son "p" variables explicativas del nivel uno; los coeficientes β_h´s con h = 1,2,..., p son constantes grupo a grupo. Este modelo (ecuación 2.10), se puede escribir también como,

$$Y_{ij} = \beta_{0j} + \sum_{h=1}^{p} \beta_h X_{hij} + \varepsilon_{ij} \quad (2.11)$$

Observe que β_{0j} es una variable aleatoria que fluctúa grupo a grupo, con $j = 1, 2, ..., m$

2.2.7 MODELO DE INTERCEPTO Y PENDIENTE ALEATORIAS

El modelo jerárquico lineal de intercepto y pendiente aleatorias también conocido como "modelo de coeficientes aleatorios", se caracteriza porque tanto el intercepto como la pendiente de la variable explicativa son variables aleatorias que fluctúan de grupo a grupo. El submodelo dentro de grupos o modelo del nivel uno viene dado por:

$$Y_{ij} = \beta_{0j} + \beta_{1j} X_{ij} + \varepsilon_{ij} \quad \begin{cases} \mathrm{i} = 1,2,..., \mathrm{n_j} \\ \mathrm{j} = 1,2,...\mathrm{m} \end{cases} \quad (2.12)$$

Tanto el intercepto β_{0j} como la pendiente β_{1j}, son consideradas como variables aleatorias (obsérvese que ambos coeficientes contienen el subíndice j), puesto que dependen del efecto adicional de la estructura anidada y/o del hecho de que las unidades del nivel dos fueron elegidas aleatoriamente; ε_{ij} es la componente aleatoria o error al nivel uno.

La modelización de los coeficientes aleatorios del modelo dentro de grupos, se basa en las características de los grupos a través de variables explicativas del nivel dos. Un modelo de variación simple tanto para el intercepto como para la pendiente se expresan como:

$$\beta_{oj} = \beta_0 + U_{0j} \quad y \quad \beta_{1j} = \beta_1 + U_{1j} \tag{2.13}$$

donde U_{oj} y U_{1j} son variables aleatorias que expresan la contribución de la estructura anidada en β_0 y β_1 respectivamente y son consideradas como las componentes de error al nivel dos; se supone que:

$$\begin{aligned} &E\left(U_{0j}\right) = E\left(U_{1j}\right) = 0 \\ &Var\left(U_{0j}\right) = \sigma^2_{U0} \\ &Var\left(U_{1j}\right) = \sigma^2_{U1} \qquad \text{y} \\ &\text{Cov}\left(\text{U}_{0\text{j}}\,;\,\text{U}_{1\text{j}}\right) = \sigma_{01} \end{aligned} \tag{2.14}$$

Si σ^2_{U0} y σ^2_{U1} son grandes, entonces se encuentran interceptos muy diferentes, al igual que pendientes, para unidades del nivel dos. La covarianza, σ_{01}, entre U_{0j} y $\text{U}_{1\text{j}}$ indica sí la relación es positiva o negativa entre el intercepto y la pendiente. Si σ_{01} es positiva entonces interceptos grandes darán pendientes grandes; si es negativa entonces interceptos grandes se relacionan con pequeñas pendientes.

Al sustituir (2.13), en el submodelo dentro de grupos o submodelo del nivel uno (ecuación 2.12), se obtiene:

$$Y_{ij} = \beta_0 + \beta_1 X_{ij} + U_{1j} X_{ij} + U_{0j} + \varepsilon_{ij} \tag{2.15}$$

Esta ecuación es conocida como modelo jerárquico lineal de dos niveles de intercepto y pendiente aleatorios considerando una variable independiente; es un caso más complejo que los modelos anteriores puesto que ambos coeficientes son variables aleatorias y pueden estar correlacionadas dentro de los grupos. De igual manera la respuesta está compuesta por:

- Una parte fija, $\left(\beta_0 + \beta_1 X_{ij}\right)$ que en este caso es la regresión lineal simple.

- Una parte aleatoria, $\left(U_{1j}X_{ij} + U_{0j} + \varepsilon_{ij}\right)$ compuesta por la suma de los componentes de error de cada nivel.

Como antes, ε_{ij} es la componente aleatoria del error al nivel 1, y se supone son variables aleatorias $IIDN\left(o,\ \sigma_{e}^{2}\right)$; U_{0j} es una componente aleatoria del error al nivel dos que representa las fluctuaciones del intercepto, β_0, grupo a grupo y se supone que son variables aleatorias $IIDN\left(0,\ \sigma_{U0}^{2}\right)$; U_{1j} es la otra componente aleatoria del nivel dos y representa las fluctuaciones del coeficiente de regresión o pendiente, β_1, grupo a grupo, también se supone que son variables aleatorias $IIDN\left(0,\ \sigma_{U1}^{2}\right)$. El término $U_{1j}X_{ij}$ se puede ver como la interacción aleatoria entre las unidades del nivel dos y la variable explicativa al nivel uno, X.

2.2.8 VARIABILIDAD EN EL INTERCEPTO Y LA PENDIENTE

La variabilidad en datos multinivel tiene una estructura más compleja que la variabilidad en datos de un único nivel; la modelación multinivel involucra varias poblaciones, una para cada nivel. En el modelo jerárquico lineal de dos niveles de intercepto y pendiente aleatoria (ecuación 2.15) una parte de la variabilidad de la respuesta, Y, es explicada por la regresión sobre la variable explicativa X; por ejemplo, mediante el término $\beta_1 X_{ij}$; las variables aleatorias U_{0j}, U_{1j} y ε_{ij} representan, cada una, diferentes partes de variabilidad no explicada en la respuesta. Se puede intentar explicar o reducir cada una de ellas. En un primer caso, al nivel uno, añadiendo en la ecuación (2.15) otras variables explicativas a ese nivel con coeficientes fijos o aleatorios, de modo que se produzca una disminución de σ_e^2. Un segundo caso, es intentar explicar parte de la variabilidad al nivel de grupo (nivel 2), añadiendo en el modelo variables explicativas de ese nivel; esas variables variarán de grupo a grupo en el nivel uno. Para ver esto, supóngase que se tienen dos variables explicativas en el nivel dos; Z_1 que puede dar cuenta de la variación en el intercepto y Z_2 que puede dar cuenta de la variación en la pendiente (Z_1 *y* Z_2 podrían ser la misma variable), entonces los submodelos entre grupos (nivel dos) se les dice que son de variación complejas, se pueden escribir como sigue:

$$\beta_{0j} = \lambda_{00} + \lambda_{01} Z_1 + U_{0j} \tag{2.16}$$

$$\beta_{1j} = \lambda_{10} + \lambda_{11} Z_2 + U_{1j} \tag{2.17}$$

sustituyendo ambas (ecuación 2.16 y 2.17) en la ecuación (2.15) se obtiene:

$$Y_{ij} = \lambda_{00} + \lambda_{01} Z_1 + \lambda_{10} X_{ij} + \lambda_{11} Z_2 X_{ij} + U_{1j} X_{ij} + U_{oj} + \varepsilon_{ij} \tag{2.18}$$

Nótese que se ha cambiado la notación de los coeficientes de regresión, por ejemplo $\beta_0 = \lambda_{00}, \beta_1 = \lambda_{10}$. Esto es para facilitar la generalización del modelo.

De ésta ecuación se puede observar que la parte fija aumentó en dos términos. El primero, $\lambda_{01} Z_1$, es visto como el efecto principal de Z_1 en la respuesta, siendo λ_{01} el coeficiente de regresión o pendiente para Z_1 y ; el segundo, $\lambda_{11} Z_2 X_{ij}$, se puede ver como una interacción entre las variables explicativas del nivel 1 y del nivel 2 y se llama término de interacción de productos cruzados.

Por otro lado, la parte aleatoria del modelo (2.18) es igual a la del modelo (2.15) solo que en el modelo con variables explicativas del nivel dos añadidas se espera que la varianza residual tanto del intercepto como de la pendiente disminuya puesto que, parte de ellas es explicada, ahora, por Z_1 *y* Z_2.

2.2.9 MODELO JERÁRQUICO LINEAL GENERAL A DOS NIVELES

El modelo jerárquico lineal de dos niveles puede ser generalizado incluyendo más variables explicativas del nivel uno con coeficientes aleatorios y/o fijos y variables del nivel dos para intentar explicar parte de la variabilidad no explicada en la respuesta.

A continuación se presenta una situación ideal para generalizar el modelo jerárquico lineal de dos niveles. Supóngase una estructura jerárquica de dos niveles con m grupos o unidades del nivel dos y n_j unidades del nivel uno dentro del j-ésimo grupo; el interés está en caracterizar una respuesta Y del nivel uno, en función de $x_1, x_2, \ldots, x_p$ (p variables explicativas del nivel

uno) y $z_1, z_2, ..., z_q$ (q variables del nivel dos). Suponga, además, que de las p variables en el nivel uno: p_1 admiten coeficientes aleatorios en el nivel 1 y en el nivel 2 $(x_1, x_2, ..., x_{p_1})$, p_2 admiten coeficientes aleatorios en el nivel 2 $(x_{p_{1+1}}, x_{p_{1+2}}, ..., x_{p_2})$ y p_3 admiten coeficientes fijos $(x_{p_{2+1}}, x_{p_{2+2}}, ..., x_{p_3})$, sabiendo que:

$p = p_1 + p_2 + p_3$ *donde* $p_1 < p$, $p_2 < p$ y $p_3 < p$ con estas consideraciones, el modelo jerárquico lineal de dos niveles se puede escribir como sigue:

$$y_{ij} = \beta_{0ij}x_{0ij} + \beta_{1ij}x_{1ij} + ... + \beta_{p_{1ij}}x_{p_{1ij}} + \beta_{p_{1+1j}}x_{p_{1+1ij}} + \beta_{p_{1+2j}}x_{p_{1+2ij}} + ... + \beta_{p_{2j}}x_{p_{2ij}} + \beta_{p_{2+1}}x_{p_{2+1ij}} + \beta_{p_{2+2}}x_{p_{2+2ij}} + ... + \beta_{p_3}x_{p_{3ij}} \quad (2.19)$$

Nótese que las variables que admiten coeficientes aleatorios en los dos niveles sus coeficientes tienen subíndice "i" y "j" (β_{1ij}); los que admiten coeficientes aleatorios solo en el nivel dos tienen el subíndice "j" $(\beta_{p_{1+1j}})$ y los de coeficientes fijos no tienen ningún subíndice salvo el que indica a la variable.

Aplicando sumatoria,

$$y_{ij} = \sum_{h=0}^{p_1} \beta_{hij}x_{hij} + \sum_{h=p_{1+1}}^{p_2} \beta_{hj}x_{hij} + \sum_{h=p_{2+1}}^{p_3} \beta_h x_{hij} \; ; con \; x_{0ij} = 1 \quad (2.20)$$

donde $h = 0, 1, ..., p$. Las modelizaciones para los coeficientes β_{hij} y β_{hj}, vienen dadas por

$$\beta_{hij} = \beta_{hj} + \varepsilon_{hij} \quad \begin{cases} i = 1,2,..., n_j \\ j = 1,2,..., m \end{cases} \quad (2.21)$$

$$\beta_{hj} = \lambda_{h0}Z_0 + \lambda_{h1}Z_1 + ... + \lambda_{hq}Z_q + U_{hj} \quad (2.22)$$

Reagrupando términos para la modelizacion β_{hj} , se obtiene

$$\beta_{hj} = \sum_{s=0}^{q} \lambda_{hs} Z_s + U_{hj} \; ; \; Z_0 = 1 \tag{2.23}$$

Sustituyendo primero la modelizacion β_{hij} , en la ecuación 1.20

$$y_{ij} = \sum_{h=0}^{p_1} \beta_{hj} x_{hij} + \sum_{h=0}^{p_1} \varepsilon_{hij} x_{hij} + \sum_{h=p_{1+1}}^{p_2} \beta_{hj} x_{hij} + \sum_{h=p_{2+1}}^{p_3} \beta_h x_{hij} \tag{2.24}$$

Reagrupando la primera y tercera sumatoria de la ecuación 2.24, se tiene

$$y_{ij} = \sum_{h=0}^{p_2} \beta_{hj} x_{hij} + \sum_{h=0}^{p_1} \varepsilon_{hij} x_{hij} + \sum_{h=p_{2+1}}^{p_3} \beta_h x_{hij} \tag{2.25}$$

Ahora sustituyendo la modelizacion β_{hj} , en la ecuación 2.25

$$y_{ij} = \sum_{h=0}^{p_2} \sum_{s=0}^{q} \lambda_{hs} Z_s x_{hij} + \sum_{h=0}^{p_2} U_{hj} x_{hij} + \sum_{h=0}^{p_1} \varepsilon_{hij} x_{hij} + \sum_{h=p_{2+1}}^{p_3} \beta_h x_{hij} \tag{2.26}$$

Ordenando se tiene el modelo jerárquico lineal general a dos niveles,

$$y_{ij} = \sum_{h=0}^{p_2} \sum_{s=0}^{q} \lambda_{hs} Z_s x_{hij} + \sum_{h=p_{2+1}}^{p_3} \beta_h x_{hij} + \sum_{h=0}^{p_2} U_{hj} x_{hij} + \sum_{h=0}^{p_1} \varepsilon_{hij} x_{hij} \tag{2.27}$$

donde,

- $\sum_{h=0}^{p_2} \sum_{s=0}^{q} \lambda_{hs} Z_s x_{hij} + \sum_{h=p_{2+1}}^{p_3} \beta_h x_{hij}$ representa la curva promedio sobre todos los grupos.

