FACULTAD DE CIENCIAS ECONÓMICAS Y DE ADMINISTRACIÓN

LICENCIATURA EN ESTADÍSTICA

La capacidad pulmonar de escolares de la ciudad de Artigas y la relación con su entorno: Una aplicación de Modelos Mixtos

Yohana Altez de Castro

Virginia Burguete Eguren

Tutores:

Ramón Álvarez Vaz

Elena Vernazza Mañán

Montevideo, Diciembre de 2018.

UNIVERSIDAD DE LA REPÚBLICA
FACULTAD DE CIENCIAS ECONÓMICAS Y DE ADMINISTRACIÓN

El tribunal docente integrado por los abajo firmantes aprueba el trabajo:

La capacidad pulmonar de escolares de la ciudad de Artigas y la relación con su entorno: Una aplicación de Modelos Mixtos

Yohana Altez de Castro - Virginia Burguete Eguren

Tutores:

Ramón Alvarez Vaz

Elena Vernazza Mañán

Licenciatura en Estadística

Puntaje .....

Tribunal

Profesor.....(nombre y firma).

Profesor.....(nombre y firma).

Profesor.....(nombre y firma).

Fecha.....

Índice general

Resumen

En la presente investigación se implementa una aplicación de la técnica de modelos mixtos para el análisis de datos longitudinales. El objetivo de este estudio es conocer como inciden determinadas variables en la determinación de la capacidad pulmonar, tomando como medida el Pico Flujo Espiratorio (PFE).

La motivación de este estudio, surge a partir del planteo realizado por la sociedad civil residente en la zona próxima a los molinos arroceros, al área Salud Ambiental y Ocupacional del Ministerio de Salud Pública, los cuales alegan aumento en patologías respiratorias a raíz de los procesos industriales de producción de arroz.

El análisis se centra en mediciones de la capacidad pulmonar de niñas y varones de entre 4 y 7 años de edad, matriculados en las escuelas urbanas de la ciudad de Artigas, en 4 tomas a lo largo de un año de monitoreo durante el período 2015–2016.

Los datos utilizados en este estudio, resultan del trabajo de campo realizado previo a la investigación, en donde se seleccionaron muestras por conglomerados, mediante un diseño sistemático ordenado por sexo y curso. Para cada niña y varón en la muestra sorteada, se tomaron mediciones de peso, talla y capacidad pulmonar, considerando además información obtenida mediante cuestionarios que le fueran entregados a los padres, donde se efectuaron consultas referidas a los antecedentes clínicos del niño o niña y las condiciones ambientales a las que están expuestos.

Se hace foco en las variables asociadas a los procesos de industrialización del arroz (etapa de monitoreo y zona de residencia) y en variables asociadas a los antecedentes del individuo (patologías respiratorias y exposición a humo de tabaco o leña).

Se analizan las mediciones de la capacidad pulmonar de la población de estudio y se contrastan con las referidas a niñas y varones de todo el territorio nacional que no poseen patologías.

Palabras claves: Afecciones respiratorias, Capacidad pulmonar, Modelos Mixtos, Molinos arroceros, Monitoreo.

Índice de figuras

Índice de tablas

Capítulo 1

Introducción

Los procesos industriales son un conjunto de operaciones necesarias para modificar las características de las materias primas. La actividad industrial genera altas cantidades de sustancias extrañas que alcanzan niveles peligrosos de contaminantes para la vida en general, ya que superan la capacidad que el ecosistema tiene para deshacerse de ellos.

Se entiende por contaminación industrial, a la emisión directa o indirecta de sustancias nocivas, tóxicas o peligrosas, desde las instalaciones o procesos industriales hacia el entorno.

Según las Guías de calidad de aire de la OMS (15), *"Las pruebas relativas al material particulado (MP) suspendido en el aire y sus efectos en la salud pública coinciden en poner de manifiesto efectos adversos para la salud con las exposiciones que experimentan actualmente las poblaciones urbanas, tanto en los países desarrollados como en desarrollo. El abanico de los efectos en la salud es amplio, pero se producen en particular en los sistemas respiratorio y cardiovascular. Se ve afectada toda la población, pero la susceptibilidad a la contaminación puede variar con la salud o la edad. Se ha demostrado que el riesgo de diversos efectos aumenta con la exposición, y hay pocas pruebas que indiquen un umbral por debajo del cual no quepa prever efectos adversos en la salud. En realidad, el nivel más bajo de la gama de concentraciones*

para las cuales se han demostrado efectos adversos no es muy superior a la concentración de fondo, que para las partículas de menos de 2,5 (MP 2,5) se ha estimado en 3-5 g/m3 tanto en los Estados Unidos como en Europa occidental. Las pruebas epidemiológicas ponen de manifiesto efectos adversos del MP tras exposiciones tanto breves como prolongadas".

En particular, los procesos industriales asociados a la elaboración del arroz producen gran cantidad de residuos sólidos, la principal fuente contaminante es el polvillo que se genera en las distintas etapas del proceso[1]. En primer lugar, en la **etapa de ingreso**, con la carga y descarga de la materia prima de los camiones, en segundo lugar en la **etapa de pre-limpieza**, donde se separan las impurezas de la materia prima, tales como pajas, cáscaras, tallos, etc., luego, en la **etapa de secado** donde el polvillo es generado desde los silos de secado del grano y sistemas de aspiración. Una vez ingresado al molino, restan las etapas de **descascarado, pulido, clasificación y empaque**, en las cuales también se producen residuos sólidos en forma de polvillo. Este material particulado no permanece únicamente en la planta, sino que se dispersa en el medio ambiente, afectando el entorno.

En Uruguay, las principales zonas de cultivo de arroz, son la **zona norte**, conformada por los departamentos de *Artigas y Salto*, la **zona centro** que integra los departamentos de *Rivera, Tacuarembó, Durazno y Río Negro* y la **zona este** compuesta por los departamentos de *Rocha, Treinta y Tres, Lavalleja* y *Cerro Largo*. Los factores que inciden en la elección de las zonas de cultivo son la disponibilidad de agua (principalemente por la Cuenca del Río Cuareim y la Cuenca de la Laguna Merín) y por factores culturales (3).

Además, cabe resaltar, que el arroz es uno de los principales productos de exportación del país. Para el año 2015, ocupaba el sexto lugar luego de la carne, celulosa, soja, productos lácteos y concentrado de bebidas, con cifras superiores a los U$S

[1]Ver proceso del arroz en Figura A.1.

360 millones, según datos del Informe Anual de comercio exterior realizado por *Uruguay XXI* (21). A nivel mundial, de acuerdo a *UN Comtrade* (United Nations International Trade Statistics Database), para el año 2015 Uruguay ocupaba el noveno lugar en el listado de países exportadores de arroz, con el 1,69 %, luego de India (26,36 %), Tailandia (22,50 %), Vietnam (13,10 %), Pakistán (10,08 %), Estados Unidos (7,03 %), Italia (2,74 %), Australia (1,77 %) y Camboya (1,71 %), y ocupa el primer lugar de América del Sur con el 36,14 % de las exportaciones de arroz de la región.

Teniendo estos aspectos en consideración, en este estudio se analiza el impacto de los procesos industriales derivados de los molinos arroceros, en las afecciones respiratorias de los habitantes de la ciudad de Artigas, en la zona norte del país.

La motivación de este estudio surge a partir del planteo realizado por la sociedad civil residente en la zona próxima a los molinos arroceros al área Salud Ambiental y Ocupacional del Ministerio de Salud Pública, los cuales perciben aumento en patologías respiratorias a raíz de los procesos industriales asociados a la producción de arroz.

Dado que las afecciones respiratorias en la población adulta pueden deberse a otras causas, tales como el tabaquismo, en el presente trabajo se analiza únicamente a niños en edad escolar. Por lo tanto, la población objetivo está formada por niñas y varones entre 4 y 7 años de edad inclusive[2]. La muestra seleccionada es de 714 escolares, se monitorean en 4 etapas a lo largo de un año (2 monitoreos dentro de zafra y 2 fuera de zafra)[3] midiendo su peso, talla y capacidad pulmonar.

Se desea estudiar la evolución del PFE en 4 etapas de monitoreo, ya que esta medida es una buena variable para determinar afecciones de tipo pulmonar, ya que las afecciones respiratorias tienden a indicar una capacidad pulmonar menor a la esperada como síntoma clínico.

[2]Responde al grupo etario con mayor sensibilidad a las variaciones de la calidad de aire.

[3]Se considera la zafra en los meses de febrero a abril.

Las hipótesis de este trabajo son:

- La disminución en los valores de la capacidad pulmonar se debe a incrementos en la producción de material particulado, asociados a períodos de zafra arrocera.
- La capacidad pulmonar de niños que residen (o concurren a escuelas) dentro de la zona industrial, es inferior a la de niños que residen (o concurren a escuelas) de otras zonas.
- La capacidad pulmonar de los escolares residentes en la ciudad de Artigas es menor que la de niñas y varones sin patologías de enfermedades respiratorias a nivel nacional.

1.1. Objetivos

El objetivo general de este trabajo es analizar el impacto de las variables de la industrialización del arroz (zafra, entorno geográfico, etc.) en la capacidad pulmonar de los escolares de la ciudad de Artigas.
Los objetivos específicos son:

1. Comparar el valor del PFE observado de niñas y varones de la ciudad de Artigas con respecto a valores percentilares de las curvas de referencia(2).
2. Analizar la exposición a contaminación intradomiciliaria y patologías a cuadros respiratorios de los individuos monitoreados.
3. Estudiar la asociación que existe entre la capacidad pulmonar medida a través del PFE con las variables: *zona de residencia* y *etapa de monitoreo*.

La presente investigación se estructura en cinco capítulos. En el presente Capítulo (Capítulo 1) se introduce al tema de interés, también se describen tanto la motivación del estudio, como las hipótesis de partida y los objetivos que se plantean. En el

Capítulo 2 se desarrollan los conceptos teóricos a utilizar, se prosigue en el Capítulo 3 donde se detallan los principales aspectos del relevamiento de datos y los métodos y herramientas utilizados. En el Capítulo 4 se exponen los resultados obtenidos, y por último, en el Capítulo 5 se plantean las conclusiones que derivan de esta investigación.

1.2. Antecedentes

El principal antecedente para esta investigación es el estudio bajo el nombre **Pico de flujo espiratorio en niños uruguayos sin enfermedad, de 3 a 13 años**(2) El objetivo de este estudio es determinar el PFE en niños uruguayos sin patología respiratoria, con edades de 3 a 13 años y describir el PFE de estos niños, en función del peso, talla, edad y sexo.
El PFE ha sido incluido en las recomendaciones de consenso sobre control y tratamiento del asma desde la década de los 90. Dado que presenta variaciones según diferentes poblaciones, se determinaron los valores de PFE de niños uruguayos sin enfermedad. Se estudiaron 362 varones y 437 niñas, con edades de 3 a 13 años cumplidos, sin antecedentes respiratorios ni utilización de medicación antiasmática en su historia previa, y que al momento del estudio no evidenciaran anormalidades del examen clínico del aparato respiratorio. Se determinó el valor percentilar 10, 50 y 90 para cada intervalo de clase y se correlacionó mediante una regresión de mínimos cuadrados con un polinomio de 2º grado. El coeficiente de correlación (R^2) para los valores de p.10, p.50 y p.90 fue de 0,99, 0,98 y de 1,00 para talla, siendo levemente inferiores para edad y peso. Los valores de PFE son mayores en los varones con respecto a las niñas.

Otro antecedente a considerar es la **Evaluación de la función respiratoria en niños de 6 a 12 años de la cuidad de Tacuarembó: potencial efecto de**

la zafra arrocera(13). El objetivo de esta investigación es comparar el estado de la función respiratoria de los niños habitantes de la ciudad de Tacuarembó durante y fuera del período de zafra de la planta y evaluar la eventual influencia de la "proximidad a la planta". Debido a planteos de percepción de los habitantes de la ciudad, por presencia de contaminantes ambientales atmosféricos identificado como material particulado proveniente de fuentes industriales, generado sobre todo por la actividad arrocera y la presencia de silos en la ciudad, se desea analizar la capacidad pulmonar dentro de zafra y en ausencia de zafra, y según la zona de proximidad de la escuela a la planta industrial.

Se toma una muestra representativa de niños. Se considera por lo tanto la ubicación de la escuela como referencia geográfica del niño, se evalúa la capacidad pulmonar en dos etapas: durante y fuera de zafra. Para evaluar la eventual influencia de la proximidad a la planta sobre el estado respiratorio de los niños, se agrupan las escuelas en función de su distancia a la planta, conformándose tres grupos (zona 1, zona 2, zona 3) y se realiza un análisis descriptivo de posible correlación de las influencia entre dichas variables

Se concluye que si bien la proximidad a la planta muestra una leve tendencia a la *"disminución"* de los PFE (aumento de las diferencias) durante la zafra, ésta no es estadísticamente significativa. Por lo tanto, no se puede concluir que la proximidad a la planta aumente la probabilidad de afectación respiratoria, presentándose modificaciones del PFE tanto en los niños que concurren a escuelas cercanas a la planta como los que no.

En el año 2015, surge la investigación **Efecto de valores faltantes en estudios longitudinales en adultos mayores**(12), en la cual se desea aplicar técnicas referentes al modelado conjunto de datos longitudinales y de sobrevida. Para ello se utilizan datos del estudio "Origins of Variance in the Old-old: Octogenarian Twins" (OCTO-Twin Study). El mismo cuenta con 351 parejas de mellizos de 80 años o más para el año 1991 en Suecia. El período de seguimiento de los individuos cons-

ta de 4 visitas (posteriores a la evaluación inicial) en períodos de 2 años donde se recabaron datos sobre memoria, capacidad funcional, y salud entre otras mediciones. Se focaliza en la evolución del resultado del "Mini-Mental State Examination" ($MMSE$). Se trata de un cuestionario con puntaje máximo de 30 puntos desarrollado por Folstein que pretende determinar el deterioro cognitivo y detectar demencia. El objetivo es estimar la velocidad de cambio del $MMSE$ y determinar los factores que puedan alterarla, prestando especial atención al proceso de fallecimiento de los individuos ya que el mismo puede sesgar los resultados. Se llega a la conclusión de que la evolución del $MMSE$ se ve afectada tanto por la edad como por la educación de las personas. Al incluir el análisis de sobrevida como parte del proceso de deterioro cognitivo, se "corrigen" los valores de los coeficientes estimados en el modelo especificado bajo Datos Perdidos al Azar (Missing at Random (MAR)). Se observa que la sobrevida de los hombres es inferior a la de las mujeres y que a mayor edad al inicio, mayor es el riesgo de fallecimiento. Adicionalmente se observa que valores bajos de $MMSE$ incrementan el riesgo de fallecimiento. Lo mismo sucede con la pendiente en la evolución del $MMSE$, al disminuir la misma (avance del deterioro) el riesgo de fallecimiento aumenta aún más. En cuanto al modelo longitudinal se observa que la trayectoria del $MMSE$, entre los individuos con educación baja, decrece más lentamente que lo indicado en principio por el modelo MAR mientras que la educación tiene un efecto mayor al pensado originalmente.

