UNIVERSIDAD NACIONAL DE INGENIERÍA

FACULTAD DE INGENIERÍA ECONÓMICA, ESTADÍSTICA Y CCSS.
ESCUELA PROFESIONAL DE INGENIERÍA ESTADÍSTICA

TÌTULO DE TESIS

Efectos del truncamiento y censura de la variable dependiente en los modelos de regresión multinivel de 2 etapas

Design of a Two-Stage Multilevel Regression Model with Truncation and Censoring of the Dependents Variables

Ejecutor:

Huanca Huamaní Milagros Luz

Asesor:

Luis Sánchez

LIMA, 2016

DEDICATORIA

Dedico esta tesis al Señor Dios por darme fortaleza en el día a día para poder culminar satisfactoriamente lo que empiezo y ánimo a no quedarme en el mismo lugar sino avanzar porque la superación es agradable para él.

Les dedico esta tesis también a mis padres Daniel Huanca Barriga y Gloria Huamaní Choquichanca por animarme a seguir adelante y apoyarme en las decisiones que tomo.

AGRADECIMIENTOS

Agradezco a mis profesores Richard Fernández por animarnos y empujarnos a la realización de la tesis en el curso Taller de Tesis. Con sus consejos y disciplina, el avance de la tesis fue bueno.

Así mismo agradezco al profesor Luis Sánchez quien no dudó en apoyarme como guía en la elaboración de la presente tesis.

Finalmente, quiero agradecer a mi amigo Manuel Valdivia por sus consejos de cómo organizarme con los tiempos y apoyo constante.

RESUMEN

Se ha demostrado que los datos truncados y censurados afectan la estimación de los parámetros por MCO en los modelos de regresión lineal(Zuehlke & Kassekert, 2008). También se sabe que los modelos multinivel constituyen la metodología de análisis más adecuada para tratar datos “jerarquizados” o “anidados” (Murillo, 2008). Los modelos multinivel son considerados ampliaciones de los modelos de regresión lineal clásicos.

Lo que se hizo en esta investigación es estudiar el efecto marginal de la variable dependiente truncadas y censuradas en los modelos de regresión multinivel de 2 etapas en los siguientes criterios: criterio de información de Akaike (AIC), criterio de información de Akaike corregido (AICC) y criterio de información de Akaike consistente (CAIC) en los modelos de regresión multinivel de 2 etapas. Esto bajo el esquema de 2 escenarios: En el primer escenario se trabajó con una muestra de 200 registros y diferentes proporciones de censura y truncamiento como son $10, 20\ y\ 30\%$. Esto teniendo en cuenta que se considera una censura alta desde el 20%. En el segundo escenario, se fijó una proporción de censura y truncamiento del 10% para diferentes tamaños muestrales como son 50, 200 y 600 registros.

Se concluyó que el hecho de que la variable dependiente se encuentra censurada o truncada tiene efectos (en diferencias porcentuales) sobre los criterios de información Akaike (AIC), criterios de información de Akaike corregido (AICC) y criterios de información de Akaike consistente (CAIC) pues estos valores tienden a disminuir hasta 12% en el caso de truncamiento y aumentar hasta 190% en el caso de censura.

Índice

CAPÍTULO I

ANTECEDENTES

1.1. Investigaciones

- **Los modelos multinivel como herramienta para la investigación educativa**(Murillo, 2008)

En el artículo presentado por Murillo en el año 2008 se pretende realizar una introducción a los modelos de regresión multinivel haciendo una aplicación en la investigación educativa de carácter cuantitativo. Así, se presenta, en primer lugar, un análisis de los fundamentos de los modelos multinivel en donde se declara que la principal característica de los modelos multinivel es que aportan un entorno natural dentro del cual se pueden comparar teorías sobre relaciones estructurales entre variables en cada uno de los niveles en los que se organizan los datos. En segundo lugar, se analiza el proceso de modelaje detalladamente, y se finaliza reflexionando sobre de las aportaciones y utilidades de esta metodología de análisis dentro de las cuales se destaca la identificación de patrones y el reconocimiento de grupos específicos de alumnos con fracaso escolar que necesitan estudios intensivos.

- **Modelos censurados, truncados y con selección muestral**(Álvarez, 2008)

Los objetivos de esta publicación son mostrar la diferencia entre muestras truncadas y censuradas, explicar por qué la estimación por MCO de un modelo lineal es sesgada e inconsistente en tales circunstancias y proponer métodos para estimar muestras en las que la variable dependiente es continua pero limitada (bien por censura o truncamiento).La solución a este problema es plantear un modelo híbrido que utilice la especificación PROBIT para investigar por qué algunas observaciones toman valor 0 y otras no y, para aquellas observaciones tales que Y* > 0, un modelo de regresión que nos cuantifique la relación.El modelo TOBIT recoge esos dos aspectos.

- **A comparative study of MLwiN, HLM, SPSS and Stata** (Martinez, 2014)

Hoy en día contamos con diferentes programas estadísticos para la estimación de los Modelos Multinivel, o Modelos Jerárquicos. En este artículo se realiza una comparación entre los cuatro programas más utilizados para ello, el MLwiN, el HLM, el SPSS y el Stata. La estrategia de análisis es hacer una ejemplificación de su uso y, a partir de la misma, describir sus ventajas y limitaciones en el desarrollo de los Modelos Multinivel. Los resultados revelan que, aunque esencialmente muestran los mismos resultados, existen diferencias en torno a la cantidad de niveles de análisis que soportan, el tratamiento de los residuos, o el método de estimación de los componentes de la varianza asignado por defecto. Con todo ello, el investigador podrá guiar sus pasos en el desarrollo de los Modelos Multinivel y tomar una decisión fundamentada en torno a qué programa utilizar en función de las necesidades propias de su estudio.

- **Multilevel models applications to the analysis of longitudinal data** (García de Yébenes, Zunzunequi, Mathieu , Rodríguez Lazo, & Otero, 2002)

Este trabajo es una introducción al análisis de medidas repetidas en estudios longitudinales. Se utiliza un marco analítico con dos etapas, ajustando modelos jerárquicos lineales con dos niveles. El primer nivel corresponde a la ocasión (tiempo) de medida y el segundo al individuo.

Estos modelos estadísticos proceden de las ciencias sociales, en las que se han utilizado durante más de 25 años para analizar datos en organizaciones con múltiples niveles. Su aplicación permite estudiar los cambios en alguna característica de interés (estado de salud o factor de riesgo) y analizar las circunstancias que explican la variabilidad en las trayectorias individuales. En este trabajo se introducen los conceptos básicos de este método: variabilidad entre individuos y dentro de cada individuo a lo largo del tiempo, modelo del nivel individual para describir la trayectoria de cada individuo y modelo «entre individuos» para describir cómo cambian las trayectorias entre individuos, efectos fijos y efectos aleatorios, modelos decrecimiento lineal y cuadrático. Para ello se ha realizado un análisis de los cambios en la función cognitiva de una cohorte de personas mayores, el estudio «Envejecer en Leganés», seguida cada dos años, entre 1993 y 1999. Se presentan los resultados de modelos ajustados para resolver las preguntas de investigación más frecuentes en la descripción y el análisis de las trayectorias de cambio individual. Por último, se comentan posibles generalizaciones de estos modelos lineales jerárquicos a situaciones en las que la variable de interés no es continua, como es el caso de las variables dependientes dicotómicas, nominales u ordinales.

Las variables independientes utilizadas en este estudio han sido la edad, el sexo y el nivel de instrucción (analfabetos, sin escolarización, primaria incompleta y primaria completa).

- **A Mixed Model Approach for Intent-to-Treat Analysis in Longitudinal Clinical Trials with Missing Values** (Hrishikesh & Hong , 2009)

El principal problema que se plantea con la falta de datos es que la distribución de los datos observados no puede ser la misma que la distribución de los datos completos.

Algunos valores perdidos pueden estar o no relacionados con las respuestas observadas. Little & Rubin (1987) clasificaron los mecanismos de los valores perdidos como por completo al azar (MCAR), perdidos al azar (MAR),y no

perdidos al azar (NMAR). MCAR es una condición en la que los valores que faltan son al azar distribuido a través de todas las observaciones. MAR es una condición en la que los valores que faltan no son al azar distribuido a través de todas las observaciones, pero son al azar distribuido en una o más submuestras. Bajo MCAR y MAR, los mecanismos de datos faltantes son a menudo referido como "ignorables"; por el contrario, la falta mecanismo de datos NMAR se refiere a menudo como "no ignorable".

En este informe, se lleva a cabo una investigación detallada de estudios de simulación para desarrollar recomendaciones para análisis ITT de longitudinal ensayo clínico controlado datos con valores perdidos. Se comparan los estimados, los tamaños (errores de tipo I), y el poder entre varios enfoques temáticos y del modelo mixto lineal general enfoque (GLMM) para diferentes proporciones de valores faltantes.

1.2. Aspectos generales

1.2.1. Historia de los modelos de regresión multinivel

El análisis multinivel es una técnica relativamente nueva estadística en investigación en ciencias sociales, aunque sus raíces se remontan a los estudios sociológicos clásicos, en especial estudio del suicidio de Durkheim. Durkheim buscó las causas del suicidio, un fenómeno muy personal e individual, en los contextos sociales del individuo. El análisis multinivel puede ser visto como una forma moderna de hacer frente a cuestiones de investigación relativos a cómo los resultados a nivel individual pueden verse como el resultado de la interacción entre factores individuales y contextuales.

El primer paso para el análisis multinivel moderna fue el aumento del contexto de análisis en los EE.UU. en la década de 1940. El análisis contextual se introdujo como una crítica del micro-perspectiva dominante en la sociología americana. El análisis contextual se hizo más establecido en la década de 1960, las técnicas estadísticas se hicieron más sofisticadas y se avanzó conceptual. Los

conceptos de proposiciones contextuales, tipología de variables por niveles y de los efectos estructurales fueron las contribuciones más influyentes.

Alrededor de 1970, el análisis contextual fue muy criticado por Robert Hauser. Mantuvo que la mayoría de supuestos de efectos contextuales carecían de sustancia y eran artefactos de modelos a nivel individual inadecuadamente especificados. En cambio, de los "efectos contextuales" se agruparon los efectos individuales. Hauser utiliza la falacia contextual término para describir este fenómeno.

Desde finales de la década de 1970, los pasos cruciales en el desarrollo de análisis multinivel se llevaron a cabo en la investigación de la escuela. Los datos sobre educación, principalmente habían sido analizados a nivel individual, haciendo caso omiso de las escuelas. Un paso innovador fue analizar cada escuela por separado. La variable dependiente podría ser una variable de resultado, tales como la puntuación en una prueba de matemáticas con las variables explicativas a nivel individual, como el sexo y el nivel socioeconómico de los padres. La estimación de modelos de regresión idénticos para cada escuela produciría un conjunto de intersecciones y los coeficientes de regresión que podrían mostrar variación sistemática por las escuelas. Esto llevó a la aproximación de esquí-como-resultados. Las pendientes (coeficientes de regresión) fueron consideradas como variables dependientes en un análisis a nivel de la escuela, con las variables explicativas a nivel escolar. Este enfoque puede ser visto como un diseño de regresión múltiple de dos etapas.

En la década de 1980, se han desarrollado varias variaciones de los modelos multinivel para evitar los problemas estadísticos del diseño de dos etapas. En Chicago, un grupo de investigadores ha desarrollado el software HLM para la estimación simultánea de "modelos lineales jerárquicos 'con dos niveles. En Londres, otros grupos de investigadores educativos desarrollaron otro programa de software para el análisis multinivel, ahora conocido como MLwiN.

En la investigación de encuesta tradicional, el individuo se ve a menudo al margen de sus contextos, mientras que la investigación cualitativa hace hincapié

en la contextualización. El análisis multinivel puede ser visto como una herramienta para la contextualización análisis estadístico cuantitativo. En muchos campos de la investigación, los procesos que han de estudiarse involucran a dos o más niveles de análisis. El aprendizaje tiene lugar en las escuelas y que no le parece bien hacer caso omiso de este hecho en el estudio de los resultados del aprendizaje. Otro ejemplo es el del campo de la salud pública, donde la famosa hipótesis de Wilkinson implica que la desigualdad del ingreso en una comunidad puede influir en la salud de los habitantes. Un tercer ejemplo es el estudio de determinación de los salarios, donde los salarios de los empleados pueden ser vistos como depende de la rentabilidad de la empresa, así como en el capital humano de los empleados. En estos ejemplos, preguntas de investigación surgen de una combinación de teorías a nivel individual y contextual.