- $\sum_{h=0}^{p_2} U_{hj} x_{hij}$ son las fluctuaciones o efectos de interacción de los grupos y las variables del nivel uno.

- $\sum_{h=0}^{p_1} \varepsilon_{hij} x_{hij}$ son las fluctuaciones o efectos individuales de x en cada unidad dentro del grupo. También describe el comportamiento de la variación al nivel uno en función de las variables explicativas.

En forma matricial

$$y_{ij} = \left(x_{0ij},...x_{p_2ij}\right)\begin{pmatrix} \gamma_{00} & \gamma_{01} & \cdots & \gamma_{0q} \\ \gamma_{10} & \gamma_{11} & \cdots & \gamma_{1q} \\ & & \cdots & \\ \gamma_{p_2 0} & \gamma_{p_2 1} & \cdots & \gamma_{p_2 q} \end{pmatrix}\begin{pmatrix} z_0 \\ z_1 \\ \vdots \\ z_q \end{pmatrix} + \left(x_{p_{2+1}ij},..., x_{p_3ij}\right)\begin{pmatrix} \gamma_{p_{2+1}} \\ \gamma_{p_{2+2}} \\ \vdots \\ \gamma_{p_3} \end{pmatrix}$$

$$y_{ij} = \left(x_{0ij},...x_{p_2ij}\right)\begin{pmatrix} U_{0j} \\ U_{1j} \\ \vdots \\ U_{p_2 j} \end{pmatrix} + \left(x_{0ij},..., x_{p_1ij}\right)\begin{pmatrix} \varepsilon_{0ij} \\ \varepsilon_{1ij} \\ \vdots \\ \varepsilon_{p_1ij} \end{pmatrix}$$

$$y_{ij} = \left(U_{0j},...U_{p_2 j}\right)^{\mathrm{T}} \approx N_{p_2}\left(\mathrm{O}, \Omega_U\right)$$

Es necesario mencionar que a partir de la ecuación 2.27, se pueden obtener los modelos multinivel ya descritos; un ejemplo para el caso del modelo vacío es el siguiente: suponga que se está interesado en estudiar el rendimiento escolar de los alumnos de un conjunto de escuelas. Este ejemplo está conformado por una estructura jerárquica de dos niveles; donde el nivel uno, son los estudiantes; el nivel dos son las escuelas y la variable dependiente es el rendimiento escolar. Al no tener variable(s) independiente(s), se tiene que

$$p = 0, \quad p_1 = 0, \quad p_2 = 0, \quad p_3 = 0 \quad y \quad q = 0$$

Sustituyendo estos valores en la ecuación 2.27,

$$y_{ij} = \sum_{h=0}^{0}\sum_{s=0}^{0} \lambda_{hs} Z_s x_{hij} + \sum_{h=_{0+1}}^{0} \beta_h x_{hij} + \sum_{h=0}^{0} U_{hj} x_{hij} + \sum_{h=0}^{0} \varepsilon_{hij} x_{hij} \qquad (2.28)$$

Por propiedad de sumatoria y dando valores a cada término,

$$y_{ij} = \lambda_{00} Z_0 x_{0ij} + U_{0j} x_{0ij} + \varepsilon_{0ij} x_{0ij} \qquad (2.29)$$

En este caso las variables al nivel uno y dos son constantes, es decir, $Z_0 = 1$ *y* $x_{0ij} = 1$ dando como resultado el modelo vacío

$$y_{ij} = \lambda_{00} + U_{0j} + \varepsilon_{0ij} \qquad (2.30)$$

Para obtener el modelo de intercepto aleatorio es necesario definir las variables explicativas de la siguiente manera: $p = p_3 = 1, \quad p_1 = p_2 = 0 \quad y \quad q = 0$, es decir, existe una única variable explicativa del nivel uno, que admite coeficiente fijo, donde al igual que en el caso anterior $Z_0 = 1$ y la variable $x_{0ij} = 1$. Si las variables explicativas son definidas tal como $p = p_2 = 1, \quad p_3 = p_1 = 0 \quad y \quad q = 0$, entonces se estará en presencia el modelo de intercepto y pendientes aleatorias.

2.2.10 SUPUESTOS DEL MODELO JERÁRQUICO LINEAL

Los supuestos del modelo jerárquico lineal de dos niveles son los siguientes:

- Los $\varepsilon_{sij}; i = 1,2,..., n_j; \quad j = 1,2,..., m$ son variables aleatorias IID $\left(0, \sigma_e^2\right)$.
- Los $\varepsilon_{0ij}, \varepsilon_{1ij}, ..., \varepsilon_{p_1 ij}$ son variables aleatorias que pueden estar relacionadas en el i-th sujeto dentro del j-th grupo.
- Existe independencia entre los errores de los distintos niveles.

- Los errores al nivel dos tienen distribución normal Multivariante con vector de media cero y matriz de covarianza constante.

2.2.11 RESIDUOS

Se sabe que el modelo lineal jerárquico (MLJ) tiene más de un término de error, al menos un término para cada nivel de la estructura jerárquica. Por ejemplo, el MLJ, para una estructura jerárquica de dos niveles, con intercepto aleatorio únicamente (ecuación 1.8), tiene dos componentes de error o residuos: los residuos al nivel uno, ε_{ij} , y los residuos al nivel dos U_{0j} . Estos se pueden predecir individualmente cuando se tienen las observaciones y los parámetros estimados del modelo. La deducción explícita de los residuos del MLJ de intercepto aleatorio (ecuación 2.8) es la siguiente. Sea y_{ij} el valor observado en la i-ésima unidad del nivel uno agrupada en la j-ésima unidad (grupo) del nivel dos y sea $\hat{y}_{ij}$ la predicción correspondiente de ese valor observado, mediante la línea de regresión ajustada al j-ésimo grupo. Entonces el residuo bruto viene dado por:

$$r_{ij} = y_{ij} - \hat{y}_{ij} \tag{2.31}$$

Además, el residuo bruto para el j-ésimo grupo (la unidad del nivel dos a la que se le ajustó la línea de regresión) denotado r_{+j}, es justamente la media de los residuos brutos anteriores sobre todas las n_j unidades del nivel uno dentro de ese j-ésimo grupo; es decir:

$$r_{+j} = \frac{\sum_{\mathrm{i}=1}^{n_j} r_{ij}}{n_j} \qquad \mathrm{j} = 1,\ldots,\mathrm{m} \tag{2.32}$$

En consecuencia, el residuo estimado para ese j-ésimo grupo se obtiene multiplicando su residuo bruto por un factor de contracción. Es decir:

$$\hat{U}_{0j} = \frac{n_j \sigma_{U0}^2}{n_j \sigma_{U0}^2 + \sigma_e^2} r_{+j} \qquad \mathrm{j} = 1,\ldots,\mathrm{m} \tag{2.33}$$

El residuo estimado para la observación y_{ij} viene dado por:

$$\begin{aligned} \hat{e}_{ij} &= r_{ij} - \hat{U}_{0j} \\ &= y_{ij} - \hat{y}_{ij} - \hat{U}_{0j} \end{aligned} \tag{2.34}$$

Estos residuos no son siempre insesgados pero sí consistentes. Goldstein (1995) presenta una generalización para la obtención de residuos multinivel en un modelo lineal jerárquico bajo una estructura de multinivel general de "s" niveles. La mayoría de los software multinivel como HLM (Bryk et. al., 1996), MLwiN (Rasbash et al., 2001), MIXREG (Hedeker y Gibbons, 1996) permiten calcular las componentes de residuos de los diferentes niveles de la jerarquía.

La importancia de los residuos multinivel, al igual que en los métodos de regresión múltiple, consiste en que la inspección de éstos mediante algún procedimiento permitirá detectar casos inusitados que pudieran tener una influencia indebida en los resultados del análisis estadístico y, también, verificar los supuestos de normalidad. El cálculo de los residuos del modelo lineal jerárquico general de dos niveles sigue el mismo procedimiento descrito anteriormente.

2.3 CONCEPTOS BASICOS DE VALORES ATIPICOS

2.3.1 INTRODUCCIÓN.

Supongamos que estamos analizando un conjunto de datos en base a un modelo existente. Podríamos encontrar que los parámetros adecuados del modelo dependen excepcionalmente de uno o más puntos o unidades de datos individuales. En este capítulo, discutiremos como identificar dichos puntos, como considerarlos a la hora de analizar los datos y como utilizarlos para esclarecer un poco la estructura subyacente de los datos y del fenómeno en investigación.

La categoría más obvia dentro de lo que se considera un punto de datos especial son los valores atípicos o puntos atípicos. Este es un punto de datos que pareciera estar fuera de orden con respecto al resto de los datos, en otras palabras es inconsistente con el modelo que encaja con el resto de los datos.

2.3.2 Valores Atípicos

La literatura sobre el estudio de valores atípicos nos permite fijar la idea sobre estos. Se presentaran tres conceptos que, consideramos, describen lo que se entiende por un valor extraño. Estos conceptos son:

i) Son aquellos valores del conjunto de datos, que se encuentran alejados del comportamiento general del resto de los datos, según Mongomery (1981).

ii) Observaciones que no pueden ser consideradas, totalmente, como una manifestación del fenómeno bajo estudio, presentado por R.J. Beckman y R.D. Cook (1983).

iii) Observaciones que proveen residuos grande, en valor absoluto, con respecto a los demás datos, después que el modelo asido ajustado, presentado por N.R. Draper y J.A. John (1978).

La siguiente figura permite ilustrar con mayor facilidad, la idea de valor atípico, definida anteriormente (i, ii y iii).

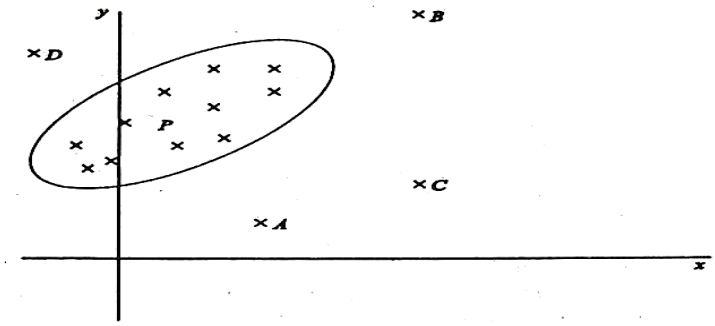


Figura 2.1 Ilustración de valores atípicos

Existen muchas definiciones de los valores discrepantes (atípicos, distantes, aberrantes, etc.); en la literatura inglesa se les denomina “outliers”. Algunas son las siguientes:

Barnety y Lewis (1994), consideran un valor discrepante como “una observación o conjunto de observaciones que parecen ser inconsistentes con el resto del conjunto de dato”.

Peña D. (2002), considera los valores atípicos como aquellas observaciones que parecen haberse generado de forma distinta al resto del conjunto de datos. Pueden ser causadas por errores de medición o una heterogeneidad intrínseca de los elementos observados.

Estas definiciones conducen a visualizar la matriz de datos como una nube de puntos en el espacio p-dimensional; $p = 1,2, \dots, k$; los valores alejados del grueso general de los datos serán los discrepantes. Entre las causas posibles que pueden ocasionar valores discrepantes, en la generación de datos, se destacan: (1) las debidas a una variación natural (característica extraña propia de la observación en la población bajo estudio); un valor discrepante de este tipo podría resultar de gran interés, que surge como consecuencia de una variación imprevista o condicionada de cambio natural de las partes del fenómeno, resultando ser pieza vital para el entendimiento de dicho fenómeno y, (2) las debidas a hechos extremos a los mismos, como por ejemplo, errores en el registro de la respuesta, ya sea en el momento en el que se recoge información o en la transcripción; ciertamente en una situación experimental donde se debe medir, observar, etc., se está condicionado a imperfecciones en los instrumentos de medida, errores en la medición, etc., recolectando así un conjunto de datos incorrectos y originando posibles valores extraños que se reflejaran luego, en el ajuste del modelo.