En 2017, se realizó la investigación denominada **Trayectoria Nutricional y desempeño escolar**(11), en la que se plantea, entre otros objetivos, analizar la evolución temporal del estado nutricional de los niños y niñas mediante la estimación de Modelos Mixtos de Clases Latentes. Considerando la infancia, como el período de vida en el que se puede presentar efectos adversos de corto y largo plazo en varias dimensiones, tales como, la educación y la salud. Este trabajo analiza la trayectoria nutricional de una cohorte de niños y niñas de Uruguay en edad escolar y su vínculo con el desempeño educativo mediante la estimación de Modelos Mixtos Conjuntos

de Clases Latentes. Se estudia el cambio con la edad del estado nutricional, medido a través del Índice de Masa Corporal (IMC). Los modelos ajustados identifican tres grupos de niños y niñas en el período escolar (entre 6 y 12 años aproximadamente), según sus trayectorias nutricionales. En cuanto a la relación entre trayectoria nutricional y desempeño escolar, la evidencia muestra que los niños abandonan a edades más tempranas que las niñas. Sin embargo, los grupos según trayectorias nutricionales no presentan diferencias en la edad de abandono escolar. Los resultados de este trabajo indican que si bien ninguno de los grupos identificados muestra características de déficit nutricional o decrecimiento del IMC en el período escolar, sí se observa un conjunto de niñas y niños con problemas de sobrepeso u obesidad que podría ser objeto de estudio en futuras investigaciones.

Otro antecedente es la investigación bajo el nombre **Modelos mixtos ¿mlne o ml4?**(4), cuyo objetivo es evaluar el efecto del extracto de Justicia secunda Vahl (Acanthaceae) sobre la glucemia de ratas adultas de experimentación con y sin carga de sacarosa. La glucemia es la medida de concentración de glucosa libre en la sangre, suero o plasma sanguíneo. La insulina es una hormona (producida por el páncreas) que toma glucosa de la sangre y la transporta al interior de las células del cuerpo donde se usa como energía. Se mide la glucemia de 4 grupos de 8 ratas cada uno, luego de 12 hs. de ayuno a cada rata. Se administran 3 tipos de tratamientos y posteriormente se hacen los controles de glucemia en todas las ratas cada 30 minutos y se analiza la incidencia en el nivel de glucemia. Este problema se encuadra dentro de los modelos lineales generalizados mixtos (GLMM), en particular en análisis de medidas repetidas. Se concluye que la inclusión de efectos aleatorios mejora la precisión del modelo frente a un modelo solo con efectos fijos, aunque se puede decidir la significación estadística de efectos fijos (pero no de los efectos aleatorios). Además se concluye, que el tipo de tratamiento no es un buen predictor al no observarse diferencias entre los 4 tipos de tratamiento.

Capítulo 2

Marco Teórico

A continuación, se describen los modelos lineales con efectos fijos que consideran una serie de supuestos sobre los residuos, estos supuestos, se irán levantando hasta llegar a los *Modelos Lineales Mixtos* (5). En el apéndice B, se detalla el diseño de muestra utilizado (20).

2.1. Modelos Lineales

Se dice que un modelo es lineal, si lo es para sus parámetros. En general, se supone que una cierta variable aleatoria es igual a un valor fijo η más una desviación aleatoria ε.

$$Y = \eta + \varepsilon \tag{2.1}$$

donde η representa la verdadera medida de la variable, es decir, la parte *determinista* de un experimento, que depende de ciertos factores. El término ε representa el *error*, es la parte del modelo no controlable debido a múltiples causas aleatorias. En particular, los modelos de la forma:

$$y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + ... + \beta_k x_{ik} + \varepsilon_i \quad i = 1, ..., n \tag{2.2}$$

con $k > 1$ variables independientes predictoras o regresoras, se denominan **modelos de regresión múltiple**. La variable cuyos datos observados es y_i denominada variable dependiente o de respuesta. Los parámetros son desconocidos y el objetivo principal es estimarlos. La evaluación de los errores ε_i, permite analizar la performance del modelo y el cumplimiento de los supuestos.

Supuestos básicos del modelo de regresión lineal

El modelo lineal definido en (2.2) supone que los errores ε_i son desviaciones que se comportan como variables aleatorias que verifican las siguientes condiciones (de Gauss-Markov):

1. $E(\varepsilon_i) = 0 \quad i = 1, ...n$
2. $Var(\varepsilon_i) = \sigma^2 \quad i = 1, ..., n$
3. $E(\varepsilon_i \varepsilon_j) = 0 \quad \forall i \neq j$

La primera condición, es la que asegura que $E(y_i) = \beta_0 + \sum_j \beta_j x_{ij}$. La segunda condición, es la llamada condición de *homocedasticidad*, en donde σ^2, varianza del modelo, es un parámetro desconocido. La tercera condición, implica que las n desviaciones son mutuamente independientes.
A continuación, se profundiza en las condiciones del modelo.

2.1.1. Homocedasticidad

Un modelo lineal para observaciones independientes y normalmente distribuidas $\mathrm{y_i (i = 1, ..., n)}$ con varianza constante puede especificarse de varias maneras. La especificación más utilizada es la representación algebraica:

$$\mathrm{y}_i = x_i^{(1)} \beta_1 + ... + x_i^{(p)} \beta_p + \varepsilon_i, \tag{2.3}$$

donde $x_i^1, ..., x_i^p (p < n)$ son valores de las covariables conocidos, $\beta_1, ..., \beta_p$ son los parámetros (desconocidos) y $\varepsilon_1, ..., \varepsilon_p$ sus errores residuales independientes, tales

que:

$$\varepsilon_i \sim N\left(0, \sigma^2\right) \tag{2.4}$$

De manera análoga se puede especificar el modelo con su representación matricial:

$$\mathrm{y}_i = \mathbf{x_i}^{'}\boldsymbol{\beta} + \varepsilon_i, \tag{2.5}$$

donde los vectores columnas $\mathbf{x}_i \equiv (x_i^{(1)}, ..., x_i^{(p)})'$ son los valores de las covariables para la i-ésima observación y $\boldsymbol{\beta} \equiv (\beta_1, ..., \beta_p)'$ son sus respectivos parámetros (efectos fijos).

De las ecuaciones (2.3) a (2.5), resulta:

$$\mathrm{E}(y_i) \equiv \mu_i = \mathbf{x_i}^{'}\boldsymbol{\beta}, \tag{2.6}$$

$$Var(y_i) = \sigma^2, \tag{2.7}$$

Y además se definen:

$$\mathbf{y} \equiv \begin{pmatrix} y_1 \\ \vdots \\ y_n \end{pmatrix}, \boldsymbol{\varepsilon} \equiv \begin{pmatrix} \varepsilon_1 \\ \vdots \\ \varepsilon_n \end{pmatrix}, \tag{2.8}$$

y

$$\mathbf{X} \equiv \begin{pmatrix} \mathbf{x}_1^{'} \\ \vdots \\ \mathbf{x}_n^{'} \end{pmatrix} = \begin{pmatrix} x_1^{(1)} & x_1^{(2)} & \cdots & x_1^{(p)} \\ \vdots & \vdots & \ddots & \vdots \\ x_n^{(1)} & x_n^{(2)} & \cdots & x_n^{(p)} \end{pmatrix} \equiv \begin{pmatrix} \mathbf{x}^{(1)} & \mathbf{x}^{(2)} & \cdots & \mathbf{x}^{(p)} \end{pmatrix}. \tag{2.9}$$

El modelo especificado en (2.3) y (2.4) puede ser expresado como

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}, \tag{2.10}$$

donde $\mathbf{X}$ es la matriz de diseño $n \times p$ y con

$$\boldsymbol{\varepsilon} \sim N_n\left(\mathbf{0}, \mathcal{R}\right) \tag{2.11}$$

donde $\mathcal{R} = \boldsymbol{\sigma}^2\mathbf{I}_n$ es la matriz de varianzas-covarianzas, con $\mathbf{I}$ matriz identidad $n \times n$. Para la matriz de diseño $\mathbf{X}$ definida en (2.9), se supone para simplificar, que es de rango completo ($p < n$) o de forma equivalente, se asume que las columnas

$\mathbf{x}^{(1)}, ..., \mathbf{x}^{(p)}$ son linealmente independientes.

Estimación por Máxima Verosimilitud Restringida:

Existen varias formas de estimar los parámetros, como ser la Estimación por Mínimos Cuadrados Ordinarios (MCO) o la Estimación por Máxima Verosimilitud (EMV), en este caso, es de interés hacer foco en la Estimación por Máxima Verosimilitud Restringida (EMVR), ya que obtiene estimadores insesgados. Cabe señalar además que los estimadores MCO para $\boldsymbol{\beta}$ y $\boldsymbol{\sigma}^2$, son equivalentes a los estimadores MVR para modelos lineales con residuos independientes y homocedásticos. Esta equivalencia no se cumple para modelos más complejos, los cuales serán considerados de aquí en adelante.

Para obtener un estimador insesgado para σ^2 se utiliza una proyección ortogonal a la estimación de $\boldsymbol{\beta}$. Esto puede hacerse si se considera la función de verosimilitud basado en un conjunto de contraste independiente $(n-p)$ de y. El logaritmo de la función de verosimilitud-restringida está dada por:

$$\omega_{MVR}(\sigma^2; y) \equiv -\frac{n-p}{2} log(\sigma^2) - \frac{1}{2\sigma^2} \sum_{i=1}^{n} r_i^2, \tag{2.12}$$

donde

$$r_i \equiv y_i \mathbf{x}_i' \left(\sum_{i=1}^{n} \mathbf{x}_i \mathbf{x}_i'\right)^{-1} \sum_{i=1}^{n} \mathbf{x}_i y_i$$

Maximizar la ecuación (2.12) con respecto a σ^2 implica encontrar el estimador MVR:

$$\widehat{\sigma}^2_{MVR} \equiv \frac{1}{n-p} \sum_{i=1}^{n} r_i^2, \tag{2.13}$$

$\widehat{\sigma}^2_{MVR}$ es un estimador insesgado para σ^2.

Diagnóstico del Modelo

Previo a hacer inferencia en el modelo, es importante verificar que se cumplen los supuestos antes mencionados sobre los errores. Esto es, verificar que los errores residuales, $\boldsymbol{\varepsilon}_i$, son independientes, homocedásticos, e incorrelacionados.

Una herramienta apropiada para el diagnóstico de los residuos, son sus respectivos gráficos, en los que se detecta la presencia o ausencia de patrones específicos y/o datos atípicos. Dichos gráficos pueden basarse en varios tipos de residuos:

Residuos crudos: Los residuos más básicos, se definen para la observación i-ésima como $\widehat{\varepsilon}_i \equiv y_i - \widehat{\mu}_i$ donde $\widehat{\mu}_i \equiv \mathbf{x}_i^{'}\widehat{\beta}$ se conoce como el valor ajustado.

Residuos a escala: Son los residuos crudos escalados, es decir, se dividen por sus desviaciones estándar reales o estimadas, de modo que su interpretación no depende de las unidades de medida de la variable dependiente. Si se conociera la desviación estándar real[1], entonces se denominan residuos a escala estandarizados. De lo contrario, si se utiliza la desviación estándar estimada $\widehat{\sigma}$, los residuos obtenidos se denominan residuos estudentizados. Esta categoría se puede subdividir en residuos estudentizados internos y residuos estudentizados externos. Cabe aclarar que $\widehat{\sigma}$ denota la estimación de σ basada en todas las observaciones, mientras que $\widehat{\sigma}_{(-1)}$ es la estimación obtenida luego de excluir la i-ésima observación de los cálculos.

Tipo de residuo	Fórmula matemática
Estandarizados por σ	$\widehat{\varepsilon}_i/\sigma$
Estudentizados internos[2]	$\widehat{\varepsilon}_i/\widehat{\sigma}$
Estudentizados externos[3]	$\widehat{\varepsilon}_i/\widehat{\sigma}_{(-1)}$

Tabla 2.1: Forma básica de residuos escalados para modelos lineales

Se observa que al reemplazar $\widehat{\sigma}$ por $\widehat{\sigma}_{(-i)}$ la técnica de residuos estudentizados externos permite que los valores atípicos se destaquen de manera más prominente en comparación con la técnica de residuos estudentizados internos.

Sea $\mathbf{H}$, de $n \times n$, matriz definida por:

$$\mathbf{H} \equiv \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}' \tag{2.14}$$

La matriz $\mathbf{H}$ representa la proyección del vector $\mathbf{y}$ en el subespacio generado por las columnas de la matriz de diseño $\mathbf{X}$. El vector de los valores predichos de $\mathbf{y}$, $\widehat{\boldsymbol{\mu}} = \mathbf{X}\widehat{\boldsymbol{\beta}}$, se puede expresar como $\widehat{\boldsymbol{\mu}} = \mathbf{H}y$. La fórmula de la matriz de varianzas y covarianzas de $\widehat{\boldsymbol{\varepsilon}}$ es:

$$Var(\widehat{\boldsymbol{\varepsilon}}) = \sigma^2(\mathbf{I}_n - \mathbf{H}). \tag{2.15}$$

[1]En la práctica, σ rara vez es conocida

donde $\mathbf{I}_n$ es matriz identidad $n \times n$. En caso de que la matriz $\mathbf{H}$ en (2.15) no sea proporcional a $\mathbf{I}_n$, los residuos crudos son potencialmente heterocedásticos y/o correlacionados, en cambio los residuos escalados, presentados en la Tabla 2.1, no abordan el tema de heterocedasticidad y/o correlación.
Si los residuos de la Tabla 2.1 presentan heterocedasticidad, se pueden escalar al utilizar la matriz $\mathbf{H}$.