CAPÍTULO II

PLANTEAMIENTO DEL PROBLEMA

2.1. Descripción del problema

Se ha demostrado que los datos truncados y censurados afectan la estimación de los parámetros por MCO en los modelos de regresión lineal. También se sabe que los modelos multinivel constituyen la metodología de análisis más adecuada para tratar datos “jerarquizados” o “anidados” (por ejemplo, los estudiantes en aulas, o las aulas en escuelas). Así, además de mejorar la calidad de los resultados, posibilita realizar análisis novedosos, tales como estimar la aportación de cada nivel de análisis o las interacciones entre variables de distintos niveles. Los modelos multinivel son, en esencia, ampliaciones de los modelos de regresión lineal clásicos; ampliaciones mediante las cuales se elaboran varios modelos de regresión para cada nivel de análisis. Lo que se hizo en esta investigación es estudiar el efecto marginal de las variables truncadas y censuradas en los modelos de regresión multinivel de 2 etapas en los siguientes

criterios: criterio de información de Akaike corregido (AICC), criterio de información de Akaike consistente (CAIC) y R^2 en los modelos de regresión multinivel de 2 etapas.

2.2. Formulación del problema

2.2.1. Problema general

- ¿Qué tanto afecta (variaciones porcentuales) la variable dependiente truncada y censurada en los siguientes indicadores: criterio de información de Akaike (AIC), criterio de información de Akaike corregido (AICC) y criterio de información de Akaike consistente (CAIC) en los modelos de regresión multinivel de 2 etapas?

2.2.2. Problemas específicos

- ¿Qué tanto afecta (variaciones porcentuales) la variable dependiente truncada en los siguientes indicadores: criterio de información de Akaike (AIC), criterio de información de Akaike corregido (AICC) y criterio de información de Akaike consistente (CAIC) en los modelos de regresión multinivel de 2 etapas?

- ¿Qué tanto afecta (variaciones porcentuales) la variable dependiente censurada en los siguientes indicadores: criterio de información de Akaike (AIC), criterio de información de Akaike corregido (AICC) y criterio de información de Akaike consistente (CAIC) en los modelos de regresión multinivel de 2 etapas?

2.3. Objetivos

2.3.1. Objetivo general

- Determinar el efecto (variaciones porcentuales) del truncamiento y censura de la variable dependiente en los siguientes indicadores: criterio de información de Akaike (AIC), criterio de información de Akaike corregido (AICC) y criterio de información de Akaike consistente (CAIC) en los modelos de regresión multinivel de 2 etapas.

2.2.2. Objetivos específicos

- Determinar el efecto marginal (variaciones porcentuales) de la variable dependiente truncada (sólo sobre observaciones no censuradas) en los siguientes indicadores: criterio de información de Akaike (AIC), criterio de información de Akaike corregido (AICC) y criterio de información de Akaike consistente (CAIC) en los modelos de regresión multinivel de 2 etapas.

- Determinar el efecto marginal (variaciones porcentuales) de la variable dependiente censurada (sobre observaciones censuradas y no censuradas) en los siguientes indicadores: criterio de información de Akaike (AIC), criterio de información de Akaike corregido (AICC) y criterio de información de Akaike consistente (CAIC) en los modelos de regresión multinivel de 2 etapas.

2.4. Hipótesis

2.4.1. Hipótesis general

- La variable dependiente truncada y censurada afecta en más del 50% los siguientes indicadores: criterio de información de Akaike corregido (AICC), criterio de información de Akaike consistente (CAIC) y R^2 en los modelos de regresión multinivel de 2 etapas la estimación de parámetros y correlación intraclase en los modelos de regresión multinivel de 2 etapas.

2.4.2. Hipótesis específicas

- La variable dependiente truncada (sólo sobre observaciones no censuradas) afecta significativamente en los siguientes indicadores: criterio de información de Akaike corregido (AICC), criterio de información de Akaike consistente (CAIC) y R^2del modelo completo en los modelos de regresión multinivel de 2 etapas con una variación porcentual mayor al 50% con respecto a los indicadores obtenidos con la base de datos completa.

- La variable dependiente censurada (sobre observaciones censuradas y no censuradas) afecta significativamente en los siguientes indicadores: criterio de información de Akaike corregido (AICC), criterio de información de Akaike consistente (CAIC) y R^2del modelo completo en los modelos de regresión multinivel de 2 etapas con una variación porcentual mayor al 50% con respecto a los indicadores obtenidos con la base de datos completa.

2.5. Justificación

Los modelos de regresión multinivel son muy utilizados en el campo de la educación. Aitkin y Longford (1986) propusieron los modelos de regresión multinivel que han marcado la investigación educativa, pues estos reconocen y manejan la organización jerárquica de los sistemas educativos (estudiantes en aula, aulas en escuelas, escuelas en países) y ofrecen resultados con una menor incidencia de los errores de estimación (p.ej. Goldstein, 2003; Raudenbush&Bryk, 2002). Sin embargo, poco se ha estudiado el efecto que pueden traer tanto los datos censurados como los truncados en las estimaciones. Por lo tanto, esta investigación beneficiará grandemente a los investigadores en el área educativa. No solo a los ministerios de educación que realizan investigación sino también a otras entidades como por ejemplo la Organización de las Naciones Unidas para la Educación, la Ciencia y la Cultura bajo el marco de acción de Dakar (2000-2015) que monitorea el avance de la educación mundial; por lo tanto, los datos censurados y truncados en una estructura jerárquica no son ajenos a los estudios que realizan.

Las variables truncadas y censuradas han sido medianamente estudiadas en los modelos de regresión lineal; sin embargo, los modelos de regresión multinivel, que pueden ser considerados como ampliaciones de los modelos de regresión lineal clásicos (F. Javier Murillo Torrecilla, Director de la Revista Iberoamericana de Evaluación Educativa, Agosto 2008) no han sido estudiados a profundidad aun cuando la variable dependiente está sujeta a censura o truncamiento. Por ende, se considera que estudiar los efectos que este tipo de casos en la variable dependiente traen en indicadores como criterio de información de Akaike corregido (AICC), criterio de información de Akaike consistente (CAIC) y R^2de los modelos de regresión multinivel de 2 etapas puede llenar un pequeño vacío en el conocimiento de la estadística con respecto a los modelos multinivel.

Los resultados presentados en la investigación son obtenidos bajo la simulación de datos aleatorios en diversos tamaños muestrales y bloques; por

ende, se considera que sus conclusiones pueden ser generalizadas a casos ya expuestos.

2.6. Matriz de Consistencia

CUADRO Nº 2.1.

MATRIZ DE CONSISTENCIA

PROBLEMA	OBJETIVOS	HIPÓTESIS	JUSTIFICACIÓN
PROBLEMA GENERAL	**OBJETIVO GENERAL**	**HIPÓTESIS GENERAL**	
· ¿Qué tanto afecta la variable dependiente truncada y censurada en los siguientes indicadores: criterio de información de Akaike corregido (AICC), criterio de información de Akaike consistente (CAIC) y R^2 en los modelos de regresión multinivel de 2 etapas?	· Determinar el efecto del truncamiento y censura de la variable dependiente en los siguientes indicadores: criterio de información de Akaike corregido (AICC), criterio de información de Akaike consistente (CAIC) y R^2 en los modelos de regresión multinivel de 2 etapas.	· La variable dependiente truncada y censurada afecta en más del 50% los siguientes indicadores: criterio de información de Akaike corregido (AICC), criterio de información de Akaike consistente (CAIC) y R^2 en los modelos de regresión multinivel de 2 etapas la estimación de parámetros y correlación intraclase en los modelos de regresión multinivel de 2 etapas.	Los modelos de regresión multinivel son muy utilizados en el campo de la educación. Aitkin y Longford (1986) propusieron los modelos de regresión multinivel que han marcado la investigación educativa, pues estos reconocen y manejan la organización jerárquica de los sistemas educati-vos (estudiantes en aula, aulas en escuelas, escuelas en países) y ofrecen resultados con una menor incidencia de los errores de estimación (p.ej. Goldstein, 2003; Raudenbush & Bryk, 2002). Sin embargo, poco se ha estudiado el efecto que pueden traer tanto los datos censurados como los truncados en las estimaciones. Por lo tanto, esta investigación beneficiará grandemente a los investigadores en el área educativa. No solo a los ministerios de educación que realizan investigación sino también a otras entidades como por ejemplo la Organización de las Naciones Unidas para la Educación, la Ciencia y la Cultura bajo el marco de acción de Dakar (2000-2015) que monitorea el avance de la educación mundial; por lo tanto, los datos censurados y truncados en una estructura jerárquica no son ajenos a los estudios que realizan.
PROBLEMAS	**OBJETIVOS**	**HIPÓTESISS**	
· ¿Qué tanto afecta (variaciones porcentuales) la variable dependiente truncada en los siguientes indicadores: criterio de información de Akaike corregido (AICC), criterio de información de Akaike consistente (CAIC) y R^2 en los modelos de regresión multinivel de 2 etapas?	· Determinar el efecto marginal de la variable dependiente truncada (sólo sobre observaciones no censuradas) en los siguientes indicadores: criterio de información de Akaike corregido (AICC), criterio de información de Akaike consistente (CAIC) y R^2en los modelos de regresión multinivel de 2 etapas.	· La variable dependiente truncada (sólo sobre observaciones no censuradas) afecta significativamente en los siguientes indicadores: criterio de información de Akaike corregido (AICC), criterio de información de Akaike consistente (CAIC) y R^2 del modelo completo en los modelos de regresión multinivel de 2 etapas con una variación porcentual mayor al 50% con respecto a los indicadores obtenidos con la base de datos completa.	
· ¿Qué tanto afecta (variaciones porcentuales) la variable dependiente censurada en los siguientes indicadores: criterio de información de Akaike corregido (AICC), criterio de información de Akaike consistente (CAIC) y R^2 en los modelos de regresión multinivel de 2 etapas?	· Determinar el efecto marginal de la variable dependiente censurada (sobre observaciones censuradas y no censuradas) en los siguientes indicadores: criterio de información de Akaike corregido (AICC), criterio de información de Akaike consistente (CAIC) y R^2en los modelos de regresión multinivel de 2 etapas.	· La variable dependiente censurada (sobre observaciones censuradas y no censuradas) afecta significativamente en los siguientes indicadores: criterio de información de Akaike corregido (AICC), criterio de información de Akaike consistente (CAIC) y R^2del modelo completo en los modelos de regresión multinivel de 2 etapascon una variación porcentual mayor al 50% con respecto a los indicadores obtenidos con la base de datos completa.	

ELABORACIÓN: PROPIA FECHA: Octubre 2016

CAPÍTULO III

MARCO TEÓRICO

3.1. Técnicas a usar

3.1.1. Modelo de regresión multinivel

A) Definiciones

A1) Modelo de regresión multinivel

El modelo de regresión multinivel se ha hecho conocido en la literatura de los investigadores bajo una variedad de nombres, tales como Modelo de coeficientes aleatorios dado por Leeuw & Kreft, 1986; Longford 1993, Modelo de componentes de varianza dado por Longford 1986, Modelos lineales jerárquicos dado por Raudenbush & Bryk en 1986, 1992.

El modelo de regresión multinivel completo asume que hay un conjunto de datos jerárquicos, con una sola variable dependiente que es medida en el nivel más bajo y variables explicativas que existen en todos los niveles. Conceptualmente el modelo puede ser visto como un sistema jerárquico de ecuaciones de regresión. Los modelos de regresión multinivel son, en esencia,

ampliaciones de los modelos de regresión lineal clásicos; ampliaciones mediante las cuales se elaboran varios modelos de regresión para cada nivel de análisis(Murillo, 2008). Con ello los modelos del primer nivel están relacionados por un modelo de segundo nivel en el que los coeficientes de regresión del nivel 1 se regresan en un segundo nivel de variables explicativas, y así sucesivamente para los diferentes niveles.

Un problema multinivel concierne a una población con estructura jerárquica. Una muestra de tal población puede ser descrita como una muestra multicéntrica (De la Cruz, 2008): primero tomamos una muestra de unidades del más alto nivel (por ejemplo, hospitales), y luego muestreamos las subunidades de las unidades disponibles (pacientes dentro de los hospitales). En tales muestras, las observaciones individuales no son completamente independientes. Por ejemplo, los pacientes en el mismo hospital tienden a ser similares entre sí, ya que pueden proceder de las mismas áreas geográficas y por tanto coincidir en varios aspectos. Como un resultado, la correlación promedio (expresada en la llamada correlación intraclase) entre las variables medidas en los pacientes del mismo hospital será más alto que la correlación promedio de las variables medidas en los pacientes de los diferentes hospitales. Las pruebas estadísticas estándar se inclinan fuertemente en la suposición de independencia de las observaciones. Si esta suposición es violada (y en los datos multinivel esto es usualmente el caso) los estimadores de los errores estándares de las pruebas estadísticas convencionales son mucho más pequeñas, y estos resultados son falsamente significativos(De la Cruz, 2008).