La presencia de los valores discrepantes tiene efectos en el análisis de los datos. Por ejemplo, si se hacen estadísticas descriptivas, la media y la varianza estarán afectadas; cuando se ajusta un modelo de regresión las estimaciones de los parámetros estarán afectadas por estos valores en cuanto a que aumenta la probabilidad de obtener estimaciones erróneas (errores atípicos y estimaciones de parámetros con valores inesperados) y un mal ajuste del modelo global. De aquí la gran preocupación en el análisis de datos con la posible presencia de valores discrepantes puesto que puede conducir a conclusiones verdaderamente reñidas con la realidad.

En el conjunto de datos puede haber uno o más valores discrepantes. Si el conjunto de datos es de un único nivel, por ejemplo mediciones en alumnos, se habla de datos discrepantes

clásicos; si los datos son multinivel, por ejemplo mediciones en alumnos dentro de escuelas, se habla de datos atípicos multinivel.

2.3.3 Rechazo de Valores Atípicos

Una vez detectados los valores extraños pueden ser rechazados o no, dependiendo de la categoría en que se encuentren. Por ejemplo será oportuno y necesario rechazar tales observaciones sólo si existe evidencia comprobada de que éstas son producto del error causado por mal funcionamiento de los instrumentos de medición o error del analista, etc., es decir, existe un debilitamiento en el conjunto de datos. También se puede rechazar un subconjunto de $1, 2, \dots, k$ observaciones sólo si ese subconjunto esta "significativamente" desviado del resto de las observaciones, haciendo salvedad que el modelo resultante se ajusta a un rango reducido de datos. Cuando el valor extraño proviene de un debilitamiento del modelo puede fácilmente ser corregido con una adecuada revisión del modelo o aplicar una transformación. Si el valor extraño proviene de la variación natural, no debe rechazarse sino más bien resaltarse para tomarlo en cuenta de manera especial.

2.3.4 Detección de Valores Atípicos en Regresión Lineal

En muchas áreas de investigación es ampliamente aplicada la regresión lineal múltiple; razón por la cual es de particular importancia detectar los valores discrepantes y mitigar su efecto en la estimación del modelo. Se debe recordar que el análisis de regresión consiste en ajustar un modelo, de algún tipo, al conjunto de datos, obtenidos de alguna experimentación en estudio, con el propósito de obtener una ecuación empírica de predicción razonablemente precisa y que describa el comportamiento de la variable respuesta, dados los valores de las variables predictoras o explicativas.

Explícitamente, supóngase que se observa una muestra tamaño "n" de una respuesta, Y, sometida al efecto de p-variables explicativas $X_1,\ X_2, \dots, X_p$; los datos se registran mediante

$$\boldsymbol{Y} = \begin{pmatrix} Y_1 \\ Y_2 \\ \vdots \\ Y_n \end{pmatrix}; \boldsymbol{X_D} = \begin{pmatrix} X_{11} & \dots & X_{1p} \\ X_{21} & \dots & X_{2p} \\ \vdots & \vdots & \vdots \\ X_{n1} & \dots & X_{np} \end{pmatrix} \tag{2.35}$$

El ajuste, a los datos, de un modelo de regresión lineal múltiple de la forma:

$$\boldsymbol{Y} = \boldsymbol{X\beta} + \in, \tag{2.36}$$

con $\boldsymbol{\beta} = \begin{pmatrix} \beta_0 \\ \vdots \\ \beta_p \end{pmatrix}$, $\boldsymbol{X} = \begin{pmatrix} 1 & X_{11} & \dots & X_{1p} \\ & X_{21} & \dots & X_{2p} \\ \vdots & \vdots & \vdots & \vdots \\ 1 & X_{n1} & \dots & X_{np} \end{pmatrix}$ y $\in = \begin{pmatrix} \in_1 \\ \in_2 \\ \vdots \\ \in_n \end{pmatrix}$, (2.37)

permite obtener una ecuación de regresión estimada de la forma:

$$\widehat{\boldsymbol{Y}} = \widehat{\boldsymbol{\beta}}\boldsymbol{X} = \boldsymbol{HY}, \tag{2.38}$$

donde

$$\widehat{\boldsymbol{\beta}} = (\boldsymbol{X'X})^{-1}\boldsymbol{X'Y} \tag{2.39}$$

$$\boldsymbol{H} = \boldsymbol{X}(\boldsymbol{X'X})^{-1}\boldsymbol{X'}, \tag{2.40}$$

$\boldsymbol{H}$ es la matriz que permite proyectar ortogonalmente el vector $\boldsymbol{Y}$ sobre el espacio generado por las variables explicativas, descritas sobre la columnas de la matriz $\boldsymbol{X}$.

Los errores (residuos) de estimación de las observaciones vienen dados mediante:

$$\hat{e}_i = y_i - \hat{y}_i; i = 1, \dots, n \tag{2.41}$$

Para la detección de outliers en regresión lineal se puede recurrir a uno o varios de los diferentes procedimientos que existen; los residuos son muy útiles en la detección de outliers. Entre los procedimientos de uso frecuente en la detección de valores discrepantes en regresión lineal podemos mencionar, el diagnostico por casos, la matriz Hat, la Distancia de

Mahalanobis, los DFITS, la Distancia de Cook, gráfica de los residuos $\hat{\varepsilon}_i$ frente a $\hat{y}_i$, entre otros.

2.4 Estadístico de Andrews y Pregibon (A.P)

En este capítulo se describen los aspectos teóricos del estadístico de Andrews y Pregibon (A.P) y se realiza una generalización de éste para la detección de valores atípicos multinivel bajo la estructura de un modelo multinivel de dos niveles con intercepto y una pendiente aleatorios.

David F. Andrews y Darly Pregibon en 1978 proponen un estadístico para examinar, en un modelo lineal, la presencia de una o más observaciones discrepantes importantes. Para la definición del estadístico ellos consideran el modelo lineal usual:

$$Y = X\beta + \varepsilon \tag{2.42}$$

donde X es la matriz de diseño β el vector de coeficientes de regresión

Antes de presentar la definición del estadístico de A.P, se definen unos operadores que facilitan la exposición.

Sea $X^* = (X : Y)$ la matriz ampliada del modelo (2.42).

2.4.1 Operador de Supresión

Se define $D_{ij\ldots}^{(k)}(\cdot)$ como un operador que suprime k filas, indicadas por los subíndices i, j, …, de X^*, X ó Y previamente a efectuar todas las operaciones indicadas en $(\cdot)$. Por ejemplo $D_i^1 Y$ es el vector $(n-1)\times 1$ de variables independientes con la i-ésima observación suprimida. Explícitamente, si

$Y = \begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{bmatrix}$, entonces $\quad D_i^1 Y = \begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_{i-1} \\ y_{i+1} \\ \vdots \\ y_n \end{bmatrix}$

$n \times 1$ $(n-1) \times 1$

k es un número natural escogido por el analista. Por ejemplo, con k = 2 podemos escribir:

$$\left[D_{3,8}^{(2)}(X^T X)\right]^{-1} \tag{2.43}$$

que es la matriz inversa del producto interno de las matrices X^T y X efectuado después de suprimir las filas terceras y octava en X.

2.4.2 Operador completo.

$\overline{D}_{ij\ldots}^{(k)}(\cdot)$ se define como el operador completo que selecciona k- filas (i, j, …) en lugar de suprimirlas.

El estadístico de A.P. se basa fundamentalmente en el determinante $|X^T X|$ que es muy útil puesto que suministra información sobre el modelo. Por ejemplo, si la supresión de una observación en X, tiene un gran o pequeño efecto en $|X^T X|$, la observación eliminada tiene un gran o una pequeña influencia en las estimaciones resultantes.

Similarmente, entre las observaciones que más se desvían de los valores ajustados, mientras que un mayor número de estas se supriman, más se reducirá la suma de cuadrado dada por:

$$RSS = \left(Y - X_\beta\right)^T \left(Y - X_\beta\right) \tag{2.44}$$

Estas dos ideas fueron combinadas por Andrews y Pregibon para calcular el cambio en el producto $RSS \times |X^T X|^T$ al suprimir una o más observaciones. Esta forma particular es conveniente estudiarla como un determinante, ya que

$$|X^{*T} X^*| = RSS \times |X^T X|^T \tag{2.45}$$

La demostración de (2.45) se puede ver en Fermín (1993).

2.4.3 Definición (estadístico A.P)

El cambio relativo debido a la supresión de k filas (i, j, …) asociada con los términos del modelo dado en (2.1), está determinado por el cociente.

$$R_{ij\ldots}^{(k)}(X^*) = \frac{D_{ij\ldots}^{k}|X^{*T}X^*|}{|X^{*T}X^*|} \tag{2.46}$$

que es precisamente el Estadístico de Andrews y Pregibon. Es una cantidad sin unidades de medida. Geométricamente $1 - \left\{R_{ij\ldots}^{(k)}(X^*)\right\}^{1/2}$ está asociada a la proporción del volumen generado por X^* atribuible a las k observaciones (i,j, …) suprimidas. Si este subconjunto de observaciones está lejos en el espacio del conjunto de todos los datos $1 - \left\{R_{ij\ldots}^{(k)}(X^*)\right\}^{1/2}$ corresponderán a una gran proporción del volumen del espacio, dando así una interpretación realista del valor atípico. De aquí, pequeños valores de $R_{ij\ldots}^{(k)}(X^*)$ son asociados con valores extraños influyentes con los parámetros estimados.

2.5 Métodos de Detección de Valores Atípicos Multinivel

Como se ha dicho existen varios procedimientos para detectar outliers multinivel, después de ajustar un modelo, mediante el análisis de residuos de los diferentes niveles de la estructura jerárquica. Entre los procedimientos jerárquicos más frecuentes están: la gráfica de los residuos puros, residuos estandarizados y supresión residual; también existen otros procedimientos más analíticos como los valores de influencia y los valores DFITS.

Para tener una visión más amplia de estos procedimientos veamos, bajo la perspectiva de un modelo multinivel de dos niveles con intercepto y una pendiente aleatorios, una descripción de éstos. Considérese una estructura jerárquica de dos niveles con m grupos o unidades del nivel dos y n_j unidades del nivel uno dentro del j-ésimo grupo; el interés está en caracterizar una respuesta, digamos *Y*, del nivel uno, en función de $X_1, X_2, \ldots, X_p$ (p

variables explicativas del nivel uno), donde X_1 admite coeficiente aleatorio en el nivel II y las demás con coeficientes fijos. Con estas consideraciones, el modelo jerárquico lineal de dos niveles de intercepto y una pendiente aleatoria se puede escribir como sigue:

$$Y_{ij} = \beta_{0ij} + \beta_{1j}X_1 + \sum_{i=2}^{p} \beta_i X_i \tag{2.47}$$

$$Y_{ij} = \beta_0 + \sum_{i=1}^{p} \beta_i X_i + \mu_{0j} + \mu_{1j}X_1 + e_{0ij} \tag{2.48}$$

Este último modelo 2.23 se le llama modelo de intercepto y pendiente aleatorios (una pendiente aleatoria); es el modelo que se usará para hacer la generalización de la distancia de Cook.

El modelo estimado para 2.23 viene dado por:

$$\hat{Y}_{ij} = \sum \hat{\beta}_0 \, x_{sij}, \tag{2.49}$$

donde

$$\hat{\beta} = \begin{bmatrix} \hat{\beta}_0 \\ \hat{\beta}_1 \\ \vdots \\ \hat{\beta}_p \end{bmatrix} = (X'V^{-1}X^{-1})X'V^{-1}Y \tag{2.50}$$

La estimación del (i, j) residuo puro del nivel I viene dado por:

$$\hat{e}_{oij} = r_{ij} - \widehat{U}_{oj} - \widehat{U}_{1j}, \qquad i = 1, \dots, n_j; j = 1, \dots, m. \tag{2.51}$$

donde $r_{ij} = y_{ij} - \hat{y}_{ij}$, $r_{+j} = \sum_{i=1}^{n_j} r_{ij}$; además, los residuos del nivel II son:

$$\hat{U}_{0j} = \frac{n_j \sigma_{U0}^2}{n_j \sigma_{U0}^2 + \sigma_e^2} r_{+j} \qquad \text{j} = 1,\dots, \text{m.} \qquad \text{y,} \tag{2.52}$$

$$\hat{U}_{1j} = \frac{n_j \sigma^2_{U1}}{n_j \sigma^2_{U1} + \sigma^2_e} r_{+j} \qquad j = 1,..., m. \tag{2.53}$$

Los residuos estandarizados se utilizan para diagnosticar unidades del nivel dos distantes u outliers. Estos residuos, en el nivel dos, vienen dados por:

$\widehat{U}'_{sj} = \frac{\widehat{U}_{sj}}{s_{us}}$, s=0, 1; siendo s_{us}el error estándar de $\widehat{U}_{sj}$.