Tipo de residuo	Ajustados por h_{ii}
Estandarizados por σ	$(\widehat{\varepsilon}_i/\sigma)/\sqrt{1-h_{ii}}$
Estudentizados internos	$(\widehat{\varepsilon}_i/\widehat{\sigma})/\sqrt{1-h_{ii}}$
Estudentizados externos[4]	$(\widehat{\varepsilon}_i/\widehat{\sigma}_{(-1)})/\sqrt{1-h_{ii}}$

Tabla 2.2: Residuos escalados que implican $h_{i,i}$ en los elementos diagonales de la matriz

La Tabla 2.2 presenta los residuos de la Tabla 2.1, escalados mediante estimaciones de error estándar que implican elementos diagonales $h_{i,i}$ de la matriz $\mathbf{H}$. La escala, en este caso, aborda el problema de heterocedasticidad de los residuos crudos, pero no elimina la correlación. Los métodos que apuntan a eliminar tanto la heterocedasticidad como la correlación de los residuos crudos $\widehat{\varepsilon}$ se conocen como métodos de recuperación de errores. La idea general en estos enfoques es transformar los residuos de tal manera que cumplan tener media cero, varianza constante y no ser correlacionados. La matriz $\mathbf{P} \equiv \mathbf{I}_n - \mathbf{H}$ de $n \times n$, es de rango incompleto. Más específicamente, si $n > p$, el rango de $\mathbf{P}$ es igual o menor que $n - p$. En consecuencia, se tiene como máximo $n-p$ residuos transformados no correlacionados. Los residuos obtenidos mediante el uso de métodos de recuperación de errores, a diferencia de los residuos crudos y escalados, pueden representar más de una observación lo que hace difícil su interpretación.

Diagnóstico Residual: En el contexto de los *Modelos Lineales*, el gráfico de diagnóstico más frecuente es el gráfico de los errores $\widehat{\varepsilon}_i$ contra valores ajustados $\widehat{\mu}_i$, éste evalúa si existe algún patrón aleatorio y variabilidad constante a lo largo del eje **x** y también se utiliza para detectar valores atípicos para la variable dependiente. Para las covariables continuas, un gráfico de dispersión de los residuos frente a los valores de la covariable también puede ser utilzado. Si se reconoce un patrón no aleatorio en el gráfico, puede indicar errores de especificación de la forma funcional de la covariable.

Otro gráfico de diagnóstico útil, es el gráfico Normal (Q-Q plot), los cuantiles de residuos ordenados se grafican contra los valores correspondientes para la distribución normal estándar. Si los residuos se distribuyen (aproximadamente) normales, la forma de la gráfica no debe desviarse de una línea recta. Por otro lado, si la distribución de los residuos es simétrica, pero con colas "más gruesas" que la normal, la forma del gráfico se verá como una "S" estirada, de lo contrario, si la distribución es sesgada, la forma del gráfico será como un "arco".

Los residuos crudos son intrínsecamente heterocedásticos y están correlacionados, por esta razón, los diagramas de dispersión se basan preferentemente en los residuos escalados, que se muestran en la Tabla 2.2 ya que tienden a eliminar la heterocedasticidad no deseada contenidas por los residuos crudos.

2.1.2. Heterocedasticidad

Hasta ahora, en el modelo lineal clásico definido en (2.3) y (2.4), se considera varianza homogénea $Var(y_i) = \sigma^2$. En esta sección, se levanta el supuesto de varianza constante y se asume que:

$$Var(y_i) = \sigma_i^2. \tag{2.16}$$

Por lo tanto, el Modelo Lineal con varianza heterogénea viene dado por:

$$y_i = x_i^{(1)}\beta_1 + ... + x_i^{(p)}\beta_p + \varepsilon_i \equiv \mathbf{x}_i^{'}\beta + \varepsilon_i \tag{2.17}$$

y

$$\varepsilon_i \sim N(0, \sigma_i^2) \tag{2.18}$$

donde ε_i son independientes, es decir, ε_i es independiente de $\varepsilon_{i'}$ para $i \neq i'$, la parte fija de *Modelos Lineales*, especificada en (2.17), es exactamente la misma que para *Modelos Lineales* clásicos con varianza homogénea definido en (2.3) y (2.4). La única diferencia entre los dos modelos es el supuesto acerca de la varianza de los residuos en (2.18) en comparación con (2.4). Por lo tanto, de forma similar a (2.6), el modelo con varianza heterogénea, definido en (2.17) y (2.18), cumple:

$$E(y_i) \equiv \mu_i = \mathbf{x}_i^{'}\beta. \tag{2.19}$$

Se observa, que el modelo contiene en total $n + p$ parámetros, n parámetros σ_i y p parámetros β. Se sabe que se tienen n observaciones, esto implica, que el modelo no es identificable. Esto puede evitarse, si se imponen restricciones adicionales a las varianzas residuales $\sigma_1^2, ..., \sigma_n^2$. Una forma simple de imponer tales restricciones es asumir ponderaciones de varianza conocidas. Otra forma más general es representar variaciones más parsimoniosas como una función de un pequeño conjunto de parámetros. Esto se puede lograr mediante el empleo de funciones de varianza que se detallan a continuación.

La forma más sencilla de introducir heterocedasticidad y al mismo tiempo reducir el número de parámetros de varianza en el modelo definido en (2.17) y (2.18), es suponer que la varianza de ε_i es igual a una proporción conocida de un parámetro (desconocido) σ^2. Más específicamente, se puede asociar con cada observación una constante conocida $w_i > 0$ y suponer que $Var(\varepsilon_i) = Var(y_i) = \sigma^2/w_i$.

$$\varepsilon_i \sim N(0, \sigma_i^2/w_i) \tag{2.20}$$

Las constantes w_i se denominan pesos "verdaderos". Cuanto mayor sea el peso para una observación dada, menor será la varianza, es decir, se registrará con mayor precisión el valor de y_i. Sin embargo, en aplicaciones de la vida real, rara vez se conocen los pesos w_i. Generalmente, se supone $w_i = 1$ para todas las observaciones,

esto es, el caso del Modelo Lineal clásico con varianza homogénea, definida por (2.3) y (2.4).

De forma más general y flexible, se puede introducir heterocedasticidad por medio de funciones de varianza:

$$\lambda(\delta, \mu, \mathbf{v}) \tag{2.21}$$

que asume valores positivos, y cumple ser continua y diferenciable con respecto a δ para todos los valores de δ.

Siendo μ un escalar y δ y $\mathbf{v}$ vectores. Entonces la varianza de los errores residuales se define de la siguiente manera:

$$Var(\varepsilon_i) = \sigma^2 \lambda^2(\delta, \mu_i, \mathbf{v}_i) \tag{2.22}$$

con μ_i definida en (2.19), σ parámetro de escala, $\mathbf{v}_i$ vector de covariables (conocidas) que definen la función de varianza para la observación i, mientras que el vector δ contiene un conjunto de parámetros de varianza, común a todas las observaciones. Debido a que la función $\lambda(\cdot)$ en el lado derecho de (2.22) involucra μ_i, la varianza también depende de β. Sin embargo, se refleja esta dependencia mediante el uso de μ_i, es decir, se considera a la dependencia de la varianza del error residual en el valor medio.

El parámetro σ, utilizado en (2.22), debe interpretarse como un parámetro de escala. Esto está en contraste con el Modelo Lineal clásico con varianza homogénea, en donde σ puede interpretarse como una desviación estándar del error residual.

Especificación: Si se considera la especificación del Modelo Lineal con función de varianza $\lambda(\cdot)$ especificada para (2.22), desde la perspectiva de una unidad de observación, se especifica la estructura media, implícita en (2.17), con el supuesto:

$$\varepsilon_i \sim \mathcal{N}(0, \sigma^2 \lambda_i^2) \tag{2.23}$$

donde

$$\lambda_i \equiv \lambda(\delta, \mu_i, \mathbf{v_i}) \tag{2.24}$$

Al usar la función de varianza $\lambda(\cdot)$, se representa la varianza de ε_i como:

$$\sigma_i^2 = \sigma^2 \lambda_i^2 \tag{2.25}$$

donde σ^2 es un parámetro escalar desconocido y λ_i definido en (2.24), depende directamente de los parámetros de varianza desconocidos δ e indirectamente en β a través de μ_i. Por ejemplo, si se supone que $\lambda(\mu_i) = \mu_i$, de (2.25) se sigue que $\sigma_i/\mu_i = \sigma$. Por lo tanto, en el contexto de este modelo, σ puede interpretarse como un *coeficiente de variación*.

Análogamente, se puede generalizar para todos los datos del modelo. Se define la matriz diagonal $\mathcal{R}$:

$$\mathcal{R} = \Lambda^2, \tag{2.26}$$

donde Λ=diag$(\lambda_1, ..., \lambda_n)$ es una matriz diagonal, con elementos definidos por (2.24). Por lo tanto, se especifica el modelo definido por (2.17) y (2.23) - (2.24), de la siguiente manera:

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}, \tag{2.27}$$

donde

$$\boldsymbol{\varepsilon} \sim N(\mathbf{0}, \sigma^2 \mathcal{R}), \tag{2.28}$$

donde $\mathbf{y}, \mathbf{X}, \boldsymbol{\beta}$ y $\boldsymbol{\varepsilon}$ se definen como en *Modelos Lineales* con varianza homogénea.

Conjunto de Funciones de Varianza

Las funciones de varianza $\lambda(\cdot)$ se pueden clasificar en los siguientes cuatro conjuntos:

1. Pesos conocidos, $\lambda(\cdot) = \lambda(\mathbf{v})$
2. Funciones de varianza que dependen de δ pero no de μ, es decir, $\lambda(\cdot) = \lambda(\delta; \mathbf{v})$
3. Funciones de varianza que dependen de δ y μ, es decir, $\lambda(\cdot) = \lambda(\delta, \mu; \mathbf{v})$
4. Funciones de varianza que dependen de μ pero no de δ, es decir, $\lambda(\cdot) = \lambda(\mu; \mathbf{v})$

A continuación, son referidos simbólicamente a los conjuntos 2 - 4 como $< \delta >$, $< \delta, \mu >$ y $< \mu >$, respectivamente.

El uso de una función de varianza de cualquiera de los conjuntos mencionados anteriormente no plantea dificultades en términos de la especificación del modelo. Sin embargo, en los modelos que implican funciones de varianza de los conjuntos $< \delta, \mu >$ o $< \mu >$, los parámetros β se comparten por la media y las estructuras de varianza. Entonces los modelos conocidos como modelos de media-varianza, requieren diferentes enfoques de estimación y técnicas de inferencia, en comparación con los modelos que involucran pesos conocidos o funciones de varianza del conjunto $< \delta >$.

Conjunto	δ	μ_i	Algoritmo de Estimación
Pesos conocidos	-	-	MCO
$< \delta >$	+	-	MV/MVR
$< \delta, \mu >$	+	+	MV/MVR - base MCG
$< \mu >$	-	+	IRLS

Tabla 2.3: Conjuntos de funciones de varianza

La Tabla 2.3 resume los conjuntos de funciones de varianza y sus métodos de estimación correspondientes.

Para profundizar en las funciones de varianza y reflejarlo en la notación, se supone que las observaciones y_i se dividen en varios estratos, indexados por s (s=1,...,S).

En este trabajo se desarrollan las funciones del conjunto $< \delta >$, la cual se usan para ajustar los modelos expuestos más adelante en el capítulo 4. Los conjuntos de funciones de varianza pueden tener esta forma:

λ_i	Descripción
$\lvert v_i \rvert^{\delta_{s_i}}$	Función de Potencia covariable v_i
δ_{s_i}	Diferentes varianzas por estratos (etapa) $\delta_1 \equiv 1, \delta_s > 0$ para $s \neq 1$

Tabla 2.4: Ejemplos de funciones de varianza a partir de los conjuntos $< \delta >$

Las funciones de varianza, presentada en la Tabla 2.4, pertenecen al grupo $< \delta >$. Es decir, dependen de la varianza covariable v_i y de los parámetros $\delta = (\delta_1, ..., \delta_S)$, y no de μ_i. Por lo tanto, son funciones de varianza independientes de la media. La función δ_{s_i} está definida para múltiples estratos (etapas).

Estimación de los parámetros

Los parámetros del modelo pueden estimarse con distintos enfoques, que dependen del tipo de función de varianza establecido en (2.22).
Las ecuaciones (2.25) y (2.28), utilizadas en la especificación del Modelo Lineal con varianza heterogénea, son importantes para los parámetros. Mientras σ se puede pensar como un parámetro de escala, los parámetros δ proporcionan información sobre la magnitud en las observaciones.
Se presenta el método de estimación para los modelos definidos mediante el uso de una función de varianza del conjunto $< \delta >$.

Optimización de verosimilitud: Se considera el modelo, definido por (2.17), (2.23) y (2.24), con la función de varianza $\lambda(\cdot)$ que pertenece al conjunto $< \delta >$, es decir,

$$\lambda_i = \lambda(\delta; \mathbf{v}_i) \tag{2.29}$$

Se observa que, en comparación con la definición general de $\lambda(\cdot)$, dada por (2.22), se consideran las funciones de varianza que dependen del vector de los parámetros de varianza δ y del vector de las covariables (conocidas) $\mathbf{v}_i$. Ejemplos de estas funciones, pueden verse en la Tabla 2.4.

Estimación por Máxima Verosimilitud: Primero se introduce el logaritmo de la función de verosimilitud completa para luego considerar el logaritmo de la verosimilitud, a partir de parámetros β y σ^2.
Log-Verosimilitud para β,σ^2 y δ.

El logaritmo de la función de verosimilitud del modelo, especificado en (2.17), (2.23) y (2.24), viene dada por:

$$l_{Full}(\beta, \sigma^2, \delta) = -\frac{n}{2}\log(\sigma^2) - \frac{1}{2}\sum_{i=1}^{n}\log(\lambda_i^2) - \frac{1}{2\sigma^2}\sum_{i=1}^{n}\lambda_i^{-2}(y_i - \mathbf{x}_i'\beta)^2 \qquad (2.30)$$

Se observa que $l_{Full}(\beta, \sigma^2, \delta)$ depende de δ a través de λ_i, definido en (2.29). Si se cumple que $\lambda_i \equiv 1$, la log-verosimilitud (2.30) resulta equivalente a la log-verosimilitud para el Modelo Lineal clásico.

Las estimaciones de los parámetros β, σ^2 y δ pueden obtenerse si se maximiza simultáneamente la función de log-verosimilitud con respecto a estos parámetros. En general, es una tarea numéricamente compleja que requiere encontrar un óptimo en un espacio de parámetros multidimensional, puede simplificarse mediante la técnica de *verosimilitud de perfil*, que se describe a continuación.

Verosimilitud de Perfil

La creación de perfiles de una función de verosimilitud se puede realizar de varias formas, se decide buscar el perfil en primer lugar de los parámetros β y luego σ^2, por lo que se asume δ conocida en (2.29) y se maximiza (2.30) con respecto a β para cada valor de δ.