El modelo de regresión multinivel completo asume que hay un conjunto de datos jerárquicos, con una sola variable dependiente que es medida en el nivel más bajo y variables explicativas que existen en todos los niveles. Conceptualmente el modelo puede ser visto como un sistema jerárquico de ecuaciones de regresión.

En la parte del análisis, nos podemos encontrar con un conjunto de problemas conceptuales. Si el análisis no es muy cuidadoso en la interpretación

de los resultados, podemos cometer la falacia del error del nivel, el cual consiste en analizar los datos en un nivel, y extraer conclusiones de otro nivel. Probablemente la falacia mejor conocida es la falacia ecológica, ésta pretende deducir relaciones para los individuos o nivel 1, cuando los resultados contextuales o nivel 2 no reproducen necesariamente al nivel individual. Existe otro tipo de falacia llamada falacia atomista, la cual propone las mismas asociaciones encontradas a nivel individual o nivel 1 como relaciones a nivel contextual o nivel 2.

Antes de profundizar mínimamente en el desarrollo formal de los modelos multinivel vamos a prestar atención a conceptos fundamentales y sus implicaciones: correlación intraclase, coeficiente fijo, coeficiente aleatorio, e interacción internivel.

A2) Problemas que resuelven los modelos de regresión multinivel

- Manejan la falta de independencia (Journal of Epidemiology, Community Health, 2001)
  - Estimación MCO ineficiente
  - Significaciones espurias (ej. 10 pacientes, a cada uno le medimos mensualmente durante 12 meses).

- Evitan falacias por interpretar efectos a nivel equivocado (Journal of Epidemiology, Community Health, 2001)
  - Ecológica (interpretar datos agregados a nivel individual)
  - Atomística (interpretar datos individuales a nivel agregado)

- Estiman el efecto de las variables “explicativas” (efectos fijos) de ambos niveles, incluyendo interacciones entre niveles (Chakraborty H, 2009)

- Estiman qué parte de la variabilidad no “explicada” (efectos aleatorios) es imputable a cada nivel (Chakraborty H, 2009)

A3) Correlación intraclase

Se entiende por correlación intraclase o autocorrelación la medida del grado de dependencia de los individuos. Es decir, es una estimación de lo que comparten, por ejemplo, los alumnos por estudiar en una misma clase o centro. Una correlación baja o cercana a cero significará que los sujetos dentro del mismo grupo son tan diferentes entre sí como los que pertenecen a otros grupos. En ese caso, la agrupación no tiene consecuencias, los grupos no son homogéneos internamente y las observaciones son independientes (requisito necesario dentro de los modelos lineales tradicionales). Si se ignora la presencia de esta correlación intraclase, los modelos resultantes son innecesaria y falsamente complejos, dado que aparecen relaciones significativas inexistentes.(Murillo, 2008)

A4) Coeficiente fijo

En los modelos de regresión clásicos los parámetros que se estiman son el intercepto (o punto de corte) y las pendientes. Desde una perspectiva clásica, estos coeficientes se asumen como fijos, es decir, comunes a todos los sujetos y son estimados a partir de los datos.(Murillo, 2008).

En los modelos multinivel se permite a los grupos desviarse de la solución central o global, tanto en el intercepto como en la pendiente. O, lo que es lo mismo, los modelos multinivel están compuestos por dos partes, una general, común a todos los contextos, que es la llamada parte fija, y otra que representa lo específico de cada contexto, que varía y que se estima a través de la varianza en los distintos niveles.(Murillo, 2008)

A5) Coeficiente aleatorio

En los modelos multinivel como unidades (agrupamientos) que definen los niveles, son vistos como efectos aleatorios, de esta forma, como muestras aleatorias de una población de estas unidades (como escuelas, centros de salud, domicilios, etc.). Estos efectos aleatorios se traducen en un modelo de coeficientes aleatorios que van a tomar en cuenta la variabilidad entre agrupamientos, desde formas simples, a través de variabilidad a nivel del intercepto, o de formas más complejas, a través de variabilidades a niveles de inclinaciones de dos rectas.(De la Cruz, 2008).

A6) Interacción internivel

La interacción internivel o la interacción entre variables que están medidas en diferentes niveles de una estructura jerárquica de datos. Ello hace referencia a la interacción que puede haber entre variables de diferentes niveles, por ejemplo, determinada metodología docente puede ser mejor con ciertos estudiantes (el llamado efecto Aptitude TreatmentInteraction-ATI), o un estilo directivo con profesores de determinadas características. La comprobación de este tipo de hipótesis necesita un modelo de análisis que no sólo dé cuenta de la estructura jerárquica de los datos, sino que también permita estimar las interacciones interniveles. (Murillo, 2008).

B) Procedimiento Formal del Modelo de regresión de 2 niveles

Como se ha señalado, los modelos multinivel son, en esencia, ampliaciones de los modelos de regresión lineal clásicos, de tal forma que en realidad son varios modelos lineales para cada nivel. Así, los modelos del primer nivel están relacionados con uno de segundo nivel en el que los coeficientes de regresión del nivel 1 se regresan en un segundo nivel de variables explicativas y así sucesivamente para los diferentes niveles.

A continuación, se presenta una ecuación de regresión lineal sencilla con dos variables independientes:

$$y_i = \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + \varepsilon_i$$

Si se permite que el intercepto pueda tomar diferentes valores en función de un segundo nivel, la ecuación quedará:

$$y_{ij} = \beta_{0j} + \beta_1 x_{1ij} + \beta_2 x_{2ij} + \varepsilon_{ij}$$

$$\beta_{0j} = \beta_0 + \mu_{0j}$$

Donde:

y_{ij}es la variable respuesta que tiene un alumno i en una escuela j

ε_{ij}es el error y se distribuye normalmente con una varianza constante e igual a σ^2_{e0},

β_{0j}es el promedio de$y$de la escuela j-ésima

$\beta_0$representa el “gran promedio” de y para la población, y

μ_{0j}es el efecto aleatorio asociado a la escuela j-ésima y se supone que tiene media cero y una varianza $\sigma^2_{\mu 0}$.

$$\mu_{0j} \sim N(0, \sigma^2_{\mu 0}), \quad \varepsilon_{ij} \sim N(0, \sigma^2_{e0})$$

$$\mathrm{Cov}(\mu_{0j}, \varepsilon_{ij}) = 0$$

$$\mathrm{Cov}(y_{i_1 j}, y_{i_2 j} / x_{ij}) = \sigma^2_{\mu 0} \geq 0$$

Dondei_1, i_2 son dos alumnos en la misma escuela j con, en general, una covarianza positiva entre las respuestas. Esta ausencia de independencia, partiendo de estos dos orígenes de variación en diferentes niveles de los datos jerárquicos (alumnos y escuelas) contradice la suposición del modelo lineal tradicional y nos conduce a considerar una nueva clase de modelos.

Si, además de hacer variar el intercepto, permitimos que las pendientes sean diferentes para cada escuela, tenemos la siguiente ecuación:

Nivel 1: $\beta_{0j} + \beta_{1j} x_{1ij} + \beta_{2j} x_{2ij} + \varepsilon_{ij}$

Nivel 2:$\beta_{0j} = \beta_0 + \mu_{0j}; \beta_{1j} = \beta_1 + \mu_{1j}; \beta_{2j} = \beta_2 + \mu_{2j}$

Con

$$\begin{bmatrix} \mu_{0j} \\ \mu_{1j} \\ \mu_{2j} \end{bmatrix} \sim N(0, \Omega_\mu) \; : \; \Omega_\mu = \begin{bmatrix} \sigma^2_{\mu 0} & & \\ \sigma^2_{\mu 10} & \sigma^2_{\mu 1} & \\ \sigma^2_{\mu 20} & \sigma^2_{\mu 21} & \sigma^2_{\mu 2} \end{bmatrix}$$

$$[e_{0ij}] \sim N(0, \Omega_e) \quad : \quad \Omega_e = [\sigma^2_{e0}]$$

B1) Modelo nulo

El modelo nulo o modelo vacío es el punto de partida del proceso modelado. Contiene únicamente una variable respuesta y la constante (o intercepto o punto de corte), es decir, ninguna variable predictora. De esta forma, el modelo posee efectos aleatorios en los dos niveles y no incluye variables explicativas en ninguno de ellos. El modelo nulo se establece como línea de base para la estimación de la varianza explicada a partir de la cual se van evaluando las aportaciones de modelos más elaborados.

Para el caso de modelo con dos niveles, la ecuación sería:

Nivel 1:$y_{ij} = \beta_{0j} + \varepsilon_{ij}$ (1)

Donde:

y_{ij}es la variable respuesta que tiene un alumno i en una escuela j

ε_{ij}es el error y se distribuye normalmente con una varianza constante e igual a σ_{e0}^2,

β_{0j}es el promedio de y de la escuela j-ésima.

Es decir, $\beta_{0j} = \mu_{yj}$

Nivel 2: $\beta_{0j} = \beta_0 + \mu_{0j}$ (2)

Donde:

$\beta_0$representa el “gran promedio” de y para la población

μ_{0j}es el efecto aleatorio asociado a la escuela j-ésima y se supone que tiene media cero y una varianza $\sigma_{\mu 0}^2$.

Entonces,$y_{ij} = \beta_0 + \mu_{0j} + \varepsilon_{ij}$ (3)

El modelo de la ecuación (3) no explica alguna varianza, sólo descompone la varianza en dos componentes independientes: σ_{e0}^2, el cual es la varianza del error del más bajo nivel ε_{ij}, y $\sigma_{\mu 0}^2$, la varianza del error del nivel más alto μ_j. Usando este modelo podemos estimar la correlación intra clase ρpor la ecuación:

$\rho = \sigma_{\mu 0}^2 / (\sigma_{\mu 0}^2 + \sigma_{e0}^2)$ (4)

La correlación intraclase ρ es un estimador de la proporción de varianza explicada en la población. La ecuación (4) establece que la correlación intraclase es igual a la proporción estimada de la varianza del nivel grupo comparada con la varianza total estimada.

La razón de verosimilitud: $-2\log_e(verosimilitud)$ servirá para ir evaluando las diferentes aportaciones al modelo.

B2) Modelo con variables explicativas

La segunda fase es la estimación del modelo con variables de ajuste. Este modelo se construye a partir del modelo nulo pero incorporándole, tanto en la

parte fija como en la aleatoria, las variables consideradas en el trabajo como variables de ajuste.

De esta forma, para el caso de tres variables de ajuste propias de la etapa 1 y una variable propia de la etapa 2, el Modelo Multinivel que se espera conseguir es el siguiente:

Nivel 1:

$$y_{ij} = \beta_{0j} + \beta_{1j}x_{1ij} + \beta_{2j}x_{2ij} + \beta_{3j}x_{3ij} + \beta_4 x_{4j} + \varepsilon_{ij} =$$

$$= \beta_{0j} + \sum_{l=1}^{3} \beta_{1j}x_{1ij} + \beta_4 \ x_{4j} + \varepsilon_{ij} \qquad (5)$$

Nivel 2:

$$\beta_{0j} = \beta_0 + \mu_{0j} \ ; \ \beta_{1j} = \beta_1 + \mu_{1j} \ ; \ \beta_{2j} = \beta_2 + \mu_{2j} \ ; \ \beta_{3j} = \beta_3 + \mu_{3j}$$

Con

$$\begin{vmatrix} \mu_{0j} \\ \mu_{1j} \\ \mu_{2j} \\ \mu_{3j} \end{vmatrix} \sim N\,(0, \Omega_u) : \ \Omega_u = \begin{bmatrix} \sigma^2_{\mu 0} & & & \\ \sigma^2_{\mu 10} & \sigma^2_{\mu 1} & & \\ \sigma^2_{\mu 20} & \sigma^2_{\mu 21} & \sigma^2_{\mu 2} & \\ \sigma^2_{\mu 30} & \sigma^2_{\mu 31} & \sigma^2_{\mu 32} & \sigma^2_{\mu 3} \end{bmatrix}$$

$$\left[e_{0ij}\right] \sim N(0, \Omega_e) : \Omega_e = \ [\sigma^2_{e0}]$$

Los pasos a seguir para estimar son los siguientes:

De esta forma, las ecuaciones se convertirán en:

Nivel 1: $y_{ij} = \beta_{0j} + \beta_1 x_{ij} + \beta_2 x_{2ij} + \beta_3 x_{3ij} + \beta_4 x_{4j} + \varepsilon_{ij} \qquad (6)$

Nivel 2: $\beta_{0j} = \beta_0 + \mu_{0j}$

Donde β_0 es la ordenada promedio de las unidades de nivel 2, $\beta_1\,\beta_3\,\beta_3\ y\,\beta_4$ son las pendientes promedio de la regresión de las unidades de nivel 1, y μ_{0j} es el incremento único del intercepto asociado a la unidad j-ésima del nivel 2.