Mientras más grande, en valor absoluto, sea el $\widehat{U}'_{sj}$ más sospechosa es la j-ésima unidad del nivel dos de ser un outliers con respecto al coeficiente aleatorio β_{sj}, $(s = 0,1)$. De hecho, Rasbash y otros (2001) establecen como procedimiento de detección de valores discrepantes un histograma de los residuos estandarizados y establecen que los residuos estandarizados menores que -2 y mayores que 2 indican que el residuo es significativamente diferente de la media de los residuos; es decir, es un residuo distante. Sin embargo, Leyland y Goldstein (2001) establecen valores fuera del rango (-3, 3) para considerar el residuo distante.

Como una crítica objetiva que hacemos a este procedimiento es que se ve más empírico y por tanto no es un planteamiento objetivo.

Por otra parte, Leyland y Goldstein (2001) establecen que los residuales de supresión permiten detectar grupos distantes. Estos autores definen los residuos de supresión mediante la siguiente expresión:

$$r'_{cj(j)} = \left(\frac{\hat{U}_{cj} - \hat{U}_{j(j)}}{s_{cj(j)}}\right), \tag{2.54}$$

$\boldsymbol{r'_{cj(j)}}$ representa el residuo de supresión de la l-ésima unidad del nivel dos correspondiente al coeficiente aleatorio β_{cj}, con c= 0, 1 y, donde $\widehat{U}_{cj}$, $\widehat{U}_{j(j)}$ y $s_{cj(j)}$ son el residuo de la respuesta estimada para la j-ésima unidad, el residuo para la respuesta estimada omitiendo la unidad sospechosa de ser outlier y el error estándar cuando se omite la unidad sospechosa de ser outlier.

De igual manera se construye el histograma de los residuos de supresión. Un valor de $\boldsymbol{r'_{cj(j)}}$ fuera del rango (-3,3) permite considerar la observación examinada como un valor atípico (Ver Leyland y Goldstein, 2001).

Los valores de apalancamiento (leverage) también se usan para diagnosticar valores discrepantes. El leverage de una unidad del nivel dos (un grupo particular) para un coeficiente aleatorio, por ejemplo el intercepto, indica que algún cambio en el intercepto, para ese grupo, tiende a mover la superficie de regresión apreciablemente hacia ese valor nuevo del intercepto. Los valores de leverage se calculan usando los elementos de la diagonal de la matriz hot, denotada ***H.*** Los términos de la diagonal de la matriz ***H*** vienen dados por***:***

$$h_{ll} = x_l(X'X)^{-1}x'_l, \tag{2.55}$$

siendo $x'_l = (x_{l1}, x_{l2}, \dots, x_{lp}), la$ l-ésima fila de $\boldsymbol{X}$.

Para el caso multinivel, el leverage para el l-ésimo grupo viene dado por:

$$\mathrm{h'_{ll}} = \mathrm{h_{ll}}/\mathrm{nca}, \tag{2.56}$$

donde nca representa el número de coeficientes aleatorios en el nivel dos.

De acuerdo con Leyland y Goldstein (2001) un leverage alto, es un indicativo de dos cuestiones: (1) que el grupo es influyente para la estimación de los parámetros del modelo a causa de el tamaño del grupo y lo disperso que están los valores de las variables explicativas

en ese grupo con respecto a los demás y (2) al grado en el cual el grupo se ajusta al modelo definido por los otros grupos, teniendo relación directa con los residuos de ese grupo. Se considera un valor h'_{ll} grande si supera a cociente $2p/m$, donde p es el número de coeficientes aleatorios y m el número de grupos. Un leverage grande, para una unidad del nivel dos, indica que la inclusión de esa unidad en el análisis producirá un incremento sustancial en la precisión de las estimaciones de los parámetros del modelo.

Rasbash et. al. (2001) hacen un histograma de los valores de leverage para cada grupo; esto permite visualizar cual grupo tiene mayor leverage para considerar un mayor análisis en éste.

Otra medida que se usa, en el ámbito multinivel, para diagnosticar valores atípicos, influyentes u outliers son los valores de influencia, denotados por DFITS. Los valores de influencia combinan los residuos y los leverages para medir el impacto de cada grupo sobre los coeficientes estimados; es decir, miden el grado de influencia que tiene, en la estimación de los parámetros del modelo, la inclusión en el conjunto de datos un grupo particular. El cálculo de los DFITS está relacionado al efecto sobre la predicción del grupo en análisis causado por la omisión de éste. Explícitamente:

$$(\text{DFITS})_j = \frac{\hat{Y}_j - \hat{Y}_{j(j)}}{s_{(j)}\sqrt{h_{jj}}}, \qquad (2.57)$$

donde $\hat{Y}_j$, $\hat{Y}_{j(j)}$ y $s_{(l)}$ son: la predicción de la respuesta para el j- ésimo grupo considerando todos los datos, la predicción de la respuesta para el j-ésimo grupo considerando todos los grupo excepto el j-ésimo y la desviación estándar de todos los datos excluyendo el j-ésimo grupo. Rasbash et. al. (2001) construyen un histograma para visualizar la magnitud de los DFITS.

Para valores de DFITS mayores que $2{,}5\sqrt{(p/m)}$, se consideran grandes e indican una elevada influencia de la unidad excluida y, por ende, debe investigarse el grupo con más detalles.

El software MLwiN 1.1 permite obtener histogramas para los residuos puros, estudentizados, valores de influencia y DFITS.

2.6 Generalización del Estadístico de Andrews y Pregibon para datos Multinivel

Antes de presentar la definición del estadístico de A.P para datos multinivel, se definen unos operadores que facilitan la exposición.

Sea $X^{j^*} = (X^* \vdots Y)$ la matriz ampliada del modelo multinivel, donde X^* representa la matriz de diseño

$$X^* = \begin{pmatrix} x^{1^*} \\ x^{2^*} \\ \vdots \\ x^{m^*} \end{pmatrix} \qquad X^{j^*} = \begin{pmatrix} 1\, x_{11}^j & \cdots & x_{1\,p}^j\, y_1^j \\ \vdots\ \vdots & \ddots & \vdots\ \vdots \\ 1 x_{nj\,1}^j & \cdots & x_{nj\,p}^j y_{nj}^j \end{pmatrix}$$

$$(N \times p + 1)$$

con $N = \sum_{j=1}^{m} n_j$, cantidad de unidades en el nivel 1 y m la cantidad de unidades en el nivel 2.

2.6.1 Operador de Supresión multinivel.

Se define $D_{i\ldots}^{(j)}(\cdot)$ como un operador que suprime k filas en la unidad j-ésima del nivel 2 , indicadas por los subíndices i, j, …, de X^* . es decir:

Sea $X^{j^*} = \begin{pmatrix} 1\, x_{11}^j & \cdots & x_{1\,p}^j\, y_1^j \\ \vdots\ \vdots & \ddots & \vdots\ \vdots \\ 1 x_{nj\,1}^j & \cdots & x_{nj\,p}^j y_{nj}^j \end{pmatrix}$ entonces, $D_{i\ldots}^{(j)}(X^{j^*}) = \begin{pmatrix} 1 & x_{11}^j & \cdots & y_1^j \\ \vdots & \vdots & \ddots & \vdots \\ 1 & X_{(i-1)1}^j & \ddots & Y_{(i-1)1}^j \\ 1 & X_{(i+1)1}^j & \ddots & Y_{(i+1)1}^j \\ \vdots & \vdots & & \vdots \\ 1 & x_{nj\,1}^j & \cdots & y_{nj}^j \end{pmatrix}$

Se define $D^j(\cdot)$ como un operador que suprime la unidad j-ésima del nivel 2 , indicadas por el subíndice j de X^* , es decir:

Sea $X^* = \begin{pmatrix} x^{1^*} \\ x^{2^*} \\ \vdots \\ x^{m^*} \end{pmatrix}$ entoces, $D^j(X^*) = \begin{pmatrix} x^{1^*} \\ \vdots \\ x^{(j-1)^*} \\ x^{(j+1)^*} \\ \vdots \\ x^{m^*} \end{pmatrix}$

2.6.2 Estadístico de Andrews y Pregibon para datos Multinivel.

El cambio relativo debido a la supresión de k filas (i, j, …) asociada con los términos del modelo multinivel, está determinado por el cociente.

$$R_i^j(X^{j*}) = \frac{D_i^j \left| X_i^{j*T} X_i^{j*} \right|}{\left| X^{j*T} X^{j*} \right|}$$

$R_i^j(X^*)$ es el estadístico de Andrews y Pregibon que se obtiene al suprimir la unida i-esima dentro del nivel j-esimo

$$R^j(X^*) = \frac{D^j \left| X^{j*T} X^{J*} \right|}{|X^{*T} X^*|}$$

$R^j(X^*)$ es el estadístico de Andrews y Pregibon que se obtiene al suprimir la unida j-esimo del nivel 2.

CAPÍTULO III

3.1.- NIVEL DE LA INVESTIGACIÓN

Según Fidas (2006) la investigación explicativa es aquella que se encarga de buscar el por qué de los hechos mediante el establecimiento de relaciones causa-efecto. En este sentido, los estudios explicativos pueden ocuparse tanto de la determinación de la causas (investigación postfacto), como de los efectos (investigación experimental), mediante la prueba de hipótesis. Sus resultados y conclusiones constituyen el nivel más profundo de conocimientos. En efecto esta investigación es de tipo explicativa puesto que tiene como fin deducir una generalización del estadístico A:P utilizado en el análisis de regresión, basándose en las teorías y estadísticos existentes para la detección de valores atípicos, tanto en el contexto multinivel como en el caso de un único nivel.

3.2.- DISEÑO DE LA INVESTIGACIÓN

Según Fidas (2006) la Investigación Documental es aquella que se basa en la obtención y análisis de datos provenientes de materiales impresos u otros tipos de documentos. Sobre la base de esto la clasificación del diseño de esta investigación es de tipo documental puesto que para darle respuesta a si es posible adaptar el estadístico de A:P como método de detección de valores discrepantes multinivel, se apoya principalmente, en trabajos previos, información y datos divulgados por medios impresos y electrónicos.

3.3.- MATERIALES Y MÉTODOS

La generalización del estadístico de A:P al caso multinivel se fundamenta en la metodología del análisis multinivel, en los métodos gráficos existentes para la detección de outliers multinivel y en el estadístico de A:P para la detección de valores discrepantes en regresión lineal múltiple. Se usan conceptos de analogía y equivalencia en la generalización del estadístico de A:P al caso multinivel. El modelo multinivel que se uso para el ajuste es el de dos niveles, de intercepto y una pendiente aleatorios.

En la ilustración de la propuesta de generalización de estadístico de A:P y comparación con los procedimientos gráficos se utilizo una base de datos, tomada de Rasbash, et. al., 2001.

Estos datos se refieren a mediciones (observaciones) realizadas en 907 estudiantes (unidades del nivel uno) pertenecientes a 18 escuelas (unidades del nivel dos) en Inglaterra. Los cálculos de estimación del estadístico de A:P Multinivel se realizaron mediante el uso de un algoritmo realizado en el programa MATLAB y el paquete estadístico MLwin 1.1

CAPÍTULO IV

4.1- PRESENTACIÓN Y ANÁLISIS DE RESULTADOS.

En este capítulo se describe la base de datos multinivel a la que se le aplicarán los métodos gráficos existentes y el estadístico de Andres y Pregibon multinivel propuesta para la detección de outlier; así mismo, se comparan los resultados para validar el estadístico de Andres y Pregibon multinivel.

4.2.- MATRIZ DE DATOS REFERIDA AL RENDIMIENTO ESTUDIANTIL (RASBASH, ET. AL., 2001).

Los datos fueron tomados de Rasbash, et. al. (2001). Éstos se refieren al rendimiento académico de 907 estudiantes en 18 diferentes escuelas en Londres. El objetivo fue modelizar el rendimiento estudiantil al finalizar el año escolar en función de la puntuación obtenida al inicio del año en una prueba de razonamiento verbal y, considerando el efecto de las escuelas y otras variables medidas en los alumnos y las escuelas. Es importante mencionar que los datos son originales de Aitkin y Longford (1986) y fueron analizados también por Goldstein, (1995). Tanto Rasbash como Goldstein concluyen, basados en los métodos gráficos multinivel, que la escuela 17 es discrepante.

Esta matriz de datos se utilizará para aplicar el estadístico de Andres y Pregibon multinivel para detectar posibles escuelas discrepantes y, alumnos discrepantes dentro de las escuelas. También se aplicarán los métodos gráficos multinivel para, por una parte corroborar si la escuela 17 es discrepante y, por otra, validar el estadístico de Andres y Pregibon multinivel como método eficiente en la detección de valores discrepantes.

La matriz de datos tiene una estructura multinivel de dos niveles: alumnos agrupados en escuelas. Los alumnos son las unidades del nivel uno (N1) y las escuelas son las unidades del nivel dos (N2).