De lo anterior implica la siguiente relación entre el valor óptimo $\widehat{\beta}$ y δ:

$$\widehat{\beta}(\delta) \equiv \left(\sum_{i=1}^{n}\lambda_i^{-2}\mathbf{x}_i\mathbf{x}_i'\right)^{-1}\sum_{i=1}^{n}\lambda_i^{-2}\mathbf{x}_i y_i. \qquad (2.31)$$

Si se sustituye (2.31) en (2.30), se obtiene la siguiente función de verosimilitud de perfil:

$$l_{MV}^*(\sigma^2, \delta) \equiv l_{Full}(\widehat{\beta}(\delta), \sigma^2, \delta) = -\frac{n}{2}\log(\sigma^2) - \frac{1}{2}\sum_{i=1}^{n}\log(\lambda_i^2) - \frac{1}{2\sigma^2}\sum_{i=1}^{n}\lambda_i^{-2}r_i^2 \quad (2.32)$$

donde

$$r_i \equiv y_i - \mathbf{x}_i'\widehat{\beta}(\delta) = y_i - \mathbf{x}_i'\left(\sum_{i=1}^{n}\lambda_i^{-2}\mathbf{x}_i\mathbf{x}_i'\right)^{-1}\sum_{i=1}^{n}\mathbf{x}_i\lambda_i^{-2}y_i. \qquad (2.33)$$

y λ_i, definido por (2.29), depende de δ. Se utiliza "*" en (2.32) para indicar que es una función de verosimilitud de perfil, la cual tiene la ventaja de no depender de β y entonces su optimización, se realiza en un espacio de parámetros de menor

dimensión.

Al maximizar $l^*_{MV}(\sigma^2, \delta)$ con respecto a σ^2 para cada valor conocido de δ conduce a la siguiente relación funcional entre el valor óptimo $\widehat{\sigma^2}$ y δ:

$$\widehat{\sigma}^2_{MV}(\delta) \equiv \sum_{i=1}^{n} \lambda_i^{-2} r_i^2 / n \tag{2.34}$$

donde $r_i \equiv r_i(\delta)$ se definen en (2.33). Si se sustituye (2.34) en (2.32) se obtiene la función de verosimilitud de perfil para δ:

$$l^*_{MV}(\delta) \equiv l^*_{MV}(\widehat{\sigma}^2(\delta), \delta) = -\frac{n}{2}\log(\widehat{\sigma}^2_{MV}) - \frac{1}{2}\sum_{i=1}^{n}\log(\lambda_i^2) - \frac{n}{2} \tag{2.35}$$

donde $\widehat{\sigma}^2 \equiv \widehat{\sigma}^2(\delta)$.

La función definido en (2.35) depende de δ no depende de β ni de σ^2. Por lo tanto, su maximización es más directa que la maximización de (2.30), (2.35).

Al maximizar $l^*_{MV}(\delta)$ con respecto a δ, se obtiene un estimador $\widehat{\delta}_{MV}$ de δ. Al reemplazar $\widehat{\delta}_{MV}$ en (2.31) y (2.34) se obtienen los siguientes estimadores $\widehat{\beta}_{MV}$ y $\widehat{\sigma}_{MV}$ de β y σ, respectivamente:

$$\widehat{\beta}_{MV} \equiv \widehat{\beta}(\widehat{\delta}_{MV}) = \left(\sum_{i=1}^{n} \widehat{\lambda}_i^{-2} \mathbf{x}_i \mathbf{x}_i^{'}\right)^{-1} \sum_{i=1}^{n} \widehat{\lambda}_i^{-2} \mathbf{x}_i y_i. \tag{2.36}$$

$$\widehat{\sigma}^2_{MV} \equiv \widehat{\sigma}^2_{MV}(\widehat{\delta}) = \sum_{i=1}^{n} \widehat{\lambda}_i^{-2} \widehat{r}_i^2 / n \tag{2.37}$$

donde $\widehat{\lambda} \equiv \lambda(\widehat{\delta}; \mathbf{v}_i)$ y $\widehat{r}_i \equiv r_i(\widehat{\delta})$ es definida en (2.33).

Al igual que para modelos con varianza homogénea (2.9) a (2.11), el estimador de máxima verosimilitud $\widehat{\sigma}^2_{MV}$ de σ^2, obtenido a partir de maximizar (2.35), es sesgado, lo mismo sucede para $\widehat{\delta}_{MV}$. Por esta razón, σ^2 y δ se estiman preferentemente al utilizar el método de máxima verosimilitud restringida (EMVR), que se describe a continuación.

Estimación de máxima verosimilitud restringida : La idea de la estimación de MVR para los modelos definidos por (2.17), (2.23) y (2.24), con función de varianza perteneciente al conjunto $< \delta >$, es similar a la utilizada en el caso del

Modelo Lineal clásico para observaciones independientes. Es decir, para obtener estimaciones insesgadas de σ^2 y δ. Esto se puede hacer al considerar la función de verosimilitud de un conjunto de contrastes independientes de $n-p$. La función de verosimilitud restringida resultante está dada por

$$\begin{aligned} l_{EMVR}(\sigma^2,\delta) \equiv & -\frac{n-p}{2}\log(\sigma^2) - \frac{1}{2}\sum_{i=1}^{n}\log(\lambda_i^2) - \frac{1}{2\sigma^2}\sum_{i=1}^{n}\lambda_i^{-2}r_i^2 \\ & - \frac{1}{2}\log\left[det\left(\sum_{i=1}^{n}\lambda_i^{-2}\mathbf{x}_i\mathbf{x}_i^{'}\right)\right] \end{aligned} \tag{2.38}$$

con $det(\cdot)$ que denota el determinante de la matriz A y r_i definido en (2.33). Podemos perfilar σ^2 desde $l_{EMVR}(\cdot)$ se observa que, para un valor conocido de δ, la función se maximiza mediante:

$$\widehat{\sigma}^2_{EMVR}(\delta) \equiv \sum_{i=1}^{n}\lambda_i^{-2}r_i^2/(n-p) \tag{2.39}$$

Al sustituir (2.39) en (2.38), se obtiene una función de verosimilitud de perfil restringida que depende solo de δ:

$$l^*_{EMVR}(\delta) \equiv l_{EMVR}(\widehat{\sigma}^2(\delta),\delta). \tag{2.40}$$

Al maximizar (2.40) con respecto a δ, se obtiene un estimador $\widehat{\delta}_{EMVR}$ de δ. El resultante $\widehat{\delta}_{EMVR}$ también se utiliza en (2.31) para calcular el estimador $\widehat{\beta}_{EMVR}$ de β.

Diagnósticos del modelo

En el caso del modelo lineal con varianza heterogénea, definido por (2.17) y (2.23) - (2.24), con función de varianza perteneciente al conjunto $< \delta >$, las herramientas de diagnóstico descritas para modelos lineales con varianza homogénea ya no son válidas. En particular, debido a la heterocedasticidad ni los residuos crudos ni a escala, presentados en las Tablas 2.1 y 2.2, pueden mostrar una dispersión de variabilidad constante cuando se grafican contra los valores predichos. No obstante, los gráficos pueden usarse para detectar patrones sistemáticos que puedan sugerir problemas

con la linealidad de los efectos de las covariables, con observaciones atípicas o que permitan también detectar patrones en la heterogeneidad de la varianza residual. Para verificar la homocedasticidad y las observaciones atípicos, los residuos de Pearson son adecuados. Se obtienen al escalar apropiadamente los residuos crudos, como se describió anteriormente.

Residuos de Pearson: Cuando describimos los modelos lineales con varianza homogénea, se consideran los residuos de escala dividiendo los residuos crudos por las estimaciones de σ Tabla 2.1. Otro conjunto de residuos a escala, que se muestra en la Tabla 2.2, implica un ajuste adicional basado en la matriz H. Como ya se mencionó, el uso de estos residuos para *Modelos Lineales* con varianza heterogénea es limitado.

Residuos de Pearson	Formulación matemática
Estandarizados por $\sqrt{Var(y_i)}$	$\widehat{\varepsilon}_i/\sqrt{Var(y_i)}$
Estudentizados internos	$\widehat{\varepsilon}_i/\sqrt{\widehat{Var}(y_i)}$
Estudentizados externos	$\widehat{\varepsilon}_i/\sqrt{\widehat{Var}(y_{-i})}$

Tabla 2.5: Ejemplos de residuos de Pearson – ML varianza heterogénea.

Un conjunto diferente de residuos escalados, también útil en el contexto de *Modelos Lineales* con varianza heterogénea, se obtiene dividiendo los residuos crudos por la desviación estándar estimada de la variable dependiente, $[\widehat{Var}(y_i)]^{1/2}$. Los residuos resultantes se denominan *estudentizados internos* y se presentan en la Tabla 2.5, por simplicidad se denominan como *residuos de Pearson*, su principal ventaja es que reducen la varianza con respecto a los residuos crudos, aunque no la eliminan por completo. Adicionalmente, tampoco se elimina la correlación entre los residuos de Pearson.

2.1.3. Modelos lineales de efectos fijos con datos correlacionados

El supuesto principal para los Modelos lineales considerados, es que las observaciones relevadas durante el estudio son independientes entre sí. Este supuesto es restrictivo en casos de estudio con datos correlacionados, como es el caso de la presente investigación. Es importante distinguir, entre *unidades de muestreo* (por ejemplo, sujetos en un estudio longitudinal) y *unidades de análisis* (por ejemplo, medidas específicas de tiempo).

En esta sección, se consideran *Modelos Lineales* más generales que permiten levantar los supuestos de independencia y homocedasticidad, denominados **Modelos Lineales con efectos fijos y errores residuales correlacionados para datos agrupados**, o simplemente **Modelos Lineales para datos correlacionados**.

El objetivo de esta sección es describir los conceptos fundamentales de la teoría de *Modelos Lineales* para datos correlacionados, en particular, se presenta la noción de estructura de correlación, el cual es un concepto general que también se aplica a los *Modelos Lineales Mixtos* que se describirán en la próxima sección.

Si se combinan estructuras de correlación con funciones de varianza, se pueden especificar formas flexibles de matrices de varianza-covarianza para un Modelo Lineal para datos correlacionados.

Especificación del Modelo

En esta sección, se especifican los *Modelos Lineales* con efectos fijos y errores residuales correlacionados para datos agrupados con estructura jerárquica. Para facilitar la presentación, se parte de datos con un solo nivel de agrupamiento, con N grupos (niveles de un factor de agrupamiento) indexados por i ($i = 1, ..., N$) y n_i observaciones por grupo indexado por j ($j = 1, ..., n_i$).

Más concretamente, se asume que para el grupo i, el modelo para una variable

dependiente continua y_i se expresa como

$$\mathbf{y}_i = \mathbf{X}_i\beta + \varepsilon_i, \tag{2.41}$$

donde

$$y \equiv \begin{pmatrix} y_{i1} \\ \vdots \\ y_{ij} \\ \vdots \\ y_{in_j} \end{pmatrix}, \boldsymbol{\varepsilon} \equiv \begin{pmatrix} \boldsymbol{\varepsilon}_{i1} \\ \vdots \\ \boldsymbol{\varepsilon}_{ij} \\ \vdots \\ \boldsymbol{\varepsilon}_{in_j} \end{pmatrix}, \tag{2.42}$$

y

$$\mathbf{X} \equiv \begin{pmatrix} x_{i1}^{1} & x_{i1}^{(2)} & \cdots & x_{i1}^{(p)} \\ \vdots & \vdots & \ddots & \vdots \\ x_{in_i}^{(1)} & x_{in_i}^{(2)} & \cdots & x_{in_i}^{(p)} \end{pmatrix} \equiv \begin{pmatrix} \mathbf{x}_i^{(1)} & \mathbf{x}_i^{(2)} & \cdots & \mathbf{x}_i^{(p)} \end{pmatrix} \tag{2.43}$$

β se define en (2.5), $\mathbf{X}_i$ es una matriz de diseño para el $i-ésimo$ grupo, y se supone que el vector de los errores residuales ε_i dentro del grupo, tiene distribución normal multivariada, esto es:

$$\varepsilon_i \sim \mathcal{N}_{n_i}(\mathbf{0}, \mathcal{R}_i) \tag{2.44}$$

donde la matriz de varianza - covarianza $\mathcal{R}_i$

$$\mathcal{R}_i = \sigma^2\mathbf{R}_i \tag{2.45}$$

σ^2 denota un parámetro de escala desconocido.

Finalmente, se supone que los vectores de errores residuales para diferentes grupos son independientes, es decir, ε_i es independiente de $\varepsilon_{i'}$ para $i \neq i'$. La media y varianza de y_i resultan:

$$E(y_{ij} \equiv \mu_{ij}) = \mathbf{x}_{ij}^{'}\beta \tag{2.46}$$

$$Var(y_i) = \sigma^2\mathbf{R}_i \tag{2.47}$$

La formulación de los modelos descritos anteriormente permite datos con más de un nivel de agrupamiento. Múltiples niveles de agrupamiento se reflejarían mediante la introducción de factores, relacionados con los diferentes niveles de grupo, en la

matriz de diseño X_i, y si se asume una forma particular de la matriz de varianza-covarianza $\mathbf{R}_i$. En lo que resta de esta sección, se hace foco en los modelos para los datos con un solo nivel de agrupamiento.

Detalles de la especificación del modelo: Es importante tener en cuenta que el Modelo Lineal con errores correlacionados, especificado por (2.41) a (2.45), no es identificable en su forma más general. Esto se debe a que su representación (2.45) no es única y porque el modelo potencialmente implica demasiados parámetros desconocidos asociados con la matriz de varianza-covarianza de los errores residuales ε_i. El problema es similar al descrito para el de *Modelos Lineales* con varianza heterogénea.

El modelo (2.41) a (2.45) puede llegar a ser identificable si imponemos restricciones adicionales a las matrices residuales de varianza-covarianza $\mathcal{R}_i$, definido en (2.45) que puede descomponerse como:

$$\mathbf{R}_i = \mathbf{\Lambda}_i \mathbf{C}_i \mathbf{\Lambda}_i \tag{2.48}$$

donde $\mathbf{\Lambda}_i$ es una matriz diagonal con elementos no negativos y $\mathbf{C}_i$ es una matriz de correlación. Al usar $\mathbf{\Lambda}_i$ en (2.48), implica heterocedasticidad, mientras la matriz $\mathbf{C}_i$ implica correlación, en ambos casos para de las observaciones dentro del grupo.

Si se utilizan conjuntos de parámetros disjuntos para $\mathbf{C}_i$ y $\mathbf{\Lambda}_i$, se usa la descomposición (2.48) para modelar $\mathbf{R}_i$. Más concretamente, se supone que los elementos de la matriz diagonal $\mathbf{\Lambda}_i$ se expresan como

$$\{\Lambda_i\}_{j,j} \equiv \lambda_{i_{j,j}} = \lambda(\mu_{ij}, \delta, \mathbf{v}_{ij}), \tag{2.49}$$

donde $\lambda(\cdot)$ es una función de varianza definida anteriormente en (2.21).