B3) Modelo Final

La construcción del modelo final incluirá tan sólo aquellas variables que han resultado significativas en la construcción del modelo verificando si se cumplen los supuestos de Homocedasticidad y Normalidad.

C) Supuestos del modelo

Los supuestos son usados para comprobar la adecuación del modelo, las violaciones a estos supuestos pueden provocar un modelo inestable, produciendo resultados totalmente diferentes y en muchos casos conclusiones opuestas al trabajar con distintas muestras.

Los modelos de regresión multinivel, como cualquier modelo de regresión, tienen algunos supuestos de partida, sin cuyo cumplimiento las estimaciones obtenidas no son correctas. El supuesto de independencia de las observaciones no se aplica en estos modelos porque la razón para realizar un análisis multinivel en primer lugar, es que las observaciones de los datos que van a ser analizados están correlacionadas. Los principales supuestos recaen sobre el error del modelo ε, y su certificación se realiza a través del análisis de los residuos $\hat{e}$. Estos supuestos son los siguientes:

C1) Media nula y varianza constante

El error tiene media nula y varianza constante, es decir, el error es homocedástico. Si la varianza de los errores no es constante a lo largo de las observaciones, la regresión es heterocedastica.

Para la comprobar si se cumple o no este supuesto se pueda hacer uso de gráficos de los residuales $e_i$con los valores ajustados correspondientes $\hat{y}_i$,

C2) Supuesto de normalidad

Como ya se ha mencionado, los modelos de regresión multinivel son una extensión de los modelos de regresión tradicionales, por ello, comparten algunos supuestos de aplicación como el hecho de que la variable dependiente se debería distribuir normalmente. (Pérez Fernández, 2013)

Con el cumplimiento del supuesto de normalidad se tiene la justificación teórica para la utilización de pruebas estadísticas que involucren a las distribuciones t, F y χ2 (de uso muy común en la parte inferencial del modelo). No obstante, el supuesto de normalidad puede no ser tan crucial cuando se emplean muestras grandes. Además, una propiedad de la distribución normal es que cualquier función lineal de variables normalmente distribuidas estará también normalmente distribuida.

La literatura referente a probar la normalidad es vasta (White & MacDonald, 1980). Al igual que en el análisis de regresión lineal, se puede comprobar la normalidad por medio de gráficos de normalidad. (Goldstein & Healy, 1995)

Un supuesto adicional en este punto para el análisis de regresión multinivel es que los interceptos y las pendientes aleatorias deben estar normalmente distribuidos

En términos matemáticos:

$[e_{0j}] \sim N(0, \Omega_e) : \ \Omega_e = [\sigma^2_{e0}]$y

$$\begin{bmatrix} \mu_{0j} \\ \mu_{10j} \\ \mu_{20j} \\ \mu_{30j} \end{bmatrix} \sim N(0, \Omega_\mu) : \ \Omega_\mu = \begin{bmatrix} \sigma^2_{\mu 0} & & & \\ \sigma^2_{\mu 10} & \sigma^2_{\mu 1} & & \\ \sigma^2_{\mu 20} & \sigma^2_{\mu 21} & \sigma^2_{\mu 2} & \\ \sigma^2_{\mu 30} & \sigma^2_{\mu 31} & \sigma^2_{\mu 32} & \sigma^2_{\mu 3} \end{bmatrix}$$

D) Estimación de parámetros

Existen varios métodos para la estimar los parámetros en los modelos de regresión multinivel (Martinez, 2014)

Uno de los métodos de estimación es el algoritmo IGLS (mínimos cuadrados generalizados) o RIGLS (mínimos cuadrados generalizados restringidos) descritos por primera vez por Goldstein (1986) y Goldstein (1989), respectivamente. Los algoritmos consideran los niveles del modelo, los coeficientes fijos de la regresión y la matriz de componentes devarianza /covarianza. El modo en el que el software MLwiN ejecuta el algoritmo es el siguiente: el algoritmo fija los componentes de la varianza en algún valor inicial y maximiza la verosimilitud sobre los coeficientes fijos (este es justo el problema de la técnica Mínimos Cuadrados Generalizados). Entonces, fijan los coeficientes por sus actuales valores y maximizan la verosimilitud sobre los componentes de la varianza.

Adicionalmente, desde la década de los años 70, los métodos de estimación de máxima verosimilitud y máxima verosimilitud restringida (MV Y MVR, respectivamente) han sido los más utilizados. MV presenta varias ventajas, incluyendo la habilidad de manejar algunas de las problemáticas de los métodos ANOVA (por ejemplo, la falta de unicidad, o las estimaciones de la varianza negativas). Ambos, MV y MVR, producen idénticos estimadores de efectos fijos. Por un lado MV toma en consideración los grados de libertad desde los efectos fijos y esto produce estimaciones de los componentes de la varianza menos equilibradas. Por otro lado, una desventaja del MVR es que el test ratio de verosimilitud no puede ser usado para comparar dos modelos con diferentes especificaciones en sus componentes fijos. Para muestras pequeñas, MVR es el método idóneo frente a MV, sin embargo, en muestras grandes las diferencias entre utilizar uno y otro son insignificantes (Snijders y Bosker, 1999). Puede decirse, por tanto, que “la utilización de uno u otro método de estimación responde más a una cuestión de gusto personal del investigador” (StataCorp, 2005:188).

El modelo mixto lineal general puede ser re-escrito como un modelo jerárquico (o modelo condicional):

$$\mathbf{y} \mid \mathbf{u} \sim N(\mathbf{X\beta} + \mathbf{Zu}, \mathbf{R})$$
$$\mathbf{u} \sim N(\mathbf{0}, \mathbf{G})$$

Es decir existe un modelo para $\mathbf{y}$ dado $\mathbf{u}$ más un modelo para $\mathbf{u}$. Esto sugiere que existen supuestos específicos sobre la dependencia de la media y la estructura de covarianza sobre las covariables en $\mathbf{X}$ y $\mathbf{Z}$. La media marginal es $\mathbf{X\beta}$ y la estructura de covarianza es $\mathbf{V = ZGZ' + R}$. Es decir que el modelo implicado para la distribución marginal o incondicional de Y es $N(\mathbf{X\beta, ZGZ' + R})$. Esta relación entre ambos modelos no se puede aplicar en general, y depende de propiedades de la distribución normal multivariada y de la linealidad del modelo.

- **Estimación de parámetros vía máximo verosimilitud para modelos mixtos**

Las estimaciones por mínimos cuadrados generalizados pueden usarse para estimar los efectos fijos del modelo mixto. Estas estimaciones se obtienen minimizando $\mathbf{(y - X\beta)'V^{-1}(y - X\beta)}$, y el estimador del vector de efectos fijos $\boldsymbol{\beta}$ es: $\hat{\boldsymbol{\beta}} = (\mathbf{X'V^{-1}X})^{-}\mathbf{X'V^{-1}y}$. Si todas las componentes de varianza en $\mathbf{V}$ son conocidas este estimador es el mejor estimador lineal insesgado (BLUE) y se corresponde con el estimador máximo verosímil. En la práctica del análisis de datos experimentales $\mathbf{V}$ usualmente es desconocida y se reemplaza por su estimador $\hat{\mathbf{V}} = \mathbf{Z\hat{G}Z'} + \hat{\mathbf{R}}$. Si se puede asumir que $\mathbf{u}$ y $\mathbf{e}$ tienen distribución normal, la mejor aproximación para la estimación se logra con métodos basados en máxima verosimilitud. Los métodos de estimación más usados son máxima verosimilitud (ML) y máxima verosimilitud restringida (REML).

La función de verosimilitud, L, puede pensarse como la probabilidad de observar los datos que tenemos si los parámetros del modelo fuesen los

postulados. Se define usando la función de densidad de las observaciones, en este caso la función normal.

La estimación de los parámetros fijos será denotada como $\hat{\boldsymbol{\beta}}_{ML}$ y la de los parámetros de la estructura de varianza como $\hat{\boldsymbol{\xi}}_{ML}$

- **Estimación de parámetros vía máximo verosimilitud restringida para modelos mixtos**

El simple ejemplo del estimador ML de la varianza σ^2 de una muestra aleatoria de variables normales, sugiere que cuando μ no es conocida y debe estimarse, dicha estimación introduce un sesgo en el estimador ML de la varianza. La pregunta entonces es, ¿cómo estimar las componentes de varianza sin tener que estimar los parámetros correspondientes a los efectos fijos? La respuesta conduce al estimador REML, sugerido por Patterson y Thompson (1971). En esta aproximación el vector de efectos fijos es eliminado de la función de verosimilitud, y por lo tanto le llamamos “verosimilitud restringida”, que nos sirve para estimar los parámetros de covarianza. Cuando los datos son balanceados, este método nos da estimadores insesgados iguales a los que nos daría un ANOVA. El estimador ML de $\boldsymbol{\xi}$, basado en $\mathbf{t}$ se llama estimador REML ($\hat{\boldsymbol{\xi}}_{REML}$). La estimación resultante del vector de efectos fijos, $\hat{\boldsymbol{\beta}}(\hat{\boldsymbol{\xi}}_{REML})$ suele denotarse por $\hat{\boldsymbol{\beta}}_{REML}$ y se obtiene usando mínimos cuadrados generalizados.

La idea del estimador REML es la siguiente: Primero se obtiene la verosimilitud basada en datos que en lugar de ser los observados son términos residuales, i.e. $\mathbf{y} - \mathbf{X}\hat{\boldsymbol{\beta}}$. Estos términos son conocidos como residuos completos ya que incluyen todas las fuentes variación aleatoria; se demuestra que los mismos son independientes de $\hat{\beta}$.

- **Propiedades del estimador de efectos fijos**

El estimador de los efectos fijos se obtiene por mínimos cuadrados generalizados usando $\hat{\xi}$ en lugar de ξ para construir $\mathbf{V}$. Si $E(\mathbf{y}) = \mathbf{X\beta}$, condicionando sobre las componentes de varianza. Este estimador es insesgado, i.e. $E(\hat{\boldsymbol{\beta}}(\hat{\xi})) = \boldsymbol{\beta}$. Luego, para obtener estimaciones insesgadas relacionadas a los efectos fijos es suficiente que la media de la respuesta sea correctamente especificada.

Condicionando sobre ξ, el estimador del vector de efectos fijos tiene covarianza independiente de la Var ($\mathbf{y}$), si se asume que la matriz Var($\mathbf{y}$) se modela correctamente como $\mathbf{V = ZGZ' + R}$. Por ello este estimador de covarianza suele llamarse “estimador naif o cándido”. La variabilidad incorporada por reemplazar las componentes de varianza por sus estimadores, no se tiene en cuenta en la construcción del estadístico de Wald que se presenta como candidato para contrastar hipótesis del tipo $H_0 : \mathbf{L\beta} = 0$, donde $\mathbf{L}$ es un arreglo de contrastes conocidos. El estadístico de Wald que se distribuye asintótica mente como una chi-cuadrado con grados de libertad iguales al rango de $\mathbf{L}$, usa la siguiente expresión de varianza:

$$\text{var}(\hat{\boldsymbol{\beta}}) = \left(\mathbf{X'V(\xi)X}\right)^{-1}$$

Luego, la prueba de Wald, solo proveerá de inferencia válida en caso de muestras grandes. Una alternativa práctica es reemplazar la distribución chi-cuadrado por una distribución F apropiada. El estadístico F para la hipótesis que contrasta efectos fijos mediante la matriz de contrastes $\mathbf{L}$, es:

$$F = \frac{\hat{\boldsymbol{\beta}}'\mathbf{L}'\left[\mathbf{L}\left(\mathbf{X}'\mathbf{V}^{-1}(\hat{\xi})\mathbf{X}\right)^{-1}\mathbf{L}'\right]^{-1}\mathbf{L}\hat{\boldsymbol{\beta}}}{rango(\mathbf{L})}$$

Bajo la hipótesis nula, la distribución de F se aproxima a la distribución F con grados de libertad en el numerador igual al rango de L. Los grados de libertad del denominador se estiman desde los datos por diversos métodos: 1) método de

containment (recomendado en modelos con efectos aleatorios y sin modelación de covarianza residual) , 2) aproximación de Sattherthwaite (casos donde existen efectos aleatorios y modelación de covarianza residual), 3) aproximación de Kenward-Roger (casos donde existen efectos aleatorios y modelación de covarianza residual), 4) Between-within (casos donde solo se modelación de covarianza residual; excepto que el tipo sea sin estructura donde se usa solo Between) y 5) Residual. Cuando existen varias observaciones por sujeto, los grados de libertad del denominador son en general muchos por lo que los tres métodos dan valores-p muy parecidos. Cuando la hipótesis es univariada, i.e. el rango de L es uno, la prueba F se reduce a la clásica prueba T.