Las variables que hemos tomado de la matriz de datos original son:

- Puntuación obtenida al final del año escolar. Ésta es la variable respuesta, se refiere a la puntuación obtenida por el alumno en el examen al final del curso; esta variable fue estandarizada. La variable está codificada como: pfinal, los datos originales y zpfinal los estandarizados. Es una variable medida al nivel de alumnos.

- Puntuación obtenida en una prueba inicial. Considerada como variable explicativa representa la puntuación obtenida por el alumno en el examen de admisión. De igual manera se trabaja con los datos estandarizados. La notación es zpadmisión. Es una variable medida al nivel de alumnos

4.3.- DETECCIÓN DE OBSERVACIONES DISCREPANTES MEDIANTE MÉTODOS GRÁFICOS EN LOS DATOS.

4.3.1.- AJUSTE DEL MODELO DE LOS DATOS.

Se tiene una matriz de datos multinivel de dos niveles: alumnos (subíndice i) anidados dentro de escuelas (subíndice j). La variable respuesta, zpfinal, medida en el nivel 1, se modela en función de la variable explicativa zpadmisión, medida en el nivel 1, considerando la estructura jerárquica o multinivel. Se usa el software MLwiN para estimar los parámetros del modelo. Se inicia destacando la estructura jerárquica y verificando su significación a través del modelo vacio; se continuará agregando la variable explicativa, estudiando su significación y el efecto de ésta a través de los grupos.

La figura 4.1 es la salida del software MLwiN 1.1 que presenta la estructura jerárquica o multinivel de los datos. Se puede apreciar que hay 18 escuelas o unidades del nivel 2 y cada escuela contiene un cierto número de alumnos; por ejemplo, la escuela 1 (L2ID: j=1) contiene 65 alumnos. Así se puede ver para cada escuela.

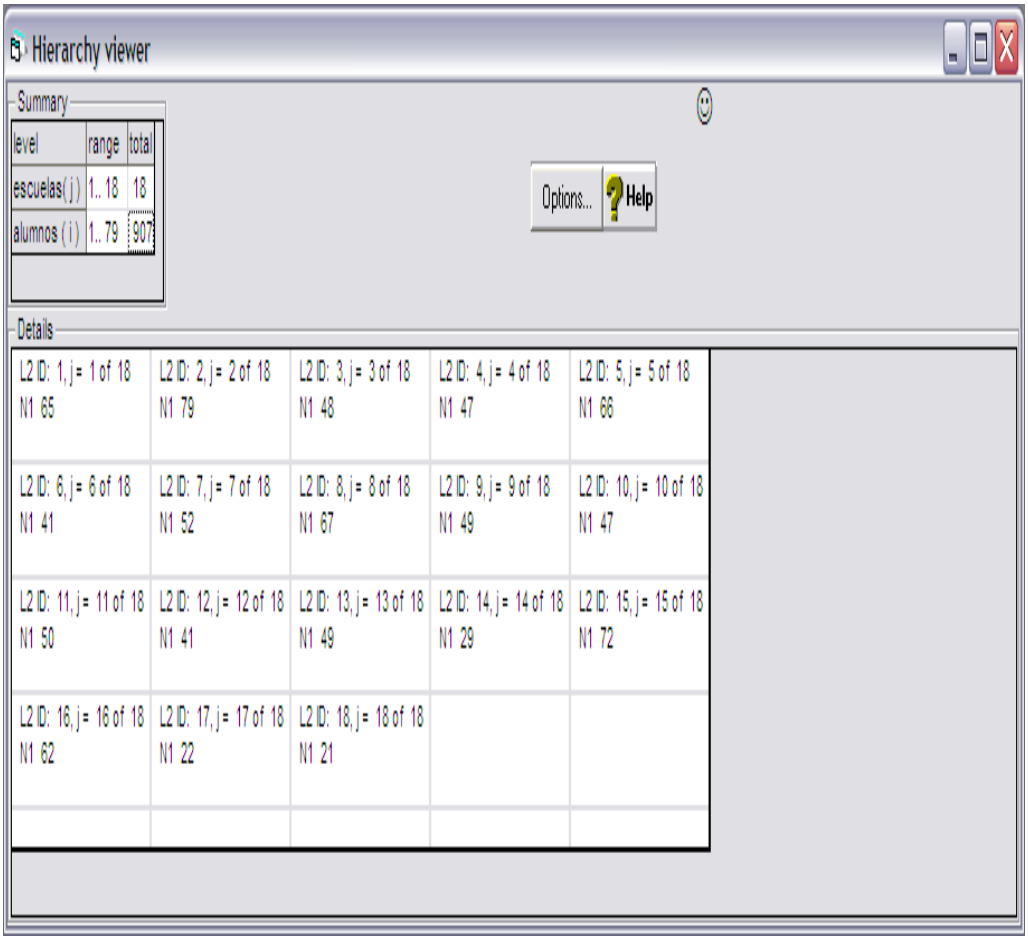


Figura 4.1: Estructura Multinivel de los datos. Salida generada por el Software MLwiN 1.1.

La figura 4.2 contiene el modelo vacio ajustado.

$$\text{zpfinal}_{ij} \sim \text{N}(XB,\ \Omega)$$

$$\text{zpfinal}_{ij} = \beta_{0ij}\text{cons}$$

$$\beta_{0ij} = 0{,}098(0{,}126) + u_{0j} + e_{0ij}$$

$$\left[u_{0j}\right] \sim \text{N}(0,\ \Omega_u)\ :\ \Omega_u = \left[0{,}268(0{,}095)\right]$$

$$\left[e_{0ij}\right] \sim \text{N}(0,\ \Omega_e)\ :\ \Omega_e = \left[0{,}760(0{,}036)\right]$$

$$-2*loglikelihood(IGLS) = 2376{,}439(907 \text{ of } 907 \text{ cases in use})$$

Figura 4.2: Modelo vacio sobre el rendimiento estudiantil en los datos. Salida generada por el Software MLwiN 1.1.

Nótese en el modelo vacio lo siguiente:

(1) el intercepto es aleatorio en ambos niveles y especifica al modelo lineal jerárquico vacío (M_0). La estimación de dicho intercepto es $\hat{\beta}_0 = 0{,}098$con un error estándar $s_{\hat{\beta}_0} = 0{,}126$. El valor del intercepto o coeficiente ajustado representa el promedio estimado global sobre todos los estudiantes; recuérdese que se está trabajando con valores tipificados. El intercepto no resulta significativo puesto que el resultado de dividir el valor estimado entre su error estándar resulta menor que el valor tabulado de la distribución normal estándar con un nivel de significancia fijo; explícitamente $\frac{0{,}098}{0{,}126} = 0{,}777$es menor que $Z_{0,\,05} = 1{,}65$.

(2) las variables aleatorias $U_{0j} y e_{0ij} conj = 1, \ldots, 23, i = 1, \ldots, n_j$;representan los residuos del nivel dos y uno respectivamente. Se supone que U_{0j}tiene distribución normal con media 0 y varianza $\Omega_u \equiv \sigma^2_{u0}$; la estimación de la varianza es $\hat{\sigma}^2_{u0} = 0{,}268$y error estándar $s_{\hat{\sigma}^2_{uo}} = 0{,}095$. Puesto que $\frac{0{,}268}{0{,}095} = 2{,}82$es mayor que $1{,}65 = z_{0{,}05}$; se concluye que la variabilidad al nivel de escuelas es significativa; es decir, zpfinal es diferente a través de las escuelas o equivalentemente las escuelas influyen de manera diferente en el rendimiento. De igual manera, se supone que e_{0ij} es normal con media cero (0) y varianza $\Omega_e \equiv \sigma^2_{e0}$; la estimación de la varianza al nivel uno es $\hat{\sigma}^2_{e0} = 0{,}760$; con un procedimiento comparativo igual a los dos anteriores se concluye que una variabilidad significativa al nivel de alumnos.

(4) la expresión -2*loglikelihood (IGLS) es conocida como la Deviance; es una medida de la falta de ajuste entre el modelo y los datos; en este modelo vacío la deviance estimada es 2376,439; dicho valor será usado como base para compararlo con modelos que incluyan variables explicativas.

La significación de la varianza al nivel dos indica la necesidad de continuar el análisis de los datos considerando lo estructura multinivel de éstos. Sería erróneo analizar los datos sin considerar el agrupamiento en las escuelas. Vamos ahora a agregar una variable explicativa zpadmisión al modelo; es decir, vamos a ajustar el modelo multinivel de intercepto y

pendiente aleatorios. Realizando este procedimiento en el software MLwiN 1.1, se tiene como salida la presentada en la figura 4.3.

Al examinar el modelo de intercepto y pendiente aleatorios (figura 4.3) se destaca lo siguiente:

- Al establecer la diferencia entre las deviances del modelo vacio y el modelo de intercepto y pendiente aleatorios resulta en 607; valor que comparado con el percentil de una distribución chi-cuadrado 3 gl, a saber 7.82, resulta ampliamente mayor; indicando que el modelo de intercepto y pendiente aleatorios se ajusta mejor a los datos.

- El efecto de zpdmisión es significativo ya que $\frac{0,696}{0,034} = 20,47$es mayor que el valor de la distribución normal estándar para cualquier nivel de significancia. Por otra parte, este efecto es diferente a través de las escuelas ya que la varianza de la pendiente de zpadmisión resulta significativa, pues $\frac{0,012}{0,007} = 1,71$> que $1,65=z_{0.05}$.

$$\text{zpfinal}_{ij} \sim N(XB, \Omega)$$
$$\text{zpfinal}_{ij} = \beta_{0ij}\text{cons} + \beta_{1j}\text{Zpadmisión}_{ij}$$
$$\beta_{0ij} = \text{-}0,002(0,032) + u_{0j} + e_{0ij}$$
$$\beta_{1j} = 0,696(0,034) + u_{1j}$$

$$\begin{bmatrix} u_{0j} \\ u_{1j} \end{bmatrix} \sim N(0, \Omega_u) : \Omega_u = \begin{bmatrix} 0,010(0,006) & \\ 0,011(0,005) & 0,012(0,007) \end{bmatrix}$$

$$\begin{bmatrix} e_{0ij} \end{bmatrix} \sim N(0, \Omega_e) : \Omega_e = \begin{bmatrix} 0,401(0,019) \end{bmatrix}$$

$$\text{-2*loglikelihood(IGLS)} = 1769,297(907 \text{ of } 907 \text{ cases in use})$$

Figura 4.3: Modelo de intercepto y pendiente aleatorios para los datos. Salida generada por el Software MLwiN 1.1.

Considerando que el modelo de intercepto y pendiente aleatorios (figura 4.3) caracteriza la respuesta en función de la explicativa, se procede a estimar, con dicho modelo, la recta promedio, y la de cada escuela; además, los residuos tanto del nivel uno como del nivel dos. La figura 4.4 presenta algunas gráficas de interés. En (a) se observa que la escuela 17, la cual se resalta en color rojo, tiene el valor más alto de intercepto y uno de los más altos valores en la pendiente; en (b) se presenta el diagrama de dispersión, también se puede ver que los estudiantes en la escuela 17 tienden a tener puntajes altos, con la posible excepción de un alumno cuyo puntaje se encuentra cerca de la media, este alumno es el numero 22.