Análogamente a (2.22), δ es un vector de parámetros de varianza y $\mathbf{v}_{ij}$ es un vector de varianzas covariadas (conocidas). Por lo tanto, (2.48) debe escribirse como:

$$\mathbf{R}_i(\mu_{ij}, \theta_R; \mathbf{v}_{ij}) = \Lambda_i(\mu_{ij}, \delta; \mathbf{v}_{ij}) \mathbf{C}_i(\varrho) \Lambda_i(\mu_{ij}, \delta; \mathbf{v}_{ij}) \tag{2.50}$$

donde $\theta_R \equiv (\delta', \varrho')'$[5] y $\mathbf{C}_i$ se especifica al utilizar un conjunto de parámetros ϱ[6].
El modelo lineal clásico, especificado anteriormente, se obtiene como un caso particular del modelo (2.41) a (2.45), con $\mathbf{R}_i$ dado por (2.50), cuando $n_i = 1$ y que $\mathbf{R}_i = 1$ para todo i.

Además, los *Modelos Lineales* para observaciones heterocedásticas independientes, especificadas anteriormente, puede verse como un caso especial del modelo (2.41) a (2.45), con $\mathbf{R}_i$ dado por (2.50), si se supone que $n_i = 1$ y que $\mathbf{R}_i = \lambda_i^2$, donde λ_i se define en (2.24).

Estructura de varianza

De manera similar al caso del Modelo Lineal para observaciones independientes con varianza heterogénea los elementos de la matriz $\mathbf{\Lambda}_i$, dados en (2.50), se definen al incorporar una función de varianza. Para datos con un solo nivel de agrupamiento, la definición de la función de varianza presentada en (2.22) se describe como

$$Var(\varepsilon_{ij}) = \sigma^2 \lambda^2(\mu_{ij}, \delta; \mathbf{v}_{ij}) \tag{2.51}$$

donde μ_{ij} es el valor medio dado en (2.46), $\mathbf{v}_{ij}$ es un vector de varianza covariadas (conocidas), δ es un vector de parámetros de covarianza y $\lambda(\cdot)$ es una función de varianza continua con respecto a δ.

Para el caso en que las funciones de varianza no dependen del valor medio, (2.51) resulta:

$$Var(\varepsilon_{ij}) = \sigma^2 \lambda^2(\delta; \mathbf{v}_{ij}) \tag{2.52}$$

[5]para simplificar la notación, a menudo se suprime el uso de θ_R, μ_{ij} y $\mathbf{v}_{ij}$ en las fórmulas
[6]Se definen más adelante.

Estructura de correlación

En esta sección, se presentan algunas matrices de estructuras de correlación $\mathbf{C}_i$, usadas para esta investigación definida en (2.50). La matriz $\mathbf{C}_i$ se especifica al suponer que el coeficiente de correlación entre dos errores residuales ε_{ij} y $\varepsilon_{ij}^{'}$ correspondientes a dos observaciones del mismo grupo i, viene dado por:

$$Corr(\varepsilon_{ij}, \varepsilon_{ij}^{'}) = h[d(\mathbf{t}_{ij}, \mathbf{t}_{ij'}), \varrho)] \tag{2.53}$$

donde ϱ es un vector de parámetros de correlación, d es una función de distancia entre los vectores de las variables de posición $\mathbf{t}_{ij}$ y $\mathbf{t}_{ij'}$ correspondientes a ε_{ij} y $\varepsilon_{ij'}$ respectivamente y $h(\cdot, \cdot)$ es una función continua con respecto a ϱ, toma valores entre -1 y 1 y $h(0, \varrho) \equiv 1$.
Las estructuras de correlación, dependerán de las distancias y funciones de correlación elegidas. Las estructuras de correlación se pueden clasificar en dos conjuntos principales:

1. Estructuras "en serie": se definen en el contexto de series temporales o datos longitudinales
2. Estructuras "espaciales": se definen con datos espacial.

Para las características de esta investigación se profundizará el primer grupo.

Estructuras de correlación en serie: Para las estructuras de correlación de este grupo, se asume que $\mathbf{t}_{ij}$ son escalares enteros positivos, es decir, $\mathbf{t}_{ij} \equiv j$, que describe la posición de la observación en una serie temporal/longitudinal.
La estructura de correlación en serie más simple es la simetría compuesta (*corCompSymm*), que asume una correlación constante entre todos los errores residuales dentro del grupo. Esto significa que

$$Corr(\epsilon_{ij}, \epsilon_{ij'}) = \varrho, \tag{2.54}$$

que corresponde a (2.53) al definir, para $j \neq j^{'}$ y $k = 1, 2, ...,$

$$h(k, \varrho) \equiv \varrho. \tag{2.55}$$

Existen otras funciones de correlación en serie y espaciales $h(\cdot, \cdot)$ que no se describen en esta investigación.

Estimación

El objetivo principal de ajustar el modelo (2.41) a (2.45) a los datos, es obtener estimaciones para los parámetros β, σ^2 y θ_R. En la presente sección, se presentan métodos para estimar los parámetros, la elección del método de estimación depende de la forma de la función de varianza. En este caso, interesa detallar los enfoques de estimación para modelos más simples que utilizan funciones de varianza del grupo $< \delta >$.

Estimación basada en la verosimilitud: En el caso en que la función de varianza pertenece al grupo $< \delta >$ se utilizan las estimaciones MV o EMVR. Por lo tanto, la función de verosimilitud de perfil para el modelo (2.41) a (2.45) está dada por:

$$\begin{aligned} l_{Full}(\beta, \sigma^2, \theta_R) = &-\frac{n}{2}\log(\sigma^2) - \frac{1}{2}\sum_{i=1}^{n}\log[det(R)_i] \\ &-\frac{1}{2\sigma^2}\sum_{i=1}^{n}(y_i - \mathbf{X}_i\beta)'\mathbf{R}_i^{-1}(y_i - \mathbf{X}_i\beta) \end{aligned} \tag{2.56}$$

Se observa que $l_{Full}(\cdot)$ depende de θ_R a través de $\mathbf{R}_i \equiv \mathbf{R}_i(\theta_R)$. Las estimaciones de los parámetros β, σ^2 y σ pueden obtenerse mediante una maximización simultánea de la función de log-verosimilitud con respecto a estos parámetros, tarea numéricamente compleja. Una alternativa es considerar el perfil β de (2.56). Con este objetivo, al suponer θ_R conocido, (2.56) se maximiza con respecto a σ^2, resulta:

$$\widehat{\beta}(\theta_R) \equiv \left(\sum_{i=1}^{N}\mathbf{X}_i'\mathbf{R}_i^{-1}\mathbf{X}_i\right)^{-1}\sum_{i=1}^{N}\mathbf{X}_i'\mathbf{R}_i^{-1}\mathbf{y}_i, \tag{2.57}$$

$$\widehat{\sigma}^2(\theta_R) \equiv \sum_{i=1}^{N} \mathbf{r}_i' \mathbf{R}_i^{-1} \mathbf{r}_i / n, \tag{2.58}$$

donde $\mathbf{r}_i \equiv \mathbf{r}_i(\theta_R) = \mathbf{y}_i - \mathbf{X}_i \widehat{\beta}(\theta_R)$ y $\mathbf{R}_i \equiv \mathbf{R}_i(\theta_R)$. Las expresiones corresponden a (2.31) y (2.34), presentadas para el LM con varianza heterogénea.

Si se sustituye (2.57) en (2.56), se obtiene la función de verosimilitud de perfil, que depende de σ^2 y θ_R:

$$l^*_{MV}(\sigma^2, \theta_R) \equiv l_{Full}(\widehat{\beta}(\theta_R), \sigma^2, \theta_R) \tag{2.59}$$

La maximización de (2.59) sobre σ^2 devuelve el estimador dado en (2.58). Al suplantar el estimador en (2.59), se obtiene una función log-verosimilitud de perfil que solo depende de θ_R:

$$l^*_{MV}(\theta_R) \equiv l_{Full}(\widehat{\beta}(\theta_R), \widehat{\sigma}^2, \theta_R) \tag{2.60}$$

Al maximizar la función, obtenemos el estimador $\widehat{\theta}_R$ de θ_R y al sustituirlo en (2.57) y (2.58), se obtienen los estimadores de β y σ^2, respectivamente.

Los estimadores de σ^2 y $\widehat{\theta}_R$ resulta sesgados, por lo tanto, σ^2 y θ_R a menudo se estiman al maximizar la siguiente función de log-verosimilitud restringida:

$$\begin{aligned} l^*_{EMVR}(\sigma^2, \theta_R) \equiv l_{Full}(\widehat{\beta}(\theta_R), \sigma^2, \theta_R) + \frac{p}{2} \log(\sigma^2) \\ -\frac{1}{2} \left[det \left(\sum_{i=1}^{N} \mathbf{X}_{ii}'^{-1} \mathbf{X}_i \right) \right] \end{aligned} \tag{2.61}$$

donde $\widehat{\beta}(\theta_R)$ es especificada en (2.57).

El parámetro σ^2 se puede perfilar a partir de la función de log-verosimilitud restringida (2.61). Es decir, σ^2 se expresa con la siguiente fórmula:

$$\widehat{\sigma}^2(\theta_R) \equiv \sum_{i=1}^{N} \mathbf{r}_i' \mathbf{R}_i^{-1} \mathbf{r}_i / (n - p), \tag{2.62}$$

que resulta de la maximización de (2.61) sobre σ^2, donde $\mathbf{r}_i$ se especifica en (2.58).

A partir de (2.62) y (2.61) se obtiene:

$$l^*_{EMVR}(\theta_R) \equiv l^*_{EMVR}(\sigma^2(\theta_R), \theta_R) \tag{2.63}$$

Al sustituir el estimador de θ_R en (2.57) y (2.62) se obtienen las estimaciones EMVR de β y σ^2.

2.1.4. Modelos lineales de efectos mixtos

En la sección anterior, se presentaron modelos con efectos fijos para datos correlacionados. Un ejemplo de datos agrupados, son los datos longitudinales con múltiples mediciones recopiladas a lo largo del tiempo para un individuo. Este es el caso de la presente investigación, datos con un solo nivel de agrupación, con N grupos (etapas) indexados por $i = 2, 3$ y 4, cada uno con n_i observaciones.

Los *Modelos Lineales Mixtos* permiten considerar correlación de las observaciones contenidas en un conjunto de datos. Además, permiten dividir efectivamente la variación general de la variable dependiente en componentes que corresponden a diferentes niveles de jerarquía de datos.

En la presente sección se describen los enfoques de estimación, herramientas de diagnóstico y métodos inferenciales utilizados para los Modelos Lineles Mixtos más frecuentes, cuya matriz de varianza-covarianza residual (condicional) es independiente del valor medio.

El modelo de efectos mixtos lineales clásicos

Para los datos jerárquicos con un único nivel de agrupamiento, se puede formular el Modelo Lineal Mixto clásico para un nivel dado de la siguiente manera:

$$\mathbf{y}_i = \mathbf{X}_i\beta + \mathbf{Z}_i\mathbf{b}_i + \varepsilon_i, \tag{2.64}$$

donde $\mathbf{y}_i$, $\mathbf{X}_i$, β_i y ε_i son: el vector de variables respuesta continuo, la matriz de diseño y el vector de errores residuales para el grupo i, especificado en (2.42) y (2.43), respectivamente. Mientras que $\mathbf{Z}_i$ y $\mathbf{b}_i$ son la matriz de covariables y su correspondiente vector de efectos aleatorios:

$$\mathbf{Z}_i \equiv \begin{pmatrix} z_{i1}^{(1)} & z_{i1}^{(2)} & \cdots & z_{i1}^{(q)} \\ \vdots & \vdots & \ddots & \vdots \\ z_{in_i}^{(1)} & z_{in_i}^{(2)} & \cdots & z_{in_i}^{(q)} \end{pmatrix} = \begin{pmatrix} \mathbf{z}_i^{(1)} & \mathbf{z}_i^{(2)} & \cdots & \mathbf{z}_i^{(q)} \end{pmatrix}, \ \mathbf{b}_i \equiv \begin{pmatrix} b_{i1} \\ \vdots \\ b_{iq} \end{pmatrix} \tag{2.65}$$

De forma similar a la matriz de diseño $\mathbf{X}_i$, la matriz $\mathbf{Z}_i$ contiene valores conocidos de q covariables, con los correspondientes efectos no observables $\mathbf{b}_i$. Además,

$$\mathbf{b}_i \sim N_q(\mathbf{0}, \mathcal{D}),\ \varepsilon_i \sim N_{n_i}(\mathbf{0}, \mathcal{R}_i),\ \text{ con } \mathbf{b}_i \perp \varepsilon_i \tag{2.66}$$

es decir que para un mismo grupo, los errores residuales ε_i son independientes de los efectos aleatorios $\mathbf{b}_i$. Este supuesto es lo que distingue un Modelo Lineal Mixto clásico de un Modelo Lineal Mixto extendido. Adicionalmente se supone, que los vectores de efectos aleatorios y errores residuales para diferentes grupos son independientes entre sí, o sea, $\mathbf{b}_i$ es independiente de $\varepsilon_{i'}$ para $i \neq i'$. También se especifica que

$$\mathcal{D} = \sigma^2\mathbf{D} \text{ y } \mathcal{R} = \sigma^2\mathbf{R}_i, \tag{2.67}$$

donde σ^2 es un parámetro de escala desconocido. En general, se supone que $\mathcal{D}$ y $\mathcal{R}$ son definidas positivas, a menos que se indique lo contrario. La representación (2.67) en su forma general, no es única. Para que resulte identificable, se especifica la estructura de la matriz $\mathbf{R}_i$ en términos de un conjunto de parámetros para una función de varianza y una matriz de correlación. Esto implica suponer restricciones en la definición de $\mathbf{R}_i$ en (2.67).

Además de los parámetros de efectos fijos β para las covariables utilizadas en la construcción de la matriz de diseño $\mathbf{X}_i$, el modelo (2.64) incluye dos componentes aleatorios: los errores residuales dentro del grupo ε_i y los efectos aleatorios $\mathbf{b}_i$ para las covariables incluidas en la matriz $\mathbf{Z}_i$. La presencia de efectos fijos y aleatorios de variables conocidas da lugar al nombre del modelo.

En muchos casos, los efectos (aleatorios) incluidos en $\mathbf{b}_i$ tienen efectos (fijos), contenidos en β. En consecuencia, la matriz $\mathbf{Z}_i$ se crea seleccionando un subconjunto de columnas apropiadas de la matriz $\mathbf{X}_i$. En tal situación, se dice que los correspondientes efectos fijos y aleatorios están "acoplados".