E) Indicadores de comparación de modelos

Uno de los estadísticos utilizados para comparar modelos de regresión multinivel es la desvianza (-2LL) (McCullag y Nelder, 1989). El resto son modificaciones de -2LL que penalizan (aumentando) su valor mediante alguna función del número de parámetros. El segundo estadístico es AIC es el criterio de información de Akaike(Akaike, 1974); el tercero AICC es el criterio de información de Akaike corregido (Hurvich y Tsai, 1989); el cuarto CAIC es el criterio de información de Akaike consistente (Bozdogan, 1987); y el quinto BIC es el criterio de información bayesiano (Schwarz, 1978).

$$AIC = -2LL + 2d$$

$$AICC = -2LL + \frac{2d}{n - d - 1}$$

$$CAIC = -2LL + d[\log(n) + 1]$$

$$BIC = -2LL + d[log(n)]$$

Donde LL se refiere al logaritmo de la verosimilitud si se utiliza el método deestimación de Máxima Verosimilitud (MV) y al logaritmo de la verosimilitudrestringida si se utiliza el método (MVR).

Si se utiliza MV, d es el número deparámetros asociados a los efectos fijos más el número de parámetros asociados a los efectos aleatorios, y n es e l número total de casos. Si se utiliza MVR, d se trata delnúmero total de parámetros asociados a los efectos aleatorios y n es el número total decasos menos el número de parámetros asociados a los efectos fijos.

Si bien estos estadísticos de ajuste global no tienen una interpretación directa, son muy útiles para comparar modelos alternativos siempre que uno de ellos incluya todos los términos del otro. La diferencia entre los estadísticos $-2LL$correspondientes ados modelos distintos se distribuye según chi-cuadrado con grados de libertad igual alnúmero de parámetros en que difieren los dos modelos comparados, por tanto, ladiferencia entre los estadísticos $-2LL$ correspondientes a dos modelos distintos puede utilizarse para valorar la ganancia que se obtiene en el ajuste cuando se añaden los efectos en que difieren ambos modelos.

Asimismo, también se puede evaluar la calidad del modelo final con otro indicador (Hox, Multilevel Analysis, 2002).

Básicamente lo que nos importa es conocer cuánta varianza del nivel 1 y del nivel 2 es explicada por el modelo. Sería un valor de su capacidad explicativa. Se estima a través del llamado Coeficiente de determinación R^2 (Longford, 1993). Si el intercepto apenas tiene varianza aleatoria la varianza total será la suma de las varianzas de los niveles 1 y 2 ($\hat{\sigma}_e^2 + \hat{\sigma}_{u0}^2$). De esta forma, podremos estimar el coeficiente de determinación total R^2, así como el coeficiente de determinación para el nivel 1, R_1^2, para el 2, R_2^2, con la siguiente fórmula:

$$R^2 = 1 - \frac{var(final)}{var(nulo)}$$

Donde $var(final)$ representa la varianza residual en el modelo cuyo poder explicativo se pretende evaluar a través de R^2 , y $var(nulo)$ es la varianza del modelo nulo.

3.2.2. Datos censurados y truncados

La estimación consistente requiere disponer de una muestra extraída de forma aleatoria y representativa de la población que se pretende estudiar. Además requiere que los estadísticos (estimadores) converjan a los parámetros poblacionales que estiman.

El problema con las muestras surge cuando se refieren a grupo de la población que no representa a la población que es objeto de estudio. En ese caso, los estimadores convergerán a las características de esa subpoblación, no a las de la población que se quiere analizar.(Álvarez, 2008).

Es posible que no observemos datos de la variable dependiente y de las variables explicativas para toda la población. En este caso, tendremos muestras censuradas o truncadas según cómo sea el tipo de limitación en la información disponible.

A) Datos censurados

Una muestra está censurada si los datos se recodifican para un subconjunto de la población (Álvarez, 2008). Para algunas observaciones, sólo se sabe quela variable es mayor (o menor) que un valor.(Albarrán Peréz, 2011).

- Censura por la derecha (o superior)

$$Y = \begin{cases} Y^*, \ Si\ Y^* < U \\ U, \ Si\ Y^* \geq U \end{cases}$$

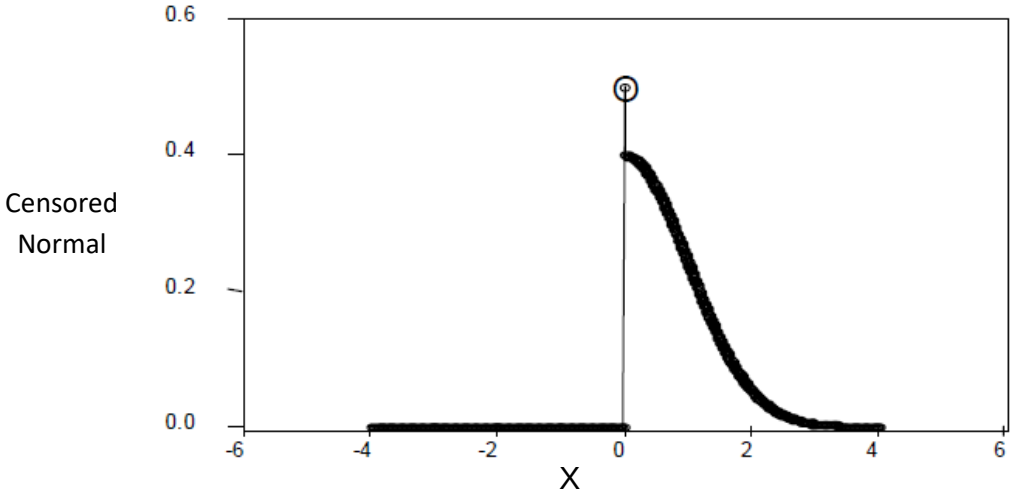
- Censura por la izquierda (o inferior)

$$Y = \begin{cases} Y^*, \ Si\ Y^* > L \\ L, \ Si\ Y^* \leq L \end{cases}$$

- La censura puede producirse por diversos motivos como por ejemplo, resulta del proceso de recogida de datos o se puede interpretar como solución de esquina en una decisión económica.

FIGURA N°3.1

Función de distribución normal con datos censurados



FUENTE: Econometric Methods with Applications in Business and Economics.

FECHA: 2010

B) Datos truncados

Una muestra está truncada si los datos sólo están disponibles para un subconjunto de la población total.

Los valores de las variables explicativas X sólo se observan cuando se observa Y. En otras palabras, los datos truncados excluyen observaciones censuradas.(Albarrán Peréz, 2011)

- Caso de censura por la derecha (o superior)

$$Y = Y^*, \ \ SiY^* < U$$

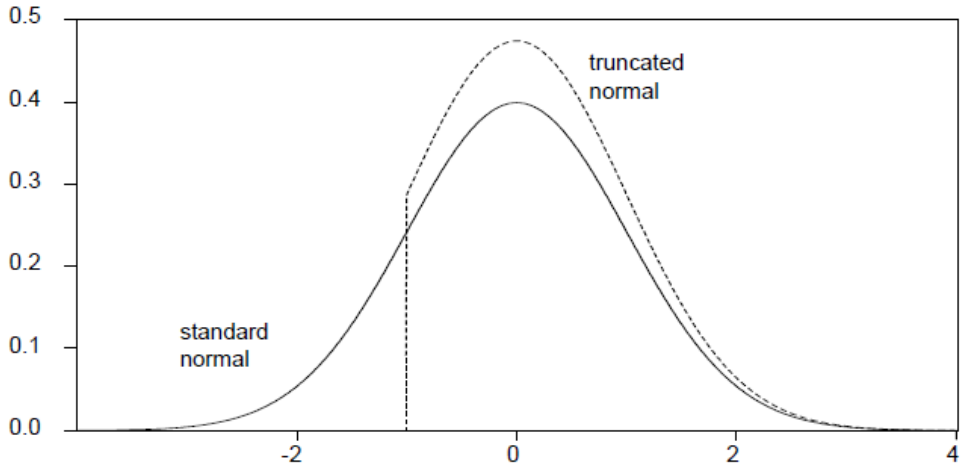
- Caso de censura por la izquierda (o inferior)

$$Y = Y^*, \ \ SiY^* > L$$

La censura/truncamiento puede considerarse como una situación en que falta información (completa) sobre la variable dependiente comparada con observar plenamente Y^*. En algunos casos el valor de censura es desconocido o diferente para cada individuo.

FIGURA N°3.2

Función de distribución normal con datos truncados en



FUENTE: Econometric Methods with Applications in Business and Economics.

FECHA: 2010

3.3. Terminología básica

- **Correlación intraclase**

Es la medida del grado de dependencia de los individuos (autocorrelación).

- **Coeficiente fijo**

Son los parámetros que se estiman en un modelo de regresión lineal clásico

- **Coeficiente aleatorio**

Son los parámetros estimados que toman en cuenta la variabilidad entre agrupamientos

- **Interacción internivel**

Es la interacción entre variables que están medidas en diferentes niveles de una estructura jerárquica de datos

- **Modelo Nulo**

Es el punto de partida del proceso modelado que contiene únicamente una variable respuesta y la constante

- **Datos censurados**

Son datos recodificados para un subconjunto de la población

- **Datos truncados**

Son datos sin considerar los datos censurados

- **Falacia ecológica**

Relaciones para los individuos o nivel 1, cuando los resultados contextuales o nivel 2 no reproducen necesariamente al nivel individual.

- **Falacia atomista**

Propone las mismas asociaciones encontradas a nivel individual o nivel 1 como relaciones a nivel contextual o nivel 2.

CAPÍTULO IV

METODOLOGÍA

4.1. Población en estudio

La población de estudio correspondiente estará determinado por la generación de 2 variables independientes con distribución normal que necesita para la generación de datos la matriz de varianzas y covarianzas que se creará en forma aleatoria. Además, el grado de variación de independencia en las ecuaciones de regresión es controlado por el coeficiente de correlación. (Zuehlke & Kassekert, 2008). Por lo tanto, se considerarán valores por encima de 0.4 para evitar la independencia de los regresores en los diferentes niveles de la base de datos.

En la generación de datos se considerarán 2 grandes escenarios que buscan responder los objetivos antes múltiples casos(Matos, Tsung, Castro, & Lachos, 2016):

- **Escenario 1:** Una muestra de tamaño $n = 200$ y diferentes proporciones de censura y truncamiento como son $10, 20\ y\ 30\%$. Esto teniendo en cuenta que se considera una censura alta desde el 20%.

- **Escenario 2:**Una proporción de censura y truncamiento del 10% y diferentes tamaños de muestra $n = 50, 200 y\ 600$.

4.2. Fuentes de información

Las bases de datos serán simuladas de acuerdo al punto 4.4.

Los escenarios propuestos tienen referencia a la publicación titulada *Heavy-tailed longitudinal regression models for censored data: A likelihood based perspective* elaborado por Larissa A. Matos, Tsung-I Lin, Luis M. Castro y Victor Hugo Lachos Dávila en enero del 2016 en donde se trabaja la censura para modelos de regresión no lineales con 2 variables independientes.

4.3. Definición de variables

Se trabajará con modelos de regresión multinivel de 2 etapas. En cada boque, se trabajará con 2 variables independientes y una variable dependiente de naturaleza continua de la siguiente manera:

$$y_{ij} = \beta_{0j} + \beta_1 x_{1j} + \beta_2 x_{2j} + e_{ij}$$

Donde,

y_{ij} es la variable respuesta continua que tiene el registro i en la categoría j

ε_{ij} es el error y se distribuye normalmente con una varianza constante e igual a 1,

β_{0j}es el vector de parámetros estimados calculado aleatoriamente.