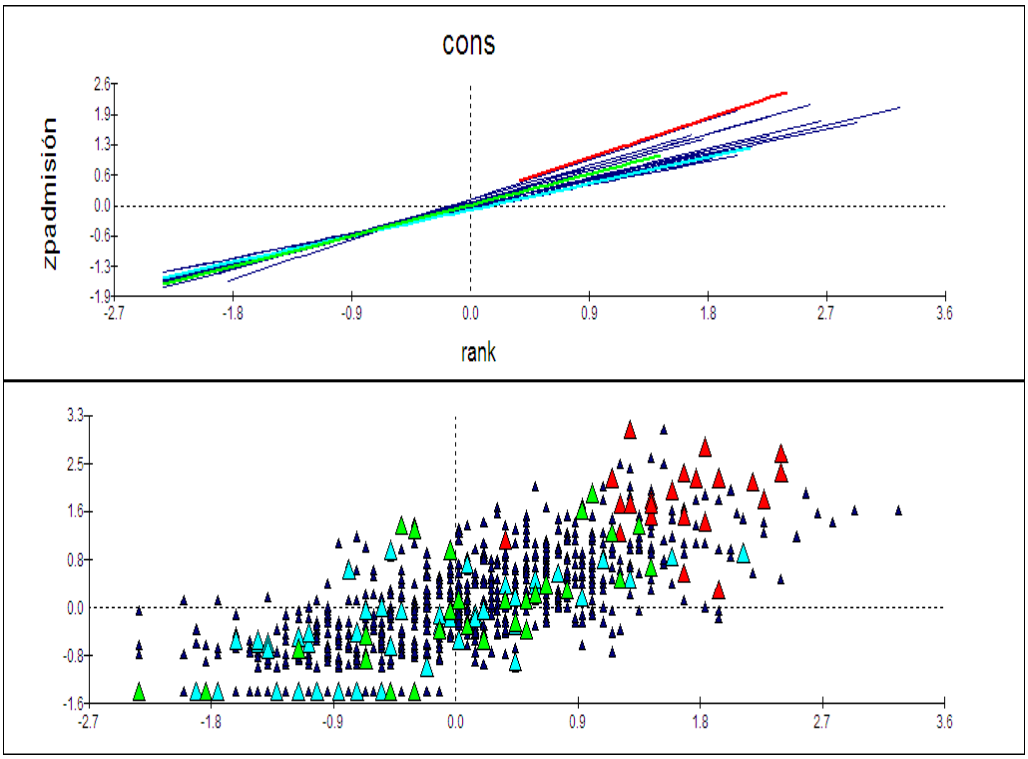


Figura 4.4: (a) Curva promedio y (b) diagrama de dispersión de las escuelas. Salida generada por el Software MLwiN 1.1

Por otra parte, una vez obtenidos los residuos de ambos niveles, interesa primeramente examinar los del nivel dos. En este nivel se tienen los residuos del intercepto y los de la pendiente. La figura 4.5 presenta un gráfico de ambos residuos (pendiente e intercepto) con intervalos de confianza del 95%. En (a) se muestra el residuo, respecto al intercepto, de cada escuela, observándose que la escuela 17 tiene el mayor residuo. En (b) se presentan los residuos, con respecto a la pendiente, para cada escuela, se observa que la escuela 17 tiene el

mayor residuo. Esto confirma que la escuela 17 que se observa remarcada en rojo tiene el mayor intercepto residual, y pendiente máxima residual.

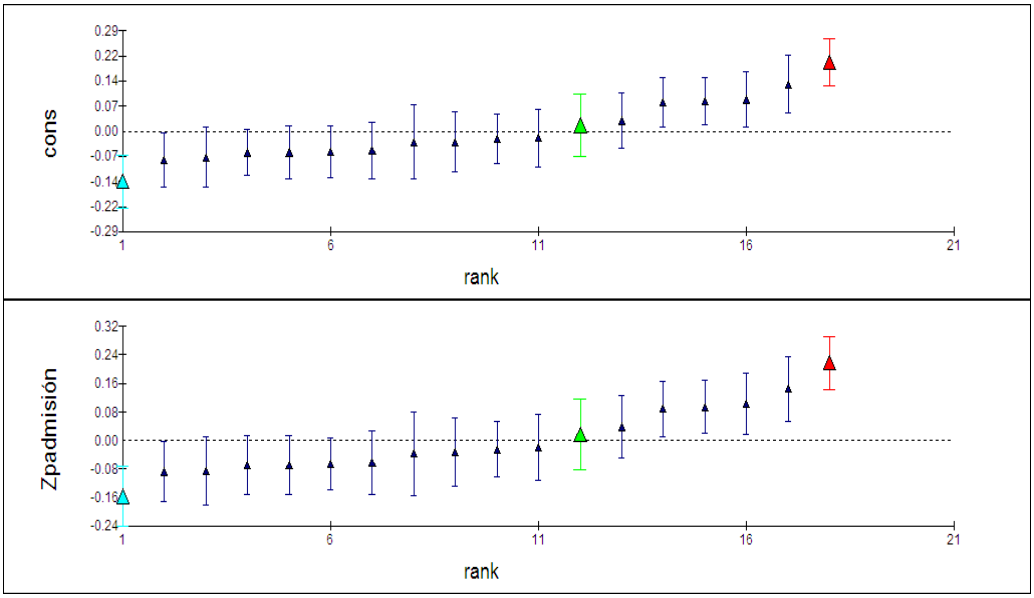


Figura 4.5: Representación de los residuos de intercepto y pendiente de las escuelas. Salida generada por el Software MLwiN 1.1

También es posible examinar la relación entre intercepto y pendiente a través de un gráfico de dispersión de ambos residuos. Se verifica que existe una relación positiva muy fuerte entre los valores de los residuos de la pendiente y el intercepto; con esto se espera que escuelas con intercepto grandes tienden a tener pendientes de la explicativa grandes. La figura 4.6 muestra el diagrama.

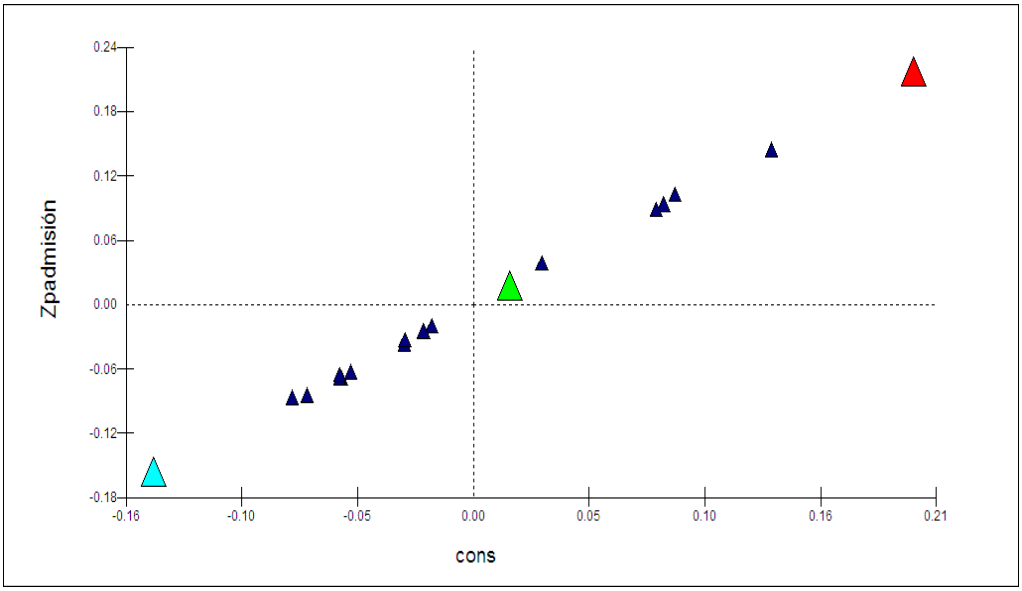


Figura 4.6: Diagrama de pares de residuos para el intercepto y la pendiente de las escuelas. Salida generada por el Software MLwiN 1.1

4.3.2.- Diagnóstico de escuelas discrepantes a través de residuos puros, residuos estandarizados, supresión residual, Leverage y DFITS.

El software MLwiN proporciona directamente una gráfica de los residuos puros, residuos estudentizados, de residuos de supresión, de los valores de influencia (leverage) y de los DFITS tanto para los residuos del intercepto como para los de la pendiente, ambos del nivel dos.. A continuación, la figura 4.7 presenta estas gráficas para los residuos del intercepto.

En el primer recuadro, resaltado como (1), se presentan los residuos puros, respecto al intercepto, de cada escuela; se puede observar en éstos que el que corresponde a la escuela 17 es el mayor (destacado con el color rojo). Es importante mencionar que para interpretar o entender el gráfico se debe prestar atención es a la magnitud del residuo, reflejada en la recta horizontal; los residuos aquí están representados por rectángulos del mismo tamaño; residuos de igual magnitud para diferentes escuelas, los rectángulos correspondientes se monta uno sobre el otro en la gráfica.

En el gráfico (2) de la figura 4.7 se presentan los residuos estandarizados; se puede observar del diagnóstico que la escuela 17 tiene el más alto residuo y supera el valor de 2, lo

que indica, tal como se describió en el capítulo 2, que la escuela puede considerarse un valor discrepante. El tercer gráfico (3) representa los valores de influencia (Leverage) para cada escuela; se observa que la escuela 6 tiene un valor aproximado de 0.35, claramente superior a $\frac{24}{18} = 0{,}22$, es posible que esta escuela sea influyente en la estimación de los parámetros fijos del modelo, se debe examinar con más detalle. La escuela 17 tiene un leverage alrededor de 0.17, lo cual es bajo.

La gráfica (4) de la figura 4.7 muestra los valores DFITS; se observa que el valor correspondiente para la escuela 17 es el más grande e igual a 0.5, que no supera el valor $() = 0.83$, con lo cual la escuela 17 no es influyente en la estimación de los parámetros.

$$2.5\sqrt{\frac{2/18}{}}$$

La gráfica (5) de la figura 4.7 muestra los residuos de supresión; se observa que para la escuela 17 este valor es aproximadamente 2.5, el cual está dentro del rango (-3, 3), indicando que la escuela no es atípica. Finalmente la gráfica (6) presenta un gráfico de dispersión de los leverages vs residuos estandarizados.

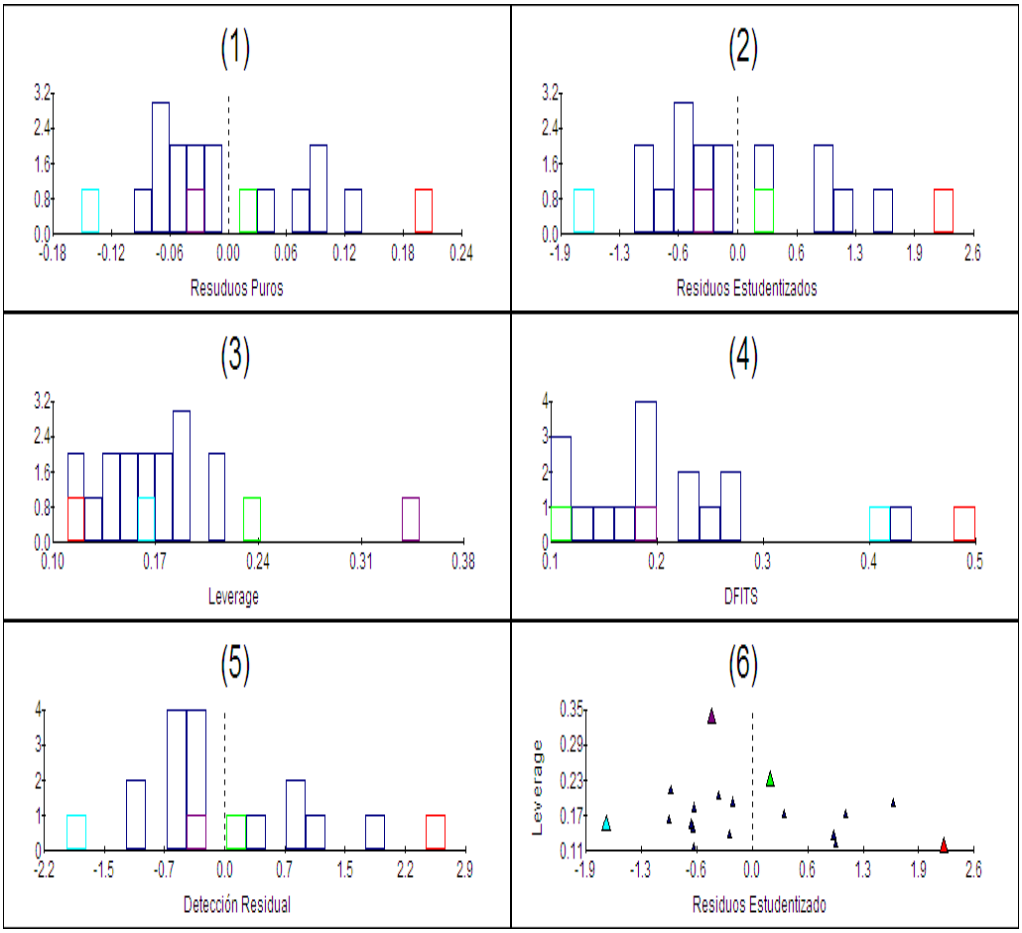


Figura 4.7: gráficas de residuos puros, residuos estandarizados, supresión residual, leverage y DFITS con respecto al intercepto de las escuelas. Salida generada por el Software MLwiN 1.1

A continuación se realiza el mismo análisis para los residuos, del nivel dos, de la pendiente de la variable explicativa (zpadmisión). La figura 4.8 presenta las graficas del (1) al (6) correspondientes a los residuos puros, residuos estandarizados, leverage, DFITS, supresión residual y el diagrama de dispersión de leverage vs residuos estandarizados.