El modelo (2.64) a (2.67) se conoce comúnmente como *modelo de dos etapas* o *modelo de dos niveles*, también llamado como *Modelo Lineal Mixto de un solo nivel*(1) porque se aplica a una jerarquía de datos definida por un único nivel de agrupación.

Un modelo de datos con dos niveles de agrupación, con observaciones agrupadas en N grupos de primer nivel (indexados por $i = 1, ..., N$), cada uno con n_i de segundo nivel (sub-) grupos (indexados por $j = 1, ..., ni$) que contiene observaciones n_{ij}, se puede escribir como

$$\mathbf{y}_{ij} = \mathbf{X}_{ij}\beta + \mathbf{Z}_{1,ij}\mathbf{b}_i + \mathbf{Z}_{2,ij}\mathbf{b}_{ij} + \varepsilon_{ij}, \tag{2.68}$$

con

$$\mathbf{b}_i \sim N_{q_1}(\mathbf{0}, \mathcal{D}_1), \quad \mathbf{b}_{ij} \sim N_{q_2}(\mathbf{0}, \mathcal{D}_2), \quad \text{y} \quad \varepsilon_{ij} \sim N_{n_{ij}}(\mathbf{0}, \mathcal{R}_{ij}),$$

donde los vectores aleatorios $\mathbf{b}_i$, $\mathbf{b}_{ij}$ y ε_{ij} son independientes entre sí. En el modelo (2.68), $\mathbf{b}_i$ son los efectos aleatorios asociados con los grupos de primer nivel, mientras que $\mathbf{b}_{ij}$ son los efectos aleatorios asociados con los grupos de segundo nivel, independientes a los efectos aleatorios de primer nivel. Las matrices de diseño $\mathbf{Z}_{1,ij}$ y $\mathbf{Z}_{2,ij}$ pueden ser idénticas[7]. Este modelo entonces, puede denominarse como un *Modelo Lineal Mixto de dos niveles.*

Especificación para todos los datos

Se puede generalizar el Modelo Lineal Mixto de un solo nivel, dado por (2.64) - (2.67), se obtiene un modelo más complejo denominado *Modeo Lineal Mixto multinivel.*

Sea $\mathbf{y} \equiv (\mathbf{y}_1^{'}, \mathbf{y}_2^{'}, ..., \mathbf{y}_N^{'})^{'}$ el vector que contiene todos los valores observados $n = \sum_{i=1}^{N} n_i$ de la variable dependiente y sea $\mathbf{b} \equiv (\mathbf{b}_1^{'}, \mathbf{b}_2^{'}, ..., \mathbf{b}_N^{'})^{'}$ y $\varepsilon \equiv (\varepsilon_1^{'}, \varepsilon_2^{'}, ..., \varepsilon_N^{'})^{'}$ los vectores de efectos aleatorios Nq y los errores residuales n, respectivamente. Se definen las matrices:

$$\mathbf{X} \equiv \begin{bmatrix} \mathbf{X}_1 \\ \mathbf{X}_2 \\ \vdots \\ \mathbf{X}_N \end{bmatrix} \quad \text{y} \quad \mathbf{Z} \equiv \begin{bmatrix} \mathbf{Z}_1 & \mathbf{0} & \cdots & \mathbf{0} \\ \mathbf{0} & \mathbf{Z}_2 & \cdots & \mathbf{0} \\ \vdots & \vdots & \ddots & \vdots \\ \mathbf{0} & \mathbf{0} & \cdots & \mathbf{Z}_N \end{bmatrix}, \tag{2.69}$$

donde $\mathbf{0}$ denota matriz con todos sus elementos nulos. Por lo general, $\mathbf{X}$ es de dimensión $n \times p$, mientras que $\mathbf{Z}$ es de dimensión $n \times Nq$.

[7] no tienen por qué serlo.

Los modelos (2.64) - (2.67) pueden generalizarse para todos los datos de la siguiente manera:

$$\mathbf{y} = \mathbf{X}\beta + \mathbf{Z}\mathbf{b} + \varepsilon, \tag{2.70}$$

con

$$\mathbf{b} \sim \mathcal{N}_{Nq}(\mathbf{0}, \sigma^2 \mathrm{D}), \quad \text{y} \quad \varepsilon \sim \mathcal{N}_n(\mathbf{0}, \sigma^2 \mathrm{R}), \tag{2.71}$$

donde

$$\mathrm{D} \equiv \mathbf{I}_N \otimes \mathbf{D} = \begin{bmatrix} \mathbf{D} & \mathbf{0} & \cdots & \mathbf{0} \\ \mathbf{0} & \mathbf{D} & \cdots & \mathbf{0} \\ \vdots & \vdots & \ddots & \vdots \\ \mathbf{0} & \mathbf{0} & \cdots & \mathbf{D} \end{bmatrix}, \quad \text{y} \quad \mathbf{R} \equiv \begin{bmatrix} \mathbf{R}_1 & \mathbf{0} & \cdots & \mathbf{0} \\ \mathbf{0} & \mathbf{R}_2 & \cdots & \mathbf{0} \\ \vdots & \vdots & \ddots & \vdots \\ \mathbf{0} & \mathbf{0} & \cdots & \mathbf{R}_N \end{bmatrix}, \tag{2.72}$$

donde $\otimes$ denota el producto Kronecker.

Vale la pena señalar, que la forma diagonal en bloques de las matrices $\mathbf{Z}$, $\mathbf{D}$ y $\mathbf{R}$, dada en (2.69), (2.71) y (2.72), respectivamente, es resultado de que el Modelo Lineal Mixto de un solo nivel, definido por (2.64) - (2.67), asume jerarquía particular de datos y efectos aleatorios, como se muestra explícitamente en (2.66). En particular, el modelo asume que los efectos aleatorios para diferentes grupos, son independientes. Sin embargo, es posible formular modelos de efectos aleatorios utilizando la representación (2.70) con las matrices no diagonales en bloques $\mathbf{Z}$, D y $\mathbf{R}$.

El modelo lineal de efectos mixtos extendido

Suponer que los errores residuales ε_i son independientes de los efectos aleatorios $\mathbf{b}_i$, como se especifica en (2.66), resulta ser demasiado restrictivo. Si se levanta este supuesto, se obtiene un *Modelo Lineal Mixto extendido*. El modelo se especifica usando (2.64) - (2.65) y reemplazando (2.66) por

$$\mathbf{b}_i \sim \mathcal{N}_q(\mathbf{0}, \mathcal{D}), \quad \text{y} \quad \varepsilon | \mathbf{b}_i \sim \mathcal{N}_n(\mathbf{0}, \sigma^2 \mathrm{R}), \tag{2.73}$$

con $\mathcal{D}$ y $\mathcal{R}_i$ definidas en (2.67), se está frente al caso de una especificación jerárquica. En el caso que ε en (2.73) sea independiente de los efectos aleatorios, entonces se

obtiene el *Modelo Lineal Mixto clásico* especificado en (2.64) - (2.67). Por lo tanto, el *Modelo Lineal Mixto extendido* permite un enfoque generalizado del *Modelo Lineal Mixto clásico*. Una especificación jerárquica de un Modelo Lineal Mixto extendido de dos niveles, definido en (2.68), equivale a suponer que:

$$\mathbf{b}_i \sim \mathcal{N}_{q_1}(\mathbf{0}, \mathcal{D}_\infty), \quad \mathbf{b}_{ij}|\mathbf{b_i} \sim \mathcal{N}_{q_2}(\mathbf{0}, \mathcal{D}_2), \quad \text{y} \quad \varepsilon_{ij}|\mathbf{b}_i, \mathbf{b}_{ij} \sim \mathcal{N}_{n_{ij}}(\mathbf{0}, \mathcal{R}_{ij}),$$

Distribuciones definidas por las variables aleatorias y y b

Tanto los *Modelos Lineales Mixtos clásicos* como los *Modelos Lineales Mixtos extendidos* introducen las variables aleatorias continuas y y b que se describen mediante funciones de densidad. El primer caso, es una distribución incondicional de efectos aleatorios $\mathbf{b}$ (no observados), definidos por (2.71). El segundo caso, es una distribución condicional de la variable dependiente (aleatoria). Se supone, que los efectos aleatorios son conocidos.

En esta sección, se describen más detalladamente las distribuciones que completan la especificación del modelo para los *Modelo Lineal Mixto clásicos* y *Modelo Lineal Mixto extendidos*. Se introducen también, las distribuciones auxiliares adicionales relacionadas con las variables aleatorias $\mathbf{y}$ y $\mathbf{b}$.

Distribución incondicional de efectos aleatorios La distribución incondicional $f_b(\mathbf{b}_i)$ de los efectos aleatorios $\mathbf{b}_i$, definida por (2.66), es una distribución normal multivariante con media cero y matriz de varianza-covarianza $\mathcal{D}$. Al considerar (2.67), se define:

$$\mathcal{D}(\sigma^2, \theta_D) = \sigma^2 \mathbf{D}(\theta_D), \tag{2.74}$$

donde θ_D es vector de parámetros que representa las varianzas (escaladas por σ^2) y las covarianzas de los elementos de $\mathbf{b}_i$. De acuerdo con (2.74), la matriz $\mathbf{D}$, utilizada para definir la matriz de varianza-covarianza de los efectos aleatorios $\mathbf{b}_i$, se parametriza utilizando un vector de parámetros θ_D. En muchos casos, se supone que dos elementos cualquiera del vector $\mathbf{b}_i$ se pueden correlacionar sin imponer restricciones a la matriz $\mathcal{D}$, excepto que es definido-positivo y simétrico. En este caso, $\mathcal{D}$ tiene una

estructura general de una matriz definida positiva, con $q(q+1)/2$ elementos distintos correspondientes a q varianzas y $q(q-1)/2$ covarianzas de los efectos aleatorios incluidos en $\mathbf{b}_i$. Por lo tanto, θ_D contiene $q(q+1)/2$ parámetros distintos. Aunque q es típicamente pequeño, la estimación de todos los parámetros puede ser difícil, por tal motivo, se puede elegir una estructura simplificada de la matriz $\mathcal{D}$, por ejemplo, se puede suponer que presenta forma diagonal[8], así, la verosimilitud dependerá de los datos disponibles y θ_D contendrá q parámetros distintos.

Distribución condicional de y dados los efectos aleatorios Se observa que, de (2.64) a (2.67), la distribución condicional para los *Modelos Lineales Mixtos* clásicos, $f_{\mathbf{y}|\mathbf{b}}(\mathbf{y}_i|\mathbf{b}_i)$, de $\mathbf{y}_i$ dada $\mathbf{b}_i$ es normal multivariada, con la media y la varianza definidas como:

$$E(\mathbf{y}_i|\mathbf{b}_i) \equiv \mu_i = \mathbf{X}_i\beta + \mathbf{Z}_i\mathbf{b}_i \tag{2.75}$$

$$Var(\mathbf{y}_i|\mathbf{b}_i) = \sigma^2\mathbf{R}_i \tag{2.76}$$

con $\mu_i \equiv (\mu_{i1}, ..., \mu_{i,n_i})'$ y

$$E(y_{ij}|\mathbf{b}_i) \equiv \mu_{ij} = \mathbf{x}_{ij}'\beta + \mathbf{z}_{ij}'\mathbf{b}_i, \tag{2.77}$$

donde $\mathbf{x}_{ij} = (x_{ij}^{(1)}, ..., x_{ij}^{(p)})'$ y $\mathbf{z}_{ij} = (z_{ij}^{(1)}, ..., z_{ij}^{(q)})'$ son vectores columna que contienen los valores de los predictores X y Z para la observación i-ésima del grupo j-ésimo. Por lo tanto, el valor medio del vector variable dependiente $\mathbf{y}_i$ condicionado a los valores (desconocidos) de los efectos aleatorios $\mathbf{b}_i$, se define mediante una combinación lineal de los vectores de las covariables X y Z incluidas en el grupo de las matrices de diseño $\mathbf{X}_i$ y $\mathbf{Z}_i$ correspondientes a los efectos fijos β y los efectos aleatorios $\mathbf{b}_i$, respectivamente. Además, la matriz de varianza-covarianza condicional de $\mathbf{y}_i$ es igual a la matriz de varianza-covarianza de los errores residuales ε_i.

De forma genérica, los Modelo Lineal Mixto no son identificables, dado que no tienen una única representación (2.67) y a que potencialmente contienen demasiados parámetros desconocidos. Para hacerlos identificables, de manera similar a la matriz $\mathcal{D}$, podemos considerar la representación de elementos de $\mathcal{R}_i$ como funciones de un

[8] esto es equivalente a suponer que todos los elementos del vector $\mathbf{b}_i$ son independientes.

conjunto limitado de parámetros θ_R, distintos de θ_D.

Para la matriz $\mathcal{R}_i$, se puede considerar la descomposición dada por (2.50), y combinarla con el uso de funciones de varianza y estructuras de correlación. Se obtiene entonces, $\mathcal{R}_i$. De esta forma, no solo se reduciría el número de parámetros del modelo, sino que la representación (2.67) se volvería identificable.

Para las funciones independientes de la media, como las del grupo $< \delta >$, al combinar el uso de la función de varianza con una estructura de correlación se establece:

$$Var(\varepsilon_i|\mathbf{b}_i) = Var(\varepsilon_i) = \sigma^2\mathbf{R}_i(\theta_R; \mathbf{v}_i), \tag{2.78}$$

Se observa que (2.78) concuerda con el supuesto de que los errores residuales ε_i son independientes de los efectos aleatorios $\mathbf{b}_i$. Por lo tanto, la especificación del modelo jerárquico con funciones de varianza independientes de la media conduce al *Modelo Lineal Mixto clásico*, con $\mathcal{R}_i = \sigma^2\mathbf{R}_i(\theta_R; \mathbf{v}_i)$.

La elección de la estructura de las matrices $\mathcal{D}$ y $\mathcal{R}_i$ o de forma equivalente, $\mathbf{D}$ y $\mathbf{R}_i$ incide en la forma de la matriz de varianza marginal del vector $\mathbf{y}_i$, implícita en el modelo (2.64) - (2.67).

Distribuciones adicionales definidas por y y b: Estas distribuciones, juegan papel fundamental en el ajuste del modelo y en la validación de los supuestos del modelo.

Distribución conjunta: La distribución conjunta $f_{y,b}(\mathbf{y}_i, \mathbf{b}_i)$ de $\mathbf{y}$ y $\mathbf{b}$ para los *Modelos Lineales Mixtos clásicos* puede especificarse como:

$$f_{y,b}(\mathbf{y}_i, \mathbf{b}_i) = f_{\mathbf{y}|\mathbf{b}}(\mathbf{y}_i|\mathbf{b}_i)f_b(\mathbf{b}_i). \tag{2.79}$$

Dado que las distribuciones de componentes, $f_b(\mathbf{b}_i)$ y $f_{\mathbf{y}|\mathbf{b}}(\mathbf{y}_i|\mathbf{b}_i)f_b(\mathbf{b}_i)$, son normales multivariadas, la distribución conjunta también es normal.