4.4. Procedimientos estadísticos

- Se trabajaron con 2escenarios. Uno con diferentes tamaños de data y un porcentaje de censura y otro con un tamaño de data y diferentes porcentajes y censuras que luego serán truncadas.

- Se crearon bloques (1 será considerado data de control en donde no se truncará ni censurará la variable dependiente, otro bloque contendrá la variable dependiente truncada y otro bloque, censurada).

- Se crearon los parámetros estimados de acuerdo a lo que ya se tiene en los pasos previos.

- Se creó el error de manera aleatoria con media 0 y varianza 1.

- Se usó el test de Mardia para comprobar normalidad de data. Supuesto importante en los modelos de regresión multinivel.

CAPÍTULO V

RESULTADOS

5.1. Análisis descriptivo de las variables

5.1.1. Análisis descriptivo del escenario 1

Una muestra de tamaño $n = 200$ y diferentes proporciones de censura y truncamiento como son $10, 20\ y\ 30\%$. Esto teniendo en cuenta que se considera una censura alta desde el 20%.

- A continuación se describen las distribuciones de las variables independientes, errores y variable dependiente.

- ✓ **Nivel 1**

Este nivel en el modelo de regresión multinivel múltiple es el mismo de un modelo re regresión múltiple.

- ✓ **Nivel 2**

La construcción de las variables que serán analizadas en el modelo de regresión multinivel se deben realizar en el nivel más alto. En la presente tesis, es el nivel 2. Este nivel presenta 2 factores: A y B. Dentro de estos factores se crean las variables independientes X_1 y X_2 que se distribuyen normalmente con media y devianza que se detalla a continuación:

- Factor A:

$$X_1 \sim N(3, 2)$$

$$X_2 \sim N(0, 1.2)$$

- Factor B:

$$X_1 \sim N(6,2)$$

$$X_2 \sim N(2, 3.5)$$

Asimismo, de genera la variable E (Error) bajo una distribución estándar de la siguiente manera: $E \sim N(0, 1)$. Esto para el conjunto total de la base.

Adicionalmente, de consideran los siguientes parámetros en la ecuación: $\beta_0 = 10, \beta_1 = 4\ y\ \beta_2 = 2.5$.

Finalmente, obtenemos la siguiente ecuación:

$$Y = 10 + 4 * X_1 + 2.5 * X_2 + E$$

- Los puntos que reciben Censura y posteriormente de truncamiento en la variable respuesta fueron seleccionados aleatoriamente dentro de cada factor en

el nivel más alto del modelo de regresión multinivel teniendo como criterio de selección que el tamaño de censura sea el mismo en cada factor del nivel 2.

CUADRO Nº 5.1

PROMEDIOS Y DEVIANZAS EN LOS BLOQUES DE CONTROL, CENSURA Y TRUNCAMIENTO PARA ESCENARIO 1

%	BLOQUES	VARIABLES	NIVEL 1		NIVEL 2			
					FACTOR A		FACTOR B	
			MU	SIG	MU	SIG	MU	SIG
0%	CONTROL	X1	4.64	2.40	3.22	1.80	6.06	2.07
		X2	1.07	2.81	-0.05	1.15	2.18	3.47
		Y	31.18	13.52	22.72	8.04	39.64	12.58
10%	CENSURA	X1	4.64	2.40	3.22	1.80	6.06	2.07
		X2	1.07	2.81	-0.05	1.15	2.18	3.47
		Y	27.13	15.84	19.69	10.39	34.58	16.86
	TRUNCAMIENTO	X1	4.50	2.34	3.09	1.81	5.91	1.93
		X2	1.00	2.86	-0.10	1.12	2.10	3.58
		Y	30.49	13.39	22.12	8.19	38.85	12.32
20%	CENSURA	X1	4.64	2.40	3.22	1.80	6.06	2.07
		X2	1.07	2.81	-0.05	1.15	2.18	3.47
		Y	24.85	17.91	17.53	11.41	32.18	20.16
	TRUNCAMIENTO	X1	4.60	2.49	3.06	1.76	6.15	2.13
		X2	1.22	2.91	-0.01	1.14	2.45	3.56
		Y	31.46	14.05	22.19	7.78	40.73	12.76
30%	CENSURA	X1	4.64	2.40	3.22	1.80	6.06	2.07
		X2	1.07	2.81	-0.05	1.15	2.18	3.47
		Y	20.90	17.67	15.63	12.62	26.16	20.30
	TRUNCAMIENTO	X1	4.50	2.33	3.15	1.88	5.84	1.94
		X2	0.93	2.86	0.07	1.18	1.80	3.68
		Y	30.29	12.93	22.65	8.40	37.92	12.16

FUENTE: Huanca Milagros FECHA: Octubre 2016

ELABORACIÓN: Propia

En el cuadro Nº 5.1 se observa la medida de tendencia central Media y la medida de dispersión Devianza en el nivel 1 y en los factores A y B del nivel 2. Esto medido en las variables independientes X1 y X2, y en la variable respuesta Y en los bloques de Control, Censura y Truncamiento del 10%, 20% y 30% de la base.

Si no se considerase un modelo de regresión multinivel, se obviaría lo visiblemente diferentes que son los comportamientos de las variables independiente en los diferentes factores del nivel 2.

5.1.2. Análisis descriptivo del escenario 2

En el escenario 2 se planteó realizar una proporción de censura y truncamiento del 10% y diferentes tamaños de muestra $n = 50, 200 y\ 600$.

- A continuación se describen las distribuciones de las variables independientes, errores y variable dependiente.

✓ **Nivel 1**

Este nivel en el modelo de regresión multinivel múltiple es el mismo de un modelo re regresión múltiple.

✓ **Nivel 2**

La construcción de las variables que serán analizadas en el modelo de regresión multinivel se deben realizar en el nivel más alto. En la presente tesis, es el nivel 2. Este nivel presenta 2 factores: A y B. Dentro de estos factores se crean las variables independientes X_1 y X_2 que se distribuyen normalmente con media y devianza que se detalla a continuación:

- Factor A:

$$X_1 \sim N(3, 1)$$

$$X_2 \sim N(15, 3)$$

- Factor B:

$$X_1 \sim N(5,1)$$

$$X_2 \sim N(10,3)$$

Asimismo, de genera la variable E (Error) bajo una distribución estándar de la siguiente manera: $E \sim N(0,1)$. Esto para el conjunto total de la base.

Adicionalmente, de consideran los siguientes parámetros en la ecuación: $\beta_0 = 10, \beta_1 = 4\ y\ \beta_2 = 2.5$.

Finalmente, obtenemos la siguiente ecuación:

$$Y = 10 + 4 * X_1 + 2.5 * X_2 + E$$

- Los puntos que reciben Censura y posteriormente de truncamiento en la variable respuesta fueron seleccionados aleatoriamente dentro de cada factor en el nivel más alto del modelo de regresión multinivel teniendo como criterio de selección que el tamaño de censura sea el mismo en cada factor del nivel 2.

CUADRO Nº 5.2

PROMEDIOS Y DEVIANZAS EN LOS BLOQUES DE CONTROL, CENSURA Y TRUNCAMIENTO PARA ESCENARIO 2 CON MUESTRA DE 50

%	BLOQUES	VARIABLES	NIVEL 1		NIVEL 2			
					FACTOR A		FACTOR B	
			MU	SIG	MU	SIG	MU	SIG
0%	CONTROL	X1	4.17	1.43	3.17	0.95	5.17	1.11
		X2	12.65	3.37	15.10	2.12	10.21	2.49
		Y	58.15	7.49	60.52	6.43	55.77	7.84
10%	CENSURA	X1	4.17	1.43	3.17	0.95	5.17	1.11
		X2	12.65	3.37	15.10	2.12	10.21	2.49
		Y	51.63	20.46	54.14	21.10	49.11	19.92
	TRUNCAMIENTO	X1	4.26	1.45	3.26	0.97	5.25	1.13
		X2	12.71	3.50	15.35	2.01	10.07	2.53
		Y	58.67	7.38	61.53	5.72	55.81	7.86

FUENTE: Huanca Milagros FECHA: Noviembre 2016

ELABORACIÓN: Propia

En el cuadro Nº 5.2 se observa la medida de tendencia central Media y la medida de dispersión Devianza en el nivel 1 y en los factores A y B del nivel 2 cuando la muestra es de 50 registros. Esto medido en las variables independientes X1 y X2, y en la variable respuesta Y en los bloques de Control, Censura y Truncamiento del 10% de la base.

CUADRO Nº 5.3

PROMEDIOS Y DEVIANZAS EN LOS BLOQUES DE CONTROL, CENSURA Y TRUNCAMIENTO PARA ESCENARIO 2 CON MUESTRA DE 200

%	BLOQUES	VARIABLES		NIVEL 1		NIVEL 2			
						FACTOR A		FACTOR B	
				MU	SIG	MU	SIG	MU	SIG
0%	CONTROL	X1		3.97	1.41	3.03	1.05	4.91	1.04
		X2		12.66	4.13	15.36	2.96	9.96	3.28
		Y		57.41	9.45	60.44	8.51	54.37	9.41
10%	CENSURA	X1		3.97	1.41	3.03	1.05	4.91	1.04
		X2	a	12.66	4.13	15.36	2.96	9.96	3.28
		Y		51.13	20.15	53.66	20.64	48.60	19.42
	TRUNCAMIENTO	X1		4.00	1.42	3.05	1.06	4.95	1.05
		X2		12.64	4.12	15.28	3.01	10.00	3.31
		Y		57.45	9.56	60.30	8.64	54.60	9.63

FUENTE: Huanca Milagros FECHA: Noviembre 2016

ELABORACIÓN: Propia

En el cuadro Nº 5.3 se observa la medida de tendencia central Media y la medida de dispersión Devianza en el nivel 1 y en los factores A y B del nivel 2 cuando la muestra es de 200 registros. Esto medido en las variables independientes X1 y X2, y en la variable respuesta Y en los bloques de Control, Censura y Truncamiento del 10% de la base.

CUADRO Nº 5.4

PROMEDIOS Y DEVIANZAS EN LOS BLOQUES DE CONTROL, CENSURA Y TRUNCAMIENTO PARA ESCENARIO 2 CON MUESTRA DE 600

%	BLOQUES	VARIABLES	NIVEL 1		NIVEL 2			
					FACTOR A		FACTOR B	
			MU	SIG	MU	SIG	MU	SIG
0%	CONTROL	X1	3.99	1.40	3.03	0.96	4.94	1.09
		X2	12.40	4.02	14.97	3.13	9.84	3.07
		Y	56.94	9.10	59.60	8.79	54.28	8.62
10%	CENSURA	X1	3.99	1.40	3.03	0.96	4.94	1.09
		X2	12.40	4.02	14.97	3.13	9.84	3.07
		Y	50.89	19.32	53.32	19.95	48.45	18.40
	TRUNCAMIENTO	X1	3.98	1.39	3.03	0.95	4.94	1.08
		X2	12.33	4.04	14.93	3.13	9.74	3.07
		Y	56.75	9.12	59.46	8.78	54.04	8.64

FUENTE: Huanca Milagros FECHA: Noviembre 2016

ELABORACIÓN: Propia

En el cuadro Nº 5.4 se observa la medida de tendencia central Media y la medida de dispersión Devianza en el nivel 1 y en los factores A y B del nivel 2 cuando la muestra es de 600 registros. Esto medido en las variables independientes X1 y X2, y en la variable respuesta Y en los bloques de Control, Censura y Truncamiento del 10% de la base.

5.2. Análisis de indicadores del modelo

5.2.1. Análisis indicadores del escenario 1

En esta etapa se consideran los 2 escenarios que ya se han comentado en el punto 4.1. y se desarrollará de acuerdo a los pasos expresados en 4.4.

✓ **Escenario 1:** Una muestra de tamaño $n = 200$ y diferentes proporciones de censura y truncamiento como son $10, 20\ y\ 30\%$. Esto teniendo en cuenta que se considera una censura alta desde el 20%.

- Distribuciones de las variables independientes.

- Factor A:

$$X_1 \sim N(3, 2)$$

$$X_2 \sim N(0, 1.2)$$

- Factor B:

$$X_1 \sim N(6, 2)$$

$$X_2 \sim N(2, 3.5)$$

Siguiendo los objetivos de la tesis en donde se busca determinar la variación porcentual de los criterios de información cuando la variable dependiente se encuentra censurada o truncada, se muestran los resultados en el cuadro Nº 5.5 de los valores de criterios de información.