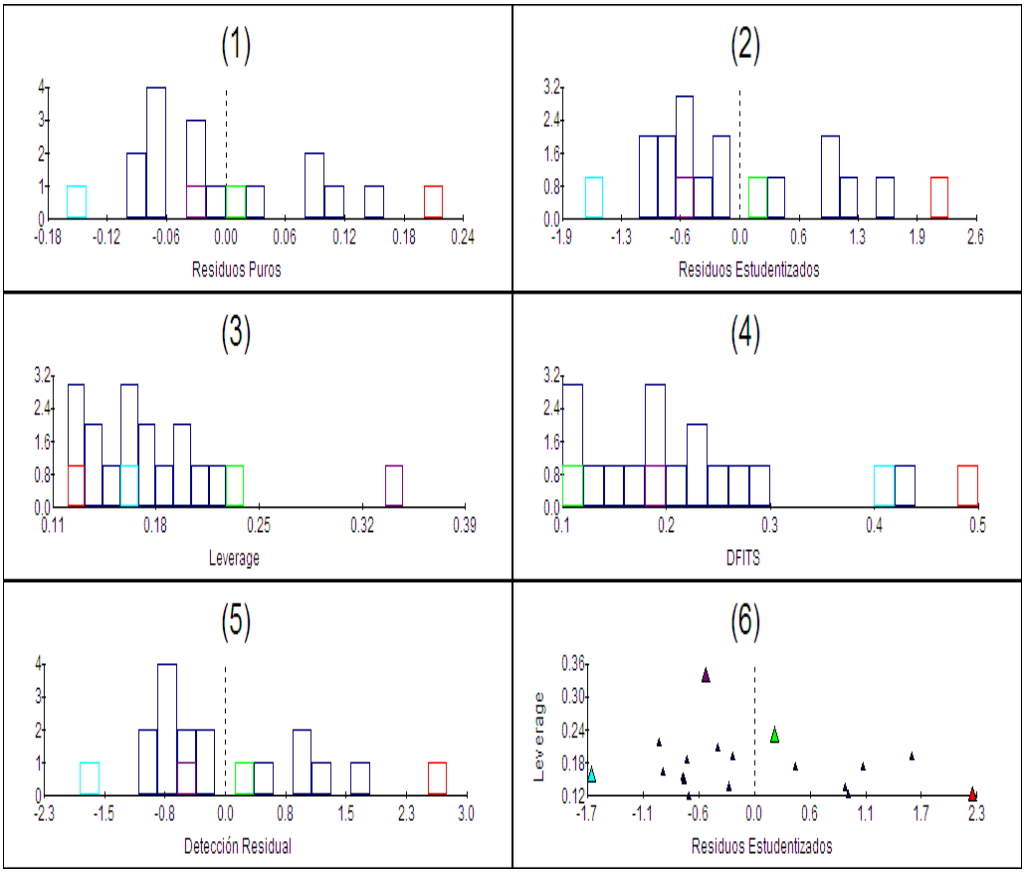


Figura 4.8: gráficas de residuos puros, residuos estandarizados, supresión residual, leverage y DFITS con respecto a la pendiente de las escuelas. Salida generada por el Software MLwiN 1.1

En los gráficos (1) y (2) de la figura 4.8 se presentan los residuos puros y estudentizados correspondientes al término aleatorio de la pendiente; se observa que estos residuos de la escuela 17 tiene los valores más grandes; esto indica que la escuela 17 es atípica con respecto a las demás. En la gráfica (3) el leverage para esta escuela es pequeño (0,12<4/18=0,22; sin embargo, en la gráfica (4) se observa que el DFITS es el más grande (0,5<$\frac{2/18}{2,5\sqrt{}} = 0,83$), pero no llega a indicar que es influyente en la estimación de los parámetros. En la gráfica (5) se observa que el valor correspondiente a la supresión residual, para esta escuela es el mayor, indicando que es un valor atípico. En cuanto al diagrama de dispersión, gráfico (6), se resalta que la escuela 17 presenta el mayor residuo estandarizado pero un pequeño leverage; no obstante, la escuela 6 presenta el mayor leverage y un residuo estandarizado pequeño.

Para complementar este diagnóstico de valores atípicos e influyentes, se realiza un análisis de los residuos al nivel de los estudiantes (nivel 1), con el fin de detectar si existe alguna unidad discrepante o que este perturbando alguna escuela. La figura 4.9 presenta el diagrama de dispersión de los residuos estudentizados contra las puntuaciones estandarizadas de zpfinal; se puede observar que los residuos de la escuela 17, remarcados con el color rojo, son muy dispersos debido a que hay uno que es considerablemente bajo con respecto a los demás. Esto permite concluir que en la escuela 17 hay al menos un alumno que es discrepante respecto a los demás.

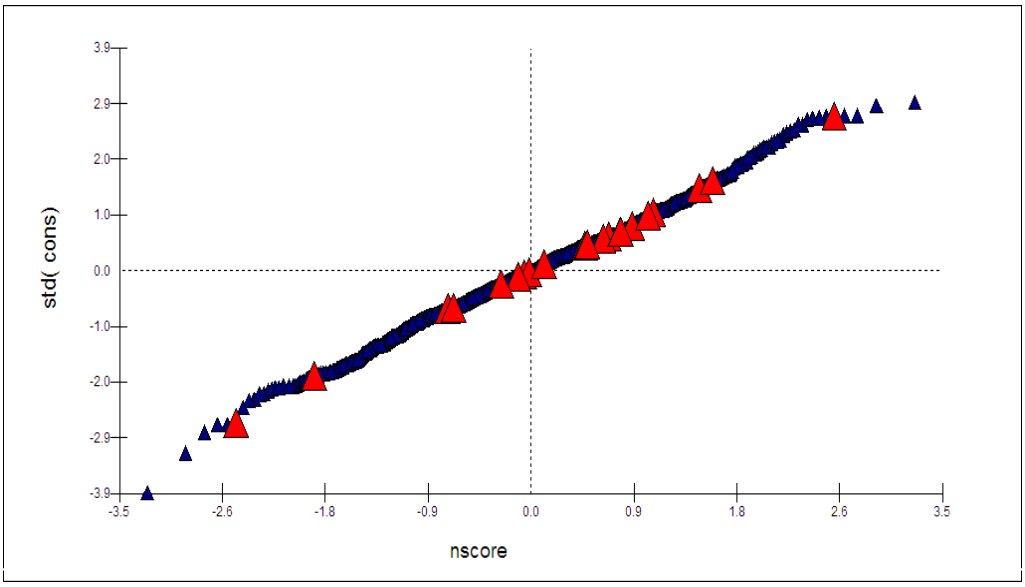


Figura 4.9: Residuos Estandarizados del nivel uno vs puntuación normal de zpfinal. Salida generada por el Software MLwiN 1.1

En resumen de los procedimientos de diagnóstico aplicados al conjunto de datos, se puede concluir que la escuela 17 es discrepante respecto a las demás, sin embargo no llega a ser preponderantemente influyente en la estimación de los parámetros del modelo; esto es tanto para el intercepto como para la pendiente

4.4 Aplicación del estadístico de Andres y Pregibon multinivel.

En esta sección se aplica el estadístico de Andres y Pregibon multinivel al conjunto de datos. Primeramente se calcula el estadístico de Andres y Pregibon multinivel a cada

estudiante para ver si alguno de éstos es influyente. En segundo lugar se calcula el estadístico de Andres y Pregibon multinivel en cada escuela.

Para realizar los cálculos del estadístico se hizo uso de un algoritmo realizado con el programa MATLB 7.8 definido de la siguiente manera:

En la tabla 1.1 se presenta el estadístico de Andres y Pregibon multinivel para cada una de las 18 escuelas y para el mayor y el menor estudiante.

Escuela	R^j	**Estudiantes**	$R_i^j maxymin$
1	0.85841155	63	0.99892678
		52	0.89817910
2	0.83421793	60	0.99964155
		42	0.87664492
3	0.90439074	18	0.99292327
		15	0.86890289
4	0.91824651	43	0.99855155
		39	0.77093263
5	0.86578549	18	0.99936315
		65	0.89471547
6	0.92412281	31	0.99170715
		2	0.80863979
7	0.89368325	39	0.99869908
		18	0.86234265
8	0.88247540	48	0.99694111
		42	0.89393232
9	0.88630220	46	0.99878384
		11	0.82283494
10	0.89152068	32	0.98644766
		41	0.88538709
11	0.90468059	43	0.99357571
		5	0.84461224

12	0.91482888	6	0.99861178
		14	0.85003741
13	0.89599812	1	0.99782647
		33	0.81419034
14	0.94243623	25	0.99442529
		9	0.76399059
15	0.86657140	52	0.99941002
		62	0.92885235
16	0.88150594	57	0.99935162
		62	0.88791869
17	0.87719773	7	0.97123136
		22	0.74470924
18	0.90469262	14	0.97470438
		13	0.76744892

Tabla 1.1: Estadístico de Andres y Pregibon para el nivel 2.

Puede observarse en la tabla 1 que el menor valor del estadístico le corresponde a las escuela 2, 1, 15 y 17, ahora si revisamos los R^jal nivel 1 se puede observar que la escuela 17 es la que posee la mayor cantidad de R^jmás pequeños en comparación con el resto de las escuela, esto significa que no hay una variación sustancial en los coeficientes al excluir cualquiera de las 18 escuelas en la estimación de los parámetros.

Es importante destacar que esta escuela es la que presenta el cuarto menor R^jy el mayor numero de R^jal nivel 1 más pequeño, indicando entonces que esta es una señal para examinarla con mayores detalles. Esto de alguna manera es coincidente con lo obtenido a través de los resultados de los residuos, el leverage, los DFITS y la supresión residual.

4.5 Comparación empírica de los métodos gráficos y el Estaditico de (A.P).

En esta sección se comparan los resultados de los métodos gráficos y el Estadístico de Andres y Pregibon. Se complementan los resultados presentados por los autores antes citados,

con una evaluación del comportamiento de los valores discrepantes y de puntos influyentes mediante la utilización del Estadístico de Andres y Pregibon Multinivel propuesta.

Se han aplicado varios procedimientos de diagnóstico de valores discrepantes e influyentes y el Estadístico de Andres y Pregibon propuesto para detectar observaciones y grupos discrepantes en la matriz de datos tomada de Rasbas et. al. (2001). Como se ha ilustrado, los residuos puros, estudentizados y la supresión residual indican que la escuela 17 es atípica (discrepante, outlier) con respecto al intercepto y a la pendiente de zpadmisión. Sin embargo, esta escuela no llega a ser preponderantemente influyente en la estimación de los parámetros del modelo, como lo indica el leverage, los DFITS y e Estadístico de Andres y Pregibon.

Por otra parte, la escuela 6 presenta un valor de leverage grande tanto para el intercepto (0,35) como para la pendiente (0,34), los cuales son claramente superior a $\frac{24}{18} = 0{,}22$, indicando que esta escuela puede ser influyente estimación de los parámetros fijos del modelo; no obstante, como los residuos de esta escuela tanto para el intercepto como para la pendiente, son bajos (-0,03) se concluye que no es influyente.

Con los resultados obtenidos mediante la aplicación del Estadístico de Andres y Pregibon propuesto, se puede concluir que no hay escuela influyente en la estimación de los parámetros del modelo y alumnos dentro de las escuelas tampoco.

CONCLUSIONES

1.- En el Análisis Multinivel se tiene establecido como procedimientos de diagnóstico de valores atípicos (distantes, outliers) en los diferentes niveles de la estructura jerárquica el análisis de los residuos puros, residuos estandarizados y de supresión residual. Estos se examinan en los diferentes niveles y se considera una unidad atípica si el residuo correspondiente a ésta en ese nivel es excesivamente grande, generalmente fuera del rango (-3; 3).

2.- En el caso del diagnostico de valores influyentes en la estimación de los parámetros del modelo se tienen como procedimientos de diagnóstico los valores de apalancamiento (Leverage) y los DFITS. Al igual que los residuos, los leverage y DFITS se calculan para cada nivel; valores grandes de leverage y residuos grandes indica que la unidad correspondiente es influyente en la estimación de los parámetros. Un Valor DFITS grande para una unidad particular indica que ésta es influyente en la estimación de los parámetros del modelo.

3.- El estadístico de Andresw y Pregibon se ha construido de forma análoga al caso de la regresión OLS, el estadistico propuesto constituye un indicador de la influencia que tiene la i-ésima observación sobre los coeficientes de regresión estimados en el modelo multinivel de dos niveles con intercepto y pendiente aleatorios. Este estadístico se basa en el cociente R_(ij…)^((k)) (X^)=(D_(ij…)^k |X^T X^ |)/|X^T X^ |

4.- En la aplicación práctica de se pudo comprobar que los resultados con el estadístico propuesto coinciden con los de leverage y DFITS en cuanto a que no hay escuela influyente en la estimación de los parámetros.

5.- Se mostró una ilustración práctica en la que se evidenció la utilidad del estadístico para detectar valores influyentes y se presento una tabla que permitió visualizar y resumir los resultados más importantes.