Distribución Marginal para y: La distribución marginal $f_y(y_i)$ de $\mathbf{y}_i$ se obtiene a partir de:

$$f_y(\mathbf{y}_i) = \int f_{y,b}(\mathbf{y}_i, \mathbf{b}_i)d\mathbf{b}_i = \int f_{\mathbf{y}|\mathbf{b}}(\mathbf{y}_i|\mathbf{b}_i)f_b(\mathbf{b}_i)d\mathbf{b}. \tag{2.80}$$

donde $f_{y,b}$ es la densidad de la distribución conjunta de $\mathbf{y}_i$ y $\mathbf{b}_i$, $f_{y|b}$ es la distribución condicional de $\mathbf{y}_i$ dada $\mathbf{b}_i$, y f_b es la densidad de la distribución incondicional de $\mathbf{b}_i$. Dado que $f_{y,b}$ y f_b son densidades de distribuciones normales multivariadas, la distribución marginal de $\mathbf{y}$ también es normal multivariada y puede derivarse analíticamente.

Estimación

Para el *Modelo Lineal Mixto clásico*, las ecuaciones (2.75) y (2.76) y (2.78) implican que la media marginal y la matriz de varianza-covarianza de $\mathbf{y}_i$ se dan de la siguiente manera:

$$E(\mathbf{y}_i) = \mathbf{X}_i\beta, \tag{2.81}$$

$$\begin{aligned} Var(\mathbf{y}_i) &\equiv \mathbf{V}_i(\sigma^2, \theta; \mathbf{v}_i) \\ &= \sigma^2\mathbf{V}_i(\theta; \mathbf{v}_i) = \sigma^2[\mathbf{Z}_i\mathbf{D}(\theta_D)\mathbf{Z}_i^{'} + \mathbf{R}_i(\theta_R; \mathbf{v}_i)] \end{aligned} \tag{2.82}$$

donde $\theta^{'} \equiv (\theta_D^{'}, \theta_R^{'})^{'}$[9]

De (2.81) y (2.82), se deduce que, marginalmente,

$$\mathbf{y}_i \sim \mathcal{N}(\mathbf{X}_i\beta, \sigma^2\mathbf{Z}_i\mathbf{D}\mathbf{Z}_i^{'} + \sigma^2\mathbf{R}_i) \tag{2.83}$$

El valor medio marginal de la variable vector dependiente $\mathbf{y}_i$, se define mediante una combinación lineal de los vectores de covariables incluidos en la matriz de diseño del grupo $\mathbf{X}_i$, con parámetros β. La matriz de varianza-covarianza de $\mathbf{y}_i$ presenta dos componentes. El primero, $\sigma^2\mathbf{Z}_i\mathbf{D}\mathbf{Z}_i^{'}$ es contribuido por los efectos aleatorios $\mathbf{b}_i$, el segundo, $\sigma^2\mathbf{R}_i$, está relacionado con los errores residuales ε_i. Por lo tanto, el modelo de efectos aleatorios especificado en (2.64) - (2.67), implica una distribución normal marginal, definida por (2.83). La matriz de varianza-covarianza de $\mathbf{y}_i$ tiene forma paramétrica muy específica, dada por (2.82).

Se observa que el modelo marginal, definido por (2.81) y (2.83)), no involucra los efectos aleatorios $\mathbf{b}_i$. Por lo tanto, la matriz $\mathcal{D}$ no tiene que ser tratada como una

[9] A modo de simplificar la notación, se suprime el uso de θ y $\mathbf{v}_i$ en las fórmulas

matriz de varianza-covarianza, por lo cual, no tiene que ser definida-positiva siempre y cuando la matriz $\mathcal{V}_i$ sea definida-positiva (definida en (2.56) como la mariz marginal de varianzas y covarianzas). Sin embargo, la matriz $\mathcal{D}$ necesita ser simétrica para garantizar que la matriz $\mathcal{V}_i$ sea simétrica. Se deduce que, aunque cada Modelo Lineal Mixto de la forma especificada en (2.64) - (2.67) implica un modelo marginal dado por (2.83), no todos los modelos de la forma (2.83) pueden interpretarse como resultado de un Modelo Lineal Mixto. Por lo tanto, los dos modelos no son equivalentes.

Se desprende que los modelos lineales con efectos fijos y errores residuales correlacionados, presentados anteriormente, son menos restrictivos que los modelos lineales con efectos mixtos. En general, los primeros no permiten hacer inferencia sobre la variabilidad que puede estar relacionada con los diferentes niveles de la jerarquía de datos.

Se destaca que los efectos de las covariables, incluidos en la matriz de diseño $\mathbf{X}_i$, se cuantifican con los mismos parámetros β en la media condicional (2.75) e incondicional (2.81). Por lo tanto, los parámetros se pueden interpretar como efectos cuantificadores a nivel de la población. Esta posibilidad de interpretación dual de los efectos fijos β es una característica única del *Modelo Lineal Mixto clásico*, dada por (2.64) - (2.67). El hecho de que el *Modelo Lineal Mixto clásico* implica el modelo marginal (2.57) es importante desde un punto de vista práctico, lo que permite la construcción de enfoques de estimación efectivos para el Modelo Lineal Mixto. Este tema se trata en la siguiente sección.

Estimación de máxima verosimilitud:

En general, la estimación de *Modelos Lineales* implica la construcción de la función de verosimilitud basada en la función de distribución de probabilidad apropiada para los datos observados. La distribución incondicional de $\mathbf{b}_i$ y la distribución condicional de $\mathbf{y}_i$ dada $\mathbf{b}_i$, que se definió anteriormente para el *Modelo Lineal Mixto clásico*, no son adecuados para construir la función de verosimilitud, porque los efectos aleatorios $\mathbf{b}_i$ no son observables. Por razones similares, la distribución conjunta de $\mathbf{y}_i$ y

$\mathbf{b}_i$ no puede utilizarse.

En cambio, la estimación de Modelo Lineal Mixto se basa en la distribución marginal de $\mathbf{y}_i$.

De hecho, coincide con la distribución dada en (2.83). Por esta razón, la estimación de los parámetros del *Modelo Lineal Mixto clásico* se puede lograr utilizando la estimación MV o MVR. En particular, la estimación de *Modelos Lineales* se basa en la logaritmo de verosimilitud marginal resultante de (2.83), que se puede expresar de la siguiente manera:

$$\begin{aligned} l_{Full}(\beta, \sigma^2, \theta) \equiv \quad & -\frac{N}{2}\log(\sigma^2) - \frac{1}{2}\sum_{i=1}^{N}\log[det(\mathbf{V}_i)] \\ & -\frac{1}{2\sigma^2}\sum_{i=1}^{N}(\mathbf{y}_i - \mathbf{X}_i\beta)'\mathbf{V}_i^{-1}(\mathbf{y}_i - \mathbf{X}_i\beta) \end{aligned} \tag{2.84}$$

donde $\mathbf{V}_i$, definido en (2.82), depende de θ. Las estimaciones de β, σ^2 y θ generalmente se obtienen usando una logaritmo de verosimilitud de perfil para θ, resulta de sustituir en (2.84) los estimadores de β y σ^2, dados por

$$\widehat{\beta}(\theta) \equiv \left(\sum_{i=1}^{N}\mathbf{X}_i'\mathbf{V}_i^{-1}\mathbf{X}_i\right)^{-1}\sum_{i=1}^{N}\mathbf{X}_i'\mathbf{V}_i^{-1}\mathbf{y}_i \tag{2.85}$$

$$\widehat{\sigma}_{MV}(\theta) \equiv \sum_{i=1}^{N}\mathbf{r}_i'\mathbf{V}_i^{-1}\mathbf{r}_i/n \tag{2.86}$$

donde $\mathbf{r}_i \equiv \mathbf{r}_i(\theta) = \mathbf{y}_i - \mathbf{X}_i\widehat{\beta}(\theta)$. Se debe tener en cuenta que las expresiones corresponden a (2.57) y (2.58). Al maximizar la función log-verosimilitud de perfil sobre θ, se obtienen estimadores de estos parámetros. Al sustituir $\widehat{\theta}$ en (2.85) y (2.86), se obtienen los estimadores correspondientes de β y σ^2, respectivamente.

Como ya se ha mencionado, las estimaciones de MV de los parámetros de varianza-covarianza están sesgadas, por lo que se considera más adecuado estimar los parámetros usando la estimación MVR. Con este fin, se considera la función de verosimilitud de perfil restringida, correspondiente a (2.61). A partir de esta función, el parámetro σ^2 resulta el correspondiente a (2.62):

$$\widehat{\sigma}_{MVR}(\theta) \equiv \sum_{i=1}^{N}\mathbf{r}_i'\mathbf{V}_i^{-1}\mathbf{r}_i/(n-p) \tag{2.87}$$

con $\mathbf{r}_i$ definido como en (2.86). Esto lleva a una función de verosimilitud de perfil restringida de registro, que solo depende de θ:

$$\begin{aligned} l_{MVR}(\theta) \equiv \quad & -\frac{n-p}{2}\log\left(\sum_{i=1}^{N}\mathbf{r}_i^{'}\mathbf{r}_i\right) - \frac{1}{2}\sum_{i=1}^{N}\log[det(\mathbf{V}_i)] \\ & -\frac{1}{2}\log\left[det\left(\sum_{i=1}^{N}\mathbf{X}_i^{'}\mathbf{V}_i^{-1}\mathbf{X}_i\right)\right] \end{aligned} \tag{2.88}$$

La maximización de (2.88) concluye un estimador de θ, que luego se suplanta en (2.85) y (2.87) para calcular estimadores de β y σ^2, respectivamente.

Capítulo 3

Trabajo de campo y caracterización de la muestra

En el presente capítulo se describe en primer lugar, el diseño muestral aplicado desarrollado en el apéndice B. Luego se detallan tanto el procedimiento del trabajo de campo y el respectivo relevamiento de datos, como los métodos y herramientas utilizadas (variables e indicadores relevados o calculados). Desde un enfoque descriptivo, se realiza una caracterización de la muestra. Por último, se evalúa el primer objetivo planteado, sobre el comportamiento del PFE de los niños en estudio con respecto a niños sin patología a nivel nacional.

Es importante destacar que en esta investigación no se cuenta con las herramientas necesarias para una correcta calibración de la muestra para poder realizar inferencia sobre la población, ya que no se obtuvieron los expansores correspondientes a las etapas dos a cuatro. Este inconveniente se debe a que la muestra fue sorteada en un marco que fue modificándose (como ser la mudanza del individuo de domicilio así como escuela) a lo largo del monitoreo.

3.1. Diseño Muestral

Se construye el **marco muestral** a partir de información proporcionada por la Administración Nacional de Educación Pública (ANEP). Esto es, a partir de los listados de preescolares y escolares, ordenados por género y curso, matriculados en todas las escuelas de la ciudad de Artigas en el año 2015.

La **población** total considerada consta de 2389 niños entre 4 y 7 años inclusive, matriculados en las veinte escuelas urbanas[1] de la ciudad de Artigas, en cursos comprendidos entre Nivel 4 y Segundo año de primaria.

El **tamaño de muestra**, se estableció de acuerdo a los costos y recursos disponibles para el monitoreo, el cual fue de 30 % de la población. Esta fracción permite estimar con distintos niveles de precisión diferentes parámetros como ser la media para variables continuas y proporciones para variables categóricas.

El tamaño de muestra se estableció en 714 niños, conformados por la totalidad de las 20 escuelas urbanas de la ciudad, con el censo de una escuela de 30 individuos (contaba con ese tamaño de población) y el resto de las escuelas (19 en total), 36 niños.

Como ya ha sido mencionado y descrito en el apéndice B, se realiza un **muestreo sistemático ordenado** por clase (nivel educativo: 4, 5, 1er y 2do año), sexo y alfabéticamente para cada una de las escuelas. Uno de los factores que incidió en la elección del diseño, fue el hecho de tener tamaño de muestra fijo por escuela (condición establecida por el Ministerio de Salud pública).

Los niños muestreados fueron objeto de la investigación en cuatro etapas de monitoreo.

[1] 18 de ellas de estrato público y 2 de estrato privado

3.2. Procedimiento de monitoreo

El monitoreo tuvo como punto de partida la entrega de consentimientos informados[2] a los padres de los niños sorteados, junto con un cuestionario realizado[3] por el Ministerio de Salud Pública (MSP). La aceptación de dicho consentimiento, implicaba la participación del niño en el estudio ademas de su propia voluntad a participar, ademas de casos que se excluyeron por falta de comprensión del procedimiento.

La recopilación de datos en campo en cada monitoreo, se realizó mediante 2 grupos de trabajo, integrados por funcionarios del MSP (doctores, licenciados en enfermería y personal administrativo), participantes residentes en la ciudad de Artigas (estudiantes de auxiliar de enfermería) y estudiantes de la licenciatura en estadística. A uno de los grupos se le asignó 11 escuelas pertenecientes a barrios próximos entre sí, mientras que el otro grupo cubría 9 escuelas, la gran mayoría ubicadas en zonas alejadas.

En cada monitoreo de relevamiento, se visitaba cada escuela en más de una oportunidad en el correr de 3 o 4 días, con el fin de localizar la mayor cantidad de niños y niñas. Cada monitoreo, constaba del relevamiento de tres variables: peso, talla y capacidad pulmonar además de los tomados del cuestionario en la primer etapa. Los aparatos utilizados como las balanzas digitales, tallímetros convencionales y flujómetros, fueron los mismos para todas las etapas. Tanto los evaluadores como los niños participantes, fueron instruidos en el procedimiento de toma de datos. Para una mejor toma y registro de los datos, y así reducir los errores de medidas entre los grupos de trabajo, se especificó un procedimiento a realizar por todos los participantes que tomaban los datos: para la medición del peso y la talla se retiraba calzado, abrigos y broches de pelo y se respetaba la postura correcta. Para la medición de la capacidad pulmonar, se solicitaba a cada participante que soplara a través de piezas bucales cilíndricas ("boquillas") descartables, colocadas en los flujómetros. Este

[2]Ver consentimiento en Apéndice A.

[3]Ver cuestionario en Apéndice A.

procedimiento se repetía en tres oportunidades, tomando un descanso prudencial entre cada una de ellas, y se tomaba registro del mayor valor alcanzado.