En el cuadro Nº 5.5 vemos que los valores de criterios de información aumentan en como mínimo en 139% en las muestras censuradas siendo este porcentaje cada vez más grande cuando el porcentaje de datos censurados aumenta.

Además, se puede observar que los valores de criterios de información disminuyen como mínimo en 11%en las muestras truncadas siendo este porcentaje cada vez más grande cuando el porcentaje de datos truncados aumenta.

De acuerdo a los criterios de información de Akaike corregido (AICC) y criterio de información de Akaike consistente (CAIC), los modelos de regresión multinivel de 2 etapas son mejores cuando la muestra es truncada en vez de censurada.

CUADRO Nº 5.5

VALORES DE CRITERIOS DE INFORMACIÓN Y VARIACIÓN PORCENTUAL EN LOS BLOQUES DE CONTROL, CENSURA Y TRUNCAMIENTOEN EL ESCENARIO 1

%	BLOQUES	CRITERIOS	VALOR	VAR. %
0%	CONTROL	Logaritmo de la verosimilitud restringido -2	607.96	0%
		Criterio de información Akaike (AIC)	617.96	0%
		Criterio de Hurvich y Tsai (AICC)	618.27	0%
		Criterio de Bozdogan (CAIC)	639.35	0%
		Criterio bayesiano de Schwarz (BIC)	634.35	0%
10%	CENSURA	Logaritmo de la verosimilitud restringido -2	1452.37	139%
		Criterio de información Akaike (AIC)	1462.37	137%
		Criterio de Hurvich y Tsai (AICC)	1462.68	137%
		Criterio de Bozdogan (CAIC)	1483.76	132%
		Criterio bayesiano de Schwarz (BIC)	1478.76	133%
	TRUNCAMIENTO	Logaritmo de la verosimilitud restringido -2	542.64	-11%
		Criterio de información Akaike (AIC)	552.64	-11%
		Criterio de Hurvich y Tsai (AICC)	552.99	-11%
		Criterio de Bozdogan (CAIC)	573.43	-10%
		Criterio bayesiano de Schwarz (BIC)	568.43	-10%
20%	CENSURA	Logaritmo de la verosimilitud restringido -2	1537.80	153%
		Criterio de información Akaike (AIC)	1547.80	150%
		Criterio de Hurvich y Tsai (AICC)	1548.12	150%
		Criterio de Bozdogan (CAIC)	1569.19	145%
		Criterio bayesiano de Schwarz (BIC)	1564.19	147%
	TRUNCAMIENTO	Logaritmo de la verosimilitud restringido -2	486.92	-20%
		Criterio de información Akaike (AIC)	496.92	-20%
		Criterio de Hurvich y Tsai (AICC)	497.33	-20%
		Criterio de Bozdogan (CAIC)	517.11	-19%
		Criterio bayesiano de Schwarz (BIC)	512.11	-19%
30%	CENSURA	Logaritmo de la verosimilitud restringido -2	1605.79	164%
		Criterio de información Akaike (AIC)	1615.79	161%
		Criterio de Hurvich y Tsai (AICC)	1616.10	161%
		Criterio de Bozdogan (CAIC)	1637.18	156%
		Criterio bayesiano de Schwarz (BIC)	1632.18	157%
	TRUNCAMIENTO	Logaritmo de la verosimilitud restringido -2	420.11	-31%
		Criterio de información Akaike (AIC)	430.11	-30%
		Criterio de Hurvich y Tsai (AICC)	430.57	-30%
		Criterio de Bozdogan (CAIC)	449.59	-30%
		Criterio bayesiano de Schwarz (BIC)	444.59	-30%

FUENTE: Huanca Milagros FECHA: Octubre 2016

ELABORACIÓN: Propia

5.2.2. Análisis indicadores del escenario 2

En el escenario 2 se planteó realizar una proporción de censura y truncamiento del 10% y diferentes tamaños de muestra $n = 50, 200 y\ 600$.

- A continuación se describen las distribuciones de las variables independientes, errores y variable dependiente.

✓ **Nivel 1**

Este nivel en el modelo de regresión multinivel múltiple es el mismo de un modelo re regresión múltiple.

✓ **Nivel 2**

La construcción de las variables que serán analizadas en el modelo de regresión multinivel se deben realizar en el nivel más alto. En la presente tesis, es el nivel 2. Este nivel presenta 2 factores: A y B. Dentro de estos factores se crean las variables independientes X_1 y X_2 que se distribuyen normalmente con media y devianza que se detalla a continuación:

- Factor A:

$$X_1 \sim N(3, 1)$$

$$X_2 \sim N(15, 3)$$

- Factor B:

$$X_1 \sim N(5,1)$$

$$X_2 \sim N(10,3)$$

Finalmente, obtenemos la siguiente ecuación:

$$Y = 10 + 4 * X_1 + 2.5 * X_2 + E$$

Siguiendo los objetivos de la tesis en donde se busca determinar la variación porcentual de los criterios de información cuando la variable dependiente se encuentra censurada o truncada, se muestran los resultados en el cuadro Nº 5.6 de los valores de criterios de información.

En el cuadro Nº 5.6 vemos los valores del Criterio de información Akaike (AIC), Criterio de Hurvich y Tsai (AICC), Criterio de Bozdogan (CAIC) y Criterio bayesiano de Schwarz (BIC) para una muestra de 50 registros. Los valores de estos criterios aumentan como mínimo en 171% en las muestras censuradas respecto al bloque de control. Ahora, también se puede observar que estos valores de criterios de información disminuyen hasta en un 13% en el bloque de truncamiento con respecto al bloque de control.

CUADRO Nº 5.6

VALORES DE CRITERIOS DE INFORMACIÓN Y VARIACIÓN PORCENTUAL EN LOS BLOQUES DE CONTROL, CENSURA Y TRUNCAMIENTO EN EL ESCENARIO 2 PARA MUESTRA DE 50

%	BLOQUES	CRITERIOS	VALOR	VAR. %
0%	CONTROL	Logaritmo de la verosimilitud restringido -2	138.45	0%
		Criterio de información Akaike (AIC)	148.45	0%
		Criterio de Hurvich y Tsai (AICC)	149.95	0%
		Criterio de Bozdogan (CAIC)	162.59	0%
		Criterio bayesiano de Schwarz (BIC)	157.59	0%
10%	CENSURA	Logaritmo de la verosimilitud restringido -2	417.20	201%
		Criterio de información Akaike (AIC)	427.20	188%
		Criterio de Hurvich y Tsai (AICC)	428.70	186%
		Criterio de Bozdogan (CAIC)	441.34	171%
		Criterio bayesiano de Schwarz (BIC)	436.34	177%
	TRUNCAMIENTO	Logaritmo de la verosimilitud restringido -2	120.81	-13%
		Criterio de información Akaike (AIC)	130.81	-12%
		Criterio de Hurvich y Tsai (AICC)	132.57	-12%
		Criterio de Bozdogan (CAIC)	144.25	-11%
		Criterio bayesiano de Schwarz (BIC)	139.25	-12%

FUENTE: Huanca Milagros FECHA: Noviembre 2016

ELABORACIÓN: Propia

En el cuadro Nº 5.7 vemos los valores del Criterio de información Akaike (AIC), Criterio de Hurvich y Tsai (AICC), Criterio de Bozdogan (CAIC) y Criterio bayesiano de Schwarz (BIC) para una muestra de 200 registros. Los valores de estos criterios aumentan como mínimo en 177% en las muestras censuradas respecto al bloque de control. Ahora, también se puede observar que estos valores de criterios de información disminuyen en un 11% en el bloque de truncamiento con respecto al bloque de control.

CUADRO Nº 5.7

VALORES DE CRITERIOS DE INFORMACIÓN Y VARIACIÓN PORCENTUAL EN LOS BLOQUES DE CONTROL, CENSURA Y TRUNCAMIENTO EN EL ESCENARIO 2 PARA MUESTRA DE 200

%	BLOQUES	CRITERIOS	VALOR	VAR. %
0%	CONTROL	Logaritmo de la verosimilitud restringido -2	619.06	0%
		Criterio de información Akaike (AIC)	629.06	0%
		Criterio de Hurvich y Tsai (AICC)	629.38	0%
		Criterio de Bozdogan (CAIC)	650.45	0%
		Criterio bayesiano de Schwarz (BIC)	645.45	0%
10%	CENSURA	Logaritmo de la verosimilitud restringido -2	1716.72	177%
		Criterio de información Akaike (AIC)	1726.72	174%
		Criterio de Hurvich y Tsai (AICC)	1727.03	174%
		Criterio de Bozdogan (CAIC)	1748.11	169%
		Criterio bayesiano de Schwarz (BIC)	1743.11	170%
	TRUNCAMIENTO	Logaritmo de la verosimilitud restringido -2	550.84	-11%
		Criterio de información Akaike (AIC)	560.84	-11%
		Criterio de Hurvich y Tsai (AICC)	561.20	-11%
		Criterio de Bozdogan (CAIC)	581.64	-11%
		Criterio bayesiano de Schwarz (BIC)	576.64	-11%

FUENTE: Huanca Milagros FECHA: Noviembre 2016

ELABORACIÓN: Propia

En el cuadro Nº 5.8 vemos los valores del Criterio de información Akaike (AIC), Criterio de Hurvich y Tsai (AICC), Criterio de Bozdogan (CAIC) y Criterio bayesiano de Schwarz (BIC) para una muestra de 600 registros. Los valores de estos criterios aumentan como mínimo en 192% en las muestras censuradas respecto al bloque de control. Ahora, también se puede observar que estos valores de criterios de información disminuyen en un 10% en el bloque de truncamiento con respecto al bloque de control.

CUADRO Nº 5.8

VALORES DE CRITERIOS DE INFORMACIÓN Y VARIACIÓN PORCENTUAL EN LOS BLOQUES DE CONTROL, CENSURA Y TRUNCAMIENTO EN EL ESCENARIO 2 PARA MUESTRA DE 600

%	BLOQUES	CRITERIOS	VALOR	VAR. %
0%	CONTROL	Logaritmo de la verosimilitud restringido -2	1768.00	0%
		Criterio de información Akaike (AIC)	1778.00	0%
		Criterio de Hurvich y Tsai (AICC)	1778.10	0%
		Criterio de Bozdogan (CAIC)	1804.95	0%
		Criterio bayesiano de Schwarz (BIC)	1799.95	0%
10%	CENSURA	Logaritmo de la verosimilitud restringido -2	5168.28	192%
		Criterio de información Akaike (AIC)	5178.28	191%
		Criterio de Hurvich y Tsai (AICC)	5178.39	191%
		Criterio de Bozdogan (CAIC)	5205.24	188%
		Criterio bayesiano de Schwarz (BIC)	5200.24	189%
	TRUNCAMIENTO	Logaritmo de la verosimilitud restringido -2	1593.22	-10%
		Criterio de información Akaike (AIC)	1603.22	-10%
		Criterio de Hurvich y Tsai (AICC)	1603.33	-10%
		Criterio de Bozdogan (CAIC)	1629.62	-10%
		Criterio bayesiano de Schwarz (BIC)	1624.62	-10%

FUENTE: Huanca Milagros FECHA: Noviembre 2016

ELABORACIÓN: Propia

CONCLUSIONES

- ✓ El hecho de que la variable dependiente se encuentra censurada o truncada tiene efectos (en diferencias porcentuales) sobre los criterios de información Akaike (AIC), criterios de información de Akaike corregido (AICC) y criterios de información de Akaike consistente (CAIC) pues estos valores tienden a disminuir hasta 12% en el caso de truncamiento y aumentar hasta 190% en el caso de censura.

- ✓ Tanto para el escenario 1 en donde la muestra se fijó con 200 registros y se realizó diferentes proporciones de truncamiento como son $10, 20\ y\ 30\%$, como para el escenario 2 en donde se fijó la proporción de truncamiento en 10% para diferentes tamaños muestrales como son 50, 200 y 600 registros, se obtuvieron valores de Criterio de información Akaike (AIC), Criterio de información de Akaike corregido (AICC) y Criterio de información de Akaike consistente (CAIC) por debajo que los valores obtenidos en el bloque de control en donde no existe ningún tipo de modificación de datos. La diferencia en valores porcentuales llega a ser hasta de 12% (cuando la muestra es de 50).