6.- En cualquier investigación estadística deben tomarse todas las medidas necesarias para evitar, en la medida de lo posible, errores en los datos. Los archivos de datos deben ser revisados varias veces para disminuir al máximo la posibilidad de errores de medición y de transcripción. Una vez obtenidos los datos y superada esa primera fase de evaluación, el investigador desea y espera no encontrarse con datos que pudiesen distorsionar

los modelos ajustados. Si esos datos existen, es indispensable reflexionar sobre la naturaleza de la información contenida en el mismo, antes de tomar cualquier decisión al respecto. El estadístico de Andresw y Pregibon propuesto en esta tesis, acompañada de las herramientas de diagnostico utilizadas usualmente en el análisis multinivel, puede ser de gran ayuda para el investigador en la identificación y manejo de los datos influyentes.

RECOMENDACIONES

1. Hacer un estudio de simulación con el estadístico de Andrews y Pregibon haciendo uso del software desarrollado, utilizando datos multinivel donde se conozcan los resultados y de esta manera verificar la efectividad del mismo.

2. Sería de gran utilidad seguir realizando estudios referentes a métodos analíticos para detectar posibles valores extraños en datos multinivel, y de esta manera el investigador podrá tener a la mano una herramienta que le permita decidir de manera precisa cuando un valor es discrepante o influyente en un conjunto de datos multinivel.

REFERENCIAS BIBLIOGRÁFICAS

Andrews, DF. And pregibon, d. (1978. “Finding the outlires tha Matter” Journal of Royal Statiscal Society,serie B, 40, 85 -93

Barnett V and Lewis T (1994). Outliers in Statistical Data: 3rd edition. New York: John Wiley.

Cook R.D. y Wang P.C (1983). “Transformations and influential Cases in Regression”, Tecnometrics, 25,337-343.

Fermín, W., P. Galindo y Martin, J. (2007). Estrategia multinivel para detectar comportamiento de abandonos en análisis de datos longitudinales.

Fermín, W., (2004). Contribuciones al Análisis de datos Longitudinales bajo una perspectiva Multinivel, Tesis Doctoral, Universidad de Salamanca, España.

Fermín, W., (1993). Detección de valores extraños en regresión lineal. Tesis de Pregrado, Universidad de Oriente Venezuela.

Goldstein, H. (1995a). Multilevel Statistical Models. London, Edward Arnold: New York, Wiley.

Langford, I. H. and Lewis, T. (1998). Outliers in multilevel models (with Discussion). J. Royal Statistical Society, A, 161: 121-160

Montgomery, D. 1991. Diseño y Análisis de Experimentos. Grupo Editorial Iberoamérica, Nebraska.

Rasbash, J., Browne, W., Goldstein, H., et al (2001) A User's Guide to MLwiN Centre for Multilevel Modelling. London: Institute of Education, University of London.

Snijders, T. & Bosker, R. (1999). Multilevel analysis. London: Sage Publications.

Van der Leeden, R. (1998). Multilevel analysis of longitudinal data. In C. C. J. H. Bijleveld & L. J. Van der Kamp (Eds.), Longitudinal data analysis (pp. 268–315). London: Sage. Conocimientos básicos requeridos: Análisis de regresión. Análisis de varianza.

ANEXOS

Algoritmo realizado con el programa MATLAB 7.8 para realizar los cálculos matriciales.

```
clc
clear
fid = fopen('TotalRk1.txt','w');
fclose(fid);
fid = fopen('TotalRk2.txt','w');
fclose(fid);
load matriztesis.mat
A = matriztesis;
OrdenA = size(A);
n2 = input('Cuantos elementos hay en el nivel dos: '    );
load UnidNivel1
% Rk2 = zeros(1,n2);
% UnidNivel1 = zeros(1,n2);
% for i = 1:n2
%     fprintf('Elmento del nivel dos, Bloque %2.0f  ',i)
%     UnidNivel1(i) = input('Unidades en el nivel uno: '   );
% end
ss = min(UnidNivel1);
Rk = zeros(1,ss);
cont = 0;
contador= zeros(1,n2);
k = 1;
for k = 1:n2
    B = A(1 + cont:cont + UnidNivel1(k),:);
    contador(k) = cont + UnidNivel1(k);
    Detm = det(B'*B);
    Rk = [];
    for n = 1:UnidNivel1(k)
        BS = B;
        BS(n,:) = [];
        DetS = det(BS'*BS);
        Rk(n) = DetS/Detm;
    end
     fid = fopen('TotalRk1.txt','a');
     fprintf(fid,'%3.4f  \n',Rk);
     fclose(fid);
    fprintf('Bloque %5.0d Rk = \n',k);
    fprintf('%2.8f \n',Rk);
    cont = cont + UnidNivel1(k);
end
Detm2 = det(A'*A);
cont2 = 0;
for k = 1:n2
    B2S = A;
    B2S( (1 + cont2):( cont2 + UnidNivel1(k) ),:) = [];
    Det2S = det(B2S'*B2S);
    Rk2 = Det2S/Detm2;
    fprintf('Nivel 2 Rk2 %5.0d  = \n',k);
    fprintf('%2.8f \n',Rk2);
    fid = fopen('TotalRk2.txt','a');
    fprintf(fid,'%3.4f  \n',Rk2);
    fclose(fid);
    cont2 = cont2 + UnidNivel1(k);
end
```

HOJAS DE METADATOS

Hoja de Metadatos para Tesis y Trabajos de Ascenso – 1/6

Título	**DETECCIÓN DE VALORES ATÍPICOS MULTINIVEL MEDIANTE UNA GENERALIZACIÓN DEL ESTADÍSTICO DE ANDREWS Y PREGIBON**
Subtítulo	

Autor(es)

Apellidos y Nombres	**Código CVLAC / e-mail**	
Lcdo. Machado Rodríguez Enrique Fernando	**CVLAC**	**13 598 124**
	e-mail	**enrique.f.machado@gmail.com**
	e-mail	
	CVLAC	
	e-mail	
	e-mail	
	CVLAC	
	e-mail	
	e-mail	
	CVLAC	
	e-mail	
	e-mail	

Palabras o frases claves:

valores discrepantes e influyentes, análisis multinivel, matriz de diseño, intercepto y pendiente aleatorios

Hoja de Metadatos para Tesis y Trabajos de Ascenso – 2/6

Líneas y sublíneas de investigación:

Área	Subárea
Escuela de Ciencias	**Departamento de Matemáticas**

Resumen (abstract):
En esta investigación se describen los diversos procedimientos para detectar valores discrepantes e influyentes en la modelización multinivel y se propone una nueva alternativa para detectar estos valores. El análisis multinivel es una metodología para analizar datos con estructura jerárquica o multinivel; estos datos se caracterizan por unidades individuales anidadas o agrupadas en niveles superiores. Al igual que en el análisis de regresión clásico donde las estimaciones se ven afectadas por valores atípicos, los datos multinivel pueden presentar uno o más valores atípicos que afectarían los coeficientes estimados del modelo de regresión jerárquico o multinivel ajustado. En análisis de regresión clásico existe una variedad de métodos analíticos y gráficos para detectar valores atípicos (ver Barnett y Lewis, 1978 citado por Montgomery, 1991) aunque éstos son una herramienta valiosa, en la mayoría de los casos es inevitable la subjetividad del investigador, por lo que la verdadera detección puede quedar en tela de juicio. Un estadístico interesante para detectar valores atípicos en regresión lineal es el de Andrews y Pregibon (1978), (ver también Fermín, 1993) el cual se basa en el determinante de la matriz de diseño con el vector de respuesta adjuntado; lo interesante de este estadístico es que detecta tanto los atípicos distantes como los atípicos influyentes en el modelo. Aquí se ha desarrollado un estadístico para la detección de unidades, del nivel dos como del nivel uno, discrepantes y/o influyentes en la estimación de los parámetros de un modelo lineal jerárquico de dos niveles de intercepto y pendiente aleatorios; el estadístico es una generalización del estadístico de Andrews y Pregibon usada en regresión clásica, al caso multinive

Hoja de Metadatos para Tesis y Trabajos de Ascenso – 3/6

Contribuidores:

Apellidos y Nombres	ROL / Código CVLAC / e-mail	
Dr. Wilmer Fermín	**ROL**	**CA** [ ] **AS** [ ] **TU** [ ] **JU** [ ]
	CVLAC	
	e-mail	**Wilmer_fermin@hotmail.com**
	e-mail	
Dr. José Fermín	**ROL**	**CA** [ ] **AS** [ ] **TU** [ ] **JU** [x]
	CVLAC	
	e-mail	**Simon.fermin@gmail.com**
	e-mail	
MSC. Erasmo Salazar	**ROL**	**CA** [ ] **AS** [ ] **TU** [ ] **JU** [x]
	CVLAC	
	e-mail	**Erasmo7772003@yahoo.com**
	e-mail	

Fecha de discusión y aprobación:

Año	Mes	Día
2012	**12**	**07**

Lenguaje: **SPA**

Hoja de Metadatos para Tesis y Trabajos de Ascenso – 4/6

Archivo(s):

Nombre de archivo	Tipo MIME
PG-Machadoe.Doc	**Aplication/word**

Alcance:

Espacial: _____ **(Opcional)**

Temporal: _____ **(Opcional)**

Título o Grado asociado con el trabajo Trabajo presentado como requisito parcial para optar al título de magister scientiarum en matemáticas

Nivel Asociado con el Trabajo: Msc -

Área de Estudio: Matemáticas

Institución(es) que garantiza(n) el Título o grado: Universidad de Oriente

Hoja de Metadatos para Tesis y Trabajos de Ascenso – 5/6

UNIVERSIDAD DE ORIENTE
CONSEJO UNIVERSITARIO
RECTORADO

CU Nº 0975

Cumaná, 04 AGO 2009

Ciudadano
Prof. JESÚS MARTÍNEZ YÉPEZ
Vicerrector Académico
Universidad de Oriente
Su Despacho

Estimado Profesor Martínez:

Cumplo en notificarle que el Consejo Universitario, en Reunión Ordinaria celebrada en Centro de Convenciones de Cantaura, los días 28 y 29 de julio de 2009, conoció el punto de agenda **"SOLICITUD DE AUTORIZACIÓN PARA PUBLICAR TODA LA PRODUCCIÓN INTELECTUAL DE LA UNIVERSIDAD DE ORIENTE EN EL REPOSITORIO INSTITUCIONAL DE LA UDO, SEGÚN VRAC Nº 696/2009"**.

Leído el oficio SIBI – 139/2009 de fecha 09-07-2009, suscrita por el Dr. Abul K. Bashirullah, Director de Bibliotecas, este Cuerpo Colegiado decidió, por unanimidad, autorizar la publicación de toda la producción intelectual de la Universidad de Oriente en el Repositorio en cuestión.

Comunicación que hago a usted a los fines consiguientes.

UNIVERSIDAD DE ORIENTE
SISTEMA DE BIBLIOTECA
RECIBIDO POR
FECHA 5/8/09 HORA 3:30

Cordialmente,

JUAN A. BOLAÑOS CURIEL
Secretario

REPÚBLICA BOLIVARIANA DE VENEZUELA * UNIVERSIDAD DE ORIENTE * SECRETARÍA CONSEJO UNIVERSITARIO

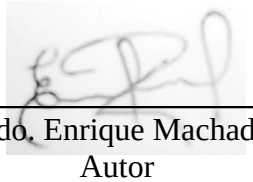
C.C: Rectora, Vicerrectora Administrativa, Decanos de los Núcleos, Coordinador General de Administración, Director de Personal, Dirección de Finanzas, Dirección de Presupuesto, Contraloría Interna, Consultoría Jurídica, Director de Bibliotecas, Dirección de Publicaciones, Dirección de Computación, Coordinación de Teleinformática, Coordinación General de Postgrado.

JABC/YGC/maruja

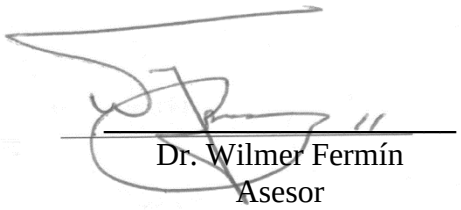
Apartado Correos 094 / Telfs: 4008042 - 4008044 / 8008045 Telefax: 4008043 / Cumaná - Venezuela

Hoja de Metadatos para Tesis y Trabajos de Ascenso- 6/6

Artículo 41 del REGLAMENTO DE TRABAJO DE PREGRADO (vigente a partir del II Semestre 2009, según comunicación CU-034-2009) : "los Trabajos de Grado son de la exclusiva propiedad de la Universidad de Oriente, y sólo podrán ser utilizados para otros fines con el consentimiento del Consejo de Núcleo respectivo, quien deberá participarlo previamente al Consejo Universitario para su autorización".



Lcdo. Enrique Machado
Autor



Dr. Wilmer Fermín
Asesor