- ✓ El efecto de la censura de la variable dependiente en los valores de Criterio de información Akaike (AIC), Criterio de información de Akaike corregido

(AICC) y Criterio de información de Akaike consistente (CAIC) tanto en el escenario 1 en donde la muestra se fijó con 200 registros y se realizó diferentes proporciones de censurade$10, 20\ y\ 30\%$, como en el escenario 2 en donde se fijó la proporción de truncamiento en 10% para diferentes tamaños muestrales como son 50, 200 y 600 registros, es muy negativo pues estos valores de criterios de información con respecto a los valores en el bloque de control en donde no se realizó censura alguna tienen una diferencia porcentual de hasta 190% (para una muestra de 600).

RECOMENDACIONES

En el caso de que se tenga la variable dependiente censurada, se recomienda no trabajar con esa variable sino realizar el truncamiento para que los valores de los criterios de información tales como el Akaike, Akaike consistente o Akaike corregido no lleguen a alterarse más del 10% con respecto a lo que debería ser.

Se recomienda seguir métodos de imputación de datos en las variables censuradas para que los registros con variable dependiente censurada puedan tener un valor en el target estimado.

COSTEO Y PRESUPUESTO

En el proceso de elaboración de tesis se ha considerado dentro del concepto de Gastos fijos al servicio de internet que tendrá una duración de 6 meses dando en total un costo de S/.600. Dentro del concepto de Gastos únicos, se detallan los costos de una laptop lenovo con procesador intel core i5, útil para el rápido desarrollo de los procesos estadísticos detallados en la tesis , un USB de 16 gigas perteneciente a la marca HP debido a su buena calidad, un cuaderno de apuntes para llevar un orden adecuado de la literatura revisada, 2 lapiceros y el encuadernado que se realizará al finalizar la tesis. Dentro de los gastos variables, se tiene en consideración que la literatura con gran relevancia en la ejecución de la tesis ha de ser impresa para un mejor manejo de apuntes, asimismo se tiene en cuenta los posibles papers que se han de comprar para enriquecer la tesis. Finalmente, el costo total de la tesis tiene la suma de S/.5,588.50 nuevos soles.

CUADRO DE COSTO Y PRESUPUESTO DE TESIS

		Costo (En soles)			
CONCEPTO	***Q***	***Unitario***		***Total***	
Gastos Fijos	*N° Meses*				
Internet	6	S/.	100.00	S/.	600.00
Gastos Únicos	*N° Unidades*				
Laptop core i5	1	S/.	4,000.00	S/.	4,000.00
USB 16Gg	1	S/.	35.00	S/.	35.00
Cuaderno de apuntes	1	S/.	1.50	S/.	1.50
Lapiceros	2	S/.	1.00	S/.	2.00
Encuadernización	1	S/.	50.00	S/.	50.00
Gastos Variables	*N° Unidades*				
Impresiones	500	S/.	0.10	S/.	50.00
Papers	5	S/.	170.00	S/.	850.00
TOTAL				***S/.***	***5,588.50***

FUENTE: Huanca Milagros FECHA: Octubre 2016
ELABORACIÓN: Propia

DIAGRAMA DE GANTT

El diagrama de Gantt que se muestra a continuación detalla las actividades a realizar para la elaboración de tesis que tiene un tiempo de elaboración de 4 meses finalizando el 15 de diciembre. Los avances de las actividades detalladas en el cuadro se realizarán de manera continua teniendo como fecha límite de revisión por el asesor de tesis cada quincena y fin de mes.

DIAGRAMA DE GANTT PARA TESIS

Fechas	Agosto	Septiembre		Octubre		Noviembre		Diciembre
Actividades	Del 16 al 31	Del 01 al 15	Del 16 al 30	Del 01 al 15	Del 16 al 31	Del 01 al 15	Del 16 al 30	Del 01 al 15
Antecedentes	x							
Problema de Investigación		x						
Marco teórico			x					
Metodología				x				
Resultados					x	x		
Conclusiones							x	
Recomendaciones							x	
Referencias bibliográficas								x
Anexos								x
Resúmen								x

FUENTE: Huanca Milagros FECHA: Octubre 2016

ELABORACIÓN: Propia

REFERENCIAS BIBLIOGRÁFICAS

Albarrán Peréz, P. (2011). Modelos para Datos Censurados y de selección muestral. Universidad de Alicante.

Álvarez, B. (2008). Modelos censurados, truncados y con selección muestral. *Econometría II.*

Chakraborty H. (2009). Mixed Model Approach for Intent-to-Treat Analysis in. *Triangle Park*.

De la Cruz, F. (2008). Modelos multinivel. *Sección de Epidemiología del Instituto de Medicina Tropical Daniel A. Carrión*.

García de Yébenes, M., Zunzunequi, M., Mathieu , M., Rodríguez Lazo, A., & Otero, A. (2002). Multilevel models applications to the analysis of longitudinal data.

Goldstein, H., & Healy, M. (1995). The Graphical Presentation of a Collection of Means.

Hrishikesh , C., & Hong , G. (2009). A Mixed Model Approach for Intent-to-Treat Analysis in Longitudinal Clinical Trials with Missing Values. *RTI International*.

Journal of Epidemiology, Community Health. (2001). *Journal of Epidemiology and Community Health*.

Martinez, C. (2014). Multilevel Analysis Software: A comparative study of MLwiN, HLM, SPSS and. *Reunido*.

Matos, L., Tsung, L., Castro, L., & Lachos, V. (2016). Heavy-tailed longitudinal regression models for censored data: A likelihood based perspective.

Murillo, F. J. (2008). Multilevel Models as a Tool for Research in Education.

Pérez Fernández, V. (2013). Los modelos multinivel en el análisis de factores de riesgo de sibilancias recurrentes en lactantes. *Universidad de Murcia*.

Zuehlke, T., & Kassekert, A. (2008). Algorithmic errors in the estimation of tobit II Models and the corresponding failure to recognize selection bias. *Department of economics, Florida State university*.

ANEXOS

Simulación de base de datos en R

Escenario 1

```
set.seed(1)
n=200
n_NA=n/2
n_NB=n/2
IndxMH<-seq(1,n,1)
library(MASS)
#Para NA
NACs10X1<-data.frame(rnorm(n_NA, 3, 2),names="A")
colnames(NACs10X1)<-c("X1","N1")
NACs10X2<-data.frame(rnorm(n_NA, 0, 1.2),names="A")
colnames(NACs10X2)<-c("X2","N1")
#NACs10E<-rnorm(n_NA, 0, 1.2)

#Para NB
NBCs10X1<-data.frame(rnorm(n_NB, 6, 2), names="B")
colnames(NBCs10X1)<-c("X1","N1")
NBCs10X2<-data.frame(rnorm(n_NB, 2, 3.5), names="B")
colnames(NBCs10X2)<-c("X2","N1")
#NBCs10E<-rnorm(n_NB, 0, 1.2)

Es1Cs10E<-data.frame(rnorm(n, 0, 1))

Cs10X1<-rbind(NACs10X1,NBCs10X1)
colnames(Cs10X1)<-c("X1","N1")
Cs10X2<-rbind(NACs10X2,NBCs10X2)
colnames(Cs10X2)<-c("X2","N1")

yEs1Cs10<-10+4*Cs10X1[,1]+2.5*Cs10X2[,1]+Es1Cs10E

DtsEs1Cs10<-cbind(Cs10X1[,2],IndxMH,yEs1Cs10,  Cs10X1[,1],Cs10X2[,1],
Es1Cs10E)
```

```
colnames(DtsEs1Cs10)<-c("N2","N1","Y","X1","X2","E")

#Con 10% de Censura
ncesA<-round(n_NA*0.3)
ncesB<-round(n_NB*0.3)

CesA<-yEs1Cs10[DtsEs1Cs10$N2=="A",1]
ubicA<-matrix(data=0,1,n_NA)
ncesAtemp<-0
while(ncesAtemp<=ncesA)
{elemento<-round(runif(1,1,n_NA))
if(ubicA[elemento]==0)
  { ncesAtemp<-ncesAtemp+1
ubicA[elemento]=1 }
}
ubicA<-data.frame(t(ubicA))
colnames(ubicA)<-"Censura10"


CesB<-yEs1Cs10[DtsEs1Cs10$N2=="B",1]
ubicB<-matrix(data=0,1,n_NB)
ncesBtemp<-0
while(ncesBtemp<=ncesB)
{elemento<-round(runif(1,1,n_NB))
if(ubicB[elemento]==0)
{ ncesBtemp<-ncesBtemp+1
ubicB[elemento]=1 }
}
ubicB<-data.frame(t(ubicB))
names(ubicB)<-"Censura10"
```

```
Censura10<-rbind(ubicA,ubicB)
names(Censura10)<-"Censura10"

Censura20<-rbind(ubicA,ubicB)
names(Censura20)<-"Censura20"

Censura30<-rbind(ubicA,ubicB)
names(Censura30)<-"Censura30"

#Base Total para escenario 1
DtsEs1<-cbind(DtsEs1Cs10,Censura10,Censura20,Censura30)
```

Escenario 2

```
set.seed(1)
n=600
n_NA=n/2
n_NB=n/2
IndxMH<-seq(1,n,1)
library(MASS)
#Para NA
NACs10X1<-data.frame(rnorm(n_NA, 3, 1),names="A")
colnames(NACs10X1)<-c("X1","N1")
NACs10X2<-data.frame(rnorm(n_NA, 15, 3),names="A")
colnames(NACs10X2)<-c("X2","N1")
#NACs10E<-rnorm(n_NA, 0, 1.2)

#Para NB
NBCs10X1<-data.frame(rnorm(n_NB, 5, 1), names="B")
colnames(NBCs10X1)<-c("X1","N1")
NBCs10X2<-data.frame(rnorm(n_NB, 10, 3), names="B")
```

```
colnames(NBCs10X2)<-c("X2","N1")
#NBCs10E<-rnorm(n_NB, 0, 1.2)

Es1Cs10E<-data.frame(rnorm(n, 0, 1))

Cs10X1<-rbind(NACs10X1,NBCs10X1)
colnames(Cs10X1)<-c("X1","N1")
Cs10X2<-rbind(NACs10X2,NBCs10X2)
colnames(Cs10X2)<-c("X2","N1")

yEs1Cs10<-10+4*Cs10X1[,1]+2.5*Cs10X2[,1]+Es1Cs10E

DtsEs1Cs10<-cbind(Cs10X1[,2],IndxMH,yEs1Cs10,  Cs10X1[,1],Cs10X2[,1],
Es1Cs10E)
colnames(DtsEs1Cs10)<-c("N2","N1","Y","X1","X2","E")

#Con pp (Sin porcentajes) de Censura
pp=0.1
ncesA<-round(n_NA*pp)
ncesB<-round(n_NB*pp)

CesA<-yEs1Cs10[DtsEs1Cs10$N2=="A",1]
ubicA<-matrix(data=0,1,n_NA)
ncesAtemp<-0
while(ncesAtemp<=ncesA)
{elemento<-round(runif(1,1,n_NA))
if(ubicA[elemento]==0)
  { ncesAtemp<-ncesAtemp+1
ubicA[elemento]=1 }
}
ubicA<-data.frame(t(ubicA))
```

```
colnames(ubicA)<-"Censura10"


CesB<-yEs1Cs10[DtsEs1Cs10$N2=="B",1]
ubicB<-matrix(data=0,1,n_NB)
ncesBtemp<-0
while(ncesBtemp<=ncesB)
{elemento<-round(runif(1,1,n_NB))
if(ubicB[elemento]==0)
{ ncesBtemp<-ncesBtemp+1
ubicB[elemento]=1 }
}
ubicB<-data.frame(t(ubicB))
names(ubicB)<-"Censura10"

Censura10<-rbind(ubicA,ubicB)
names(Censura10)<-"Censura10"


#Base Total para escenario 2
n50Es2<-cbind(DtsEs1Cs10,Censura10)
n200Es2<-cbind(DtsEs1Cs10,Censura10)
n600Es2<-cbind(DtsEs1Cs10,Censura10)

write.table(n50Es2, file="n50Es2.txt", append=FALSE, sep='\t')
write.table(n200Es2, file="n200Es2.txt", append=FALSE, sep='\t')
write.table(n600Es2, file="n600Es2.txt", append=FALSE, sep='\t')


cor(DtsEs1[DtsEs1$N2=="A",]$X1,DtsEs1[DtsEs1$N2=="A",]$X2)
```

```
cor(DtsEs1[DtsEs1$N2=="A",]$X1,DtsEs1[DtsEs1$N2=="B",]$X1)

cor(DtsEs1[DtsEs1$N2=="A",]$X2,DtsEs1[DtsEs1$N2=="B",]$X2)
```