3.3. Relevamiento de datos

La digitalización de los datos, se realizó mediante el programa EpiData (10), programa de software libre de gran utilidad para el ingreso de datos, se centra en la entrada y documentación de datos provenientes de cuestionarios o formularios de registro y tiene la ventaja de que permite exportar los datos digitalizados en formatos conocidos, lo que facilita la lectura en programas de análisis estadísticos.
Una de las finalidades del uso de esta herramienta es facilitar el trabajo de los digitadores y minimizar el ingreso de errores al tipear los datos. Esto se logra, con la programación de restricciones, por ejemplo opciones de lista en las respuesta de resultados (ej.: 1=No, 2=Sí), listas de texto como "etiquetas indizadas" o rangos de valores en campos numéricos.
Una vez que el formulario de entrada de datos está listo, es fácil definir qué datos se pueden ingresar en los diferentes campos. Asimismo, se pueden agregar variables adicionales a las monitoreadas en un formulario, como la identificación del digitador, importante al momento de la validación de los datos. El análisis de los datos fue realizado con el programa **R** (17) a través de la interfaz de usuario **RStudio**(18), utilizando como manual de referencia el **R for Data Science**(Grolemund y Wickham) y las librerías `nlmeU`(5), `nlme`(16), `dplyr`(29), `tidyverse`(25), `tibble`(14), `purrr`(8), `ggplot2`(24), `readr` (31), `stringr`(27), `forcats`(26), `tidyr`(Wickham y Henry), `dplyr`(29), `MASS`(22), `readxl`(28), `reshape`(23) y `lattice`(19).

3.4. Variables relevadas

Se clasificaron las variables de análisis en cuatro grupos, tal como se listan a continuación.

Primer grupo: 19 variables relacionadas a la información personal del niño/niña (nombre, cédula, sexo, etc.), que se recopilan de los listados de ANEP y del primer bloque de preguntas del cuestionario[4] realizado en la primer etapa de monitoreo.
Segundo grupo: 31 variables que surgen de los restantes bloques del cuestionario, presentan información acerca de los antecedentes de salud del niño/niña (diagnóstico médico, uso de medicamentos, etc.), antecedentes ambientales (fumadores dentro del hogar, métodos para calefaccionarse y cocinar, etc.) y otras variables de interés, las cuales también fueron tomadas únicamente en la primer etapa de monitioreo.
Tercer grupo: 21 variables resultantes de cada fase de monitoreo (peso, talla, PFE, digitador, etc.).
Cuarto grupo: variables adicionales creadas a partir de las anteriores, fundamentales para contextualizar la información:

- Edad: Se crea la variable *edad*, a partir de la fecha de nacimiento que resulta de los listados de ANEP. Para el caso de los participantes con fecha de nacimiento faltante, que no pudo corroborarse en campo, se imputaron fechas de nacimiento auxiliares de acuerdo al curso en que estaba matriculado (por ejemplo los niños matriculados en primer año se les asignó fechas de nacimiento correspondiente a 6 años de edad en el momento de la primera etapa).

- Manzana censal: Variable creada a partir de los domicilios particulares de cada niño/niña y escuela, para lo cual se tomó como guía, información del Instituto Nacional de Estadística (9).

[4]Ver cuestionario en Apéndice A.

- Zonas geográficas: Se definieron 5 zonas de acuerdo a la ubicación geográfica de la ciudad de Artigas (ver Figura 3.1). La zona *Cerros*, incluye los barrios ubicados en los cerros San Eugenio, Ejido, Signorelli y Pintadito. La zona *Industrial*, se definió como el área comprendida hasta 1 kilómetro de distancia del molino arrocero. Las zonas *Centro* y *Microcentro*, se determinaron a partir de la avenida principal de la ciudad (Av. Cnel. Carlos Lecueder) y sus alrededores. Por último, la zona *Periferia* se definió considerando la proximidad con el río Quaraí, que rodea parte de la ciudad.

  Una vez definidas las zonas, quedaron determinadas las manzanas censales incluidas en cada una de ellas. A partir de la manzana previamente asignada a cada niño/niña y escuela, se construyeron las variables *"zona de residencia"* y *"zona de la escuela"* respectivamente.

3.4.1. Indicadores

Del cuestionario entregado a los padres, donde se consulta acerca de la historia clínica del menor y de las condiciones ambientales del hogar, se priorizaron las preguntas con contenido objetivo y relevante para el presente estudio. Con ellas, se construyeron los siguientes indicadores:

- Indicador de patología: Este indicador se creó a partir de las variables *diagnóstico médico* y *uso de medicamentos*. En primer lugar, se depuró la variable *uso de medicamentos*, ya que solo eran de relevancia para este estudio los medicamentos asociados a afecciones respiratorias o alergias[5]. Para el caso del *diagnóstico médico*, no fue necesario la depuración, porque la pregunta ya estaba orientada a diagnóstico de enfermedades respiratorias. Luego, a partir de las dos variables validadas, se creó el *indicador de patología*, que toma los siguientes valores:

[5]La clasificación de los medicamentos fue validada por las doctoras del MSP.

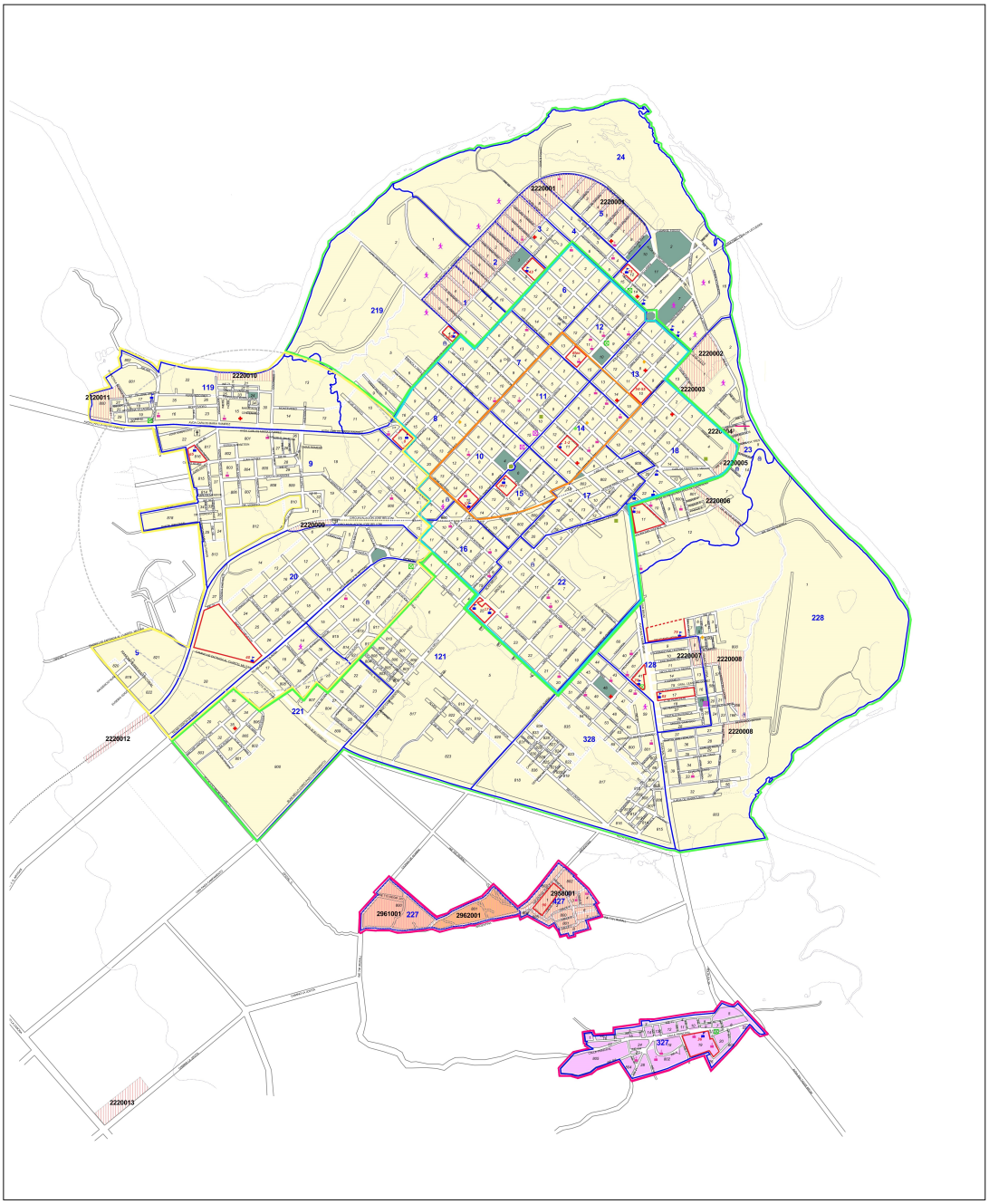


Figura 3.1: Mapa de las zonas determinadas en la ciudad de Artigas.

$$Indicador\ de\ Patología = \begin{cases} 1 & Si\ alguna\ de\ las\ variables\ es\ 1 \\ 0 & Si\ todas\ las\ variables\ son\ 0 \\ NA & Si\ todas\ las\ variables\ son\ NA \end{cases}$$

- Indicador de exposición: Este indicador se creó a partir de dos variables recabadas del cuestionario, *Fuma alguien en la casa* y *leña*. El indicador de exposición es afirmativo (igual a 1) en los casos que el niño se declara estar expuesto a contaminación intradomiciliaria:

$$Indicador\ de\ Exposición = \begin{cases} 1 & \text{Si alguna de las variables es 1} \\ 0 & \text{Si todas las variables son 0} \\ NA & \text{Si todas las variables son NA} \end{cases}$$

3.5. Características de la muestra

En esta sección, se presentan las características de los datos de la muestra relevada en cada una de las etapas.

La tabla a continuación (Tabla 3.1), presenta la cantidad de respuestas obtenidas en cada etapa, además de las fechas y características de la misma:

	Fecha	Período	Niños monitoreados
Etapa 1	Mayo 2015	Post-zafra	491 (68,8 %)
Etapa 2	Setiembre 2015	Fuera de zafra	577 (80,8 %)
Etapa 3	Diciembre 2015	Fuera de zafra	560 (78,4 %)
Etapa 4	Abril 2016	Zafra	596 (83,5 %)

Tabla 3.1: Cantidad de respuestas por etapa.

En la Tabla 3.1, se puede observar que en la primera etapa fue en la que se relevaron menor cantidad de datos, el principal factor pudo estar asociado al desconocimien-

to del estudio. Le sigue la tercera etapa, donde las razones estuvieron asociadas principalmente a condiciones climáticas (hubo alerta roja por fuertes vientos y precipitaciones) y adicionalmente, debido a que la fecha elegida era próxima al cierre del año escolar. Ambos factores influyeron en la inasistencia de gran parte de los escolares al centro de estudio.

Además del enfoque desde las distintas fases de monitoreo, se pueden analizar las respuestas desde los casos individuales. Esto es, el estudio de los patrones perdidos que se presenta en la Tabla 3.2:

+ + + +	+ + + -	+ + - +	+ - + +	- + + +	+ + - -	+ - - +	+ - + -
373	20	47	21	100	7	7	5
- + + -	- + - +	- - + +	+ - - -	- + - -	- - + -	- - - +	- - - -
9	15	31	11	6	1	2	59

Tabla 3.2: Perfil de patrones perdidos.

De la Tabla 3.2, donde "+" representa respuesta y "-" representa patrón faltante, se observa que más del 52 % (373) de los niños y niñas sorteados (714), estuvieron presentes en todas las instancias del monitoreo, mientras que el 8 % (59 participantes) nunca fue monitoreado. Un punto a destacar, es que hubo 100 casos de no respuesta en la primera estapa que lograron "captarse" en la segunda etapa. Otro punto importante, es el caso de los 47 niños ausentes en la tercera etapa, debido a las razones antes mencionadas.

Con las mediciones obtenidas de capacidad pulmonar, esto es los valores de PFE, se realizan diagramas de caja para analizar, entre otras cosas, la dispersión de los datos. En la Figura 3.2 se puede apreciar que para el caso de las niñas (gráfico de la izquierda), las cuatro mediciones de PFE presentan mayor concentración antes del segundo cuartil. Además, se observa que la primera etapa es la que presenta

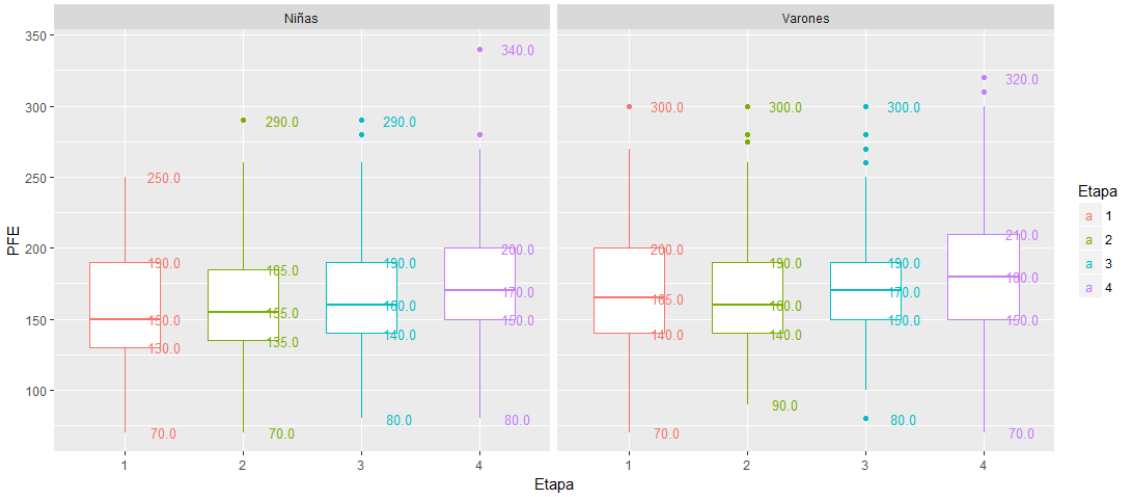


Figura 3.2: Diagrama de caja de PFE por sexo según etapa.

mayor dispersión en el rango intercuartílico y desde la segunda etapa en adelante el comportamiento del primer, segundo y tercer cuartil muestran un crecimiento moderado constante. Por otro lado, los valores mínimos se presentan estables entre 70 y 80 l/min y los valores máximos tienen grandes incrementos, reflejando la mejor medición en la cuarta etapa (340 l/min). Para el caso de los varones (gráfico de la derecha), los diagramas reflejan mayor simetría de los datos en comparación con las niñas. Se observa también, una clara desmejora en la segunda etapa para el tercer y cuarto cuartil. Las mejores mediciones se presentaron constantes en las primeras tres etapas y se destaca un incremento en la última toma de datos.

Otra herramienta para analizar el comportamiento del PFE a lo largo del monitoreo, son los “gráficos spaghetti”, diseñados para visualizar flujos de datos en un sistema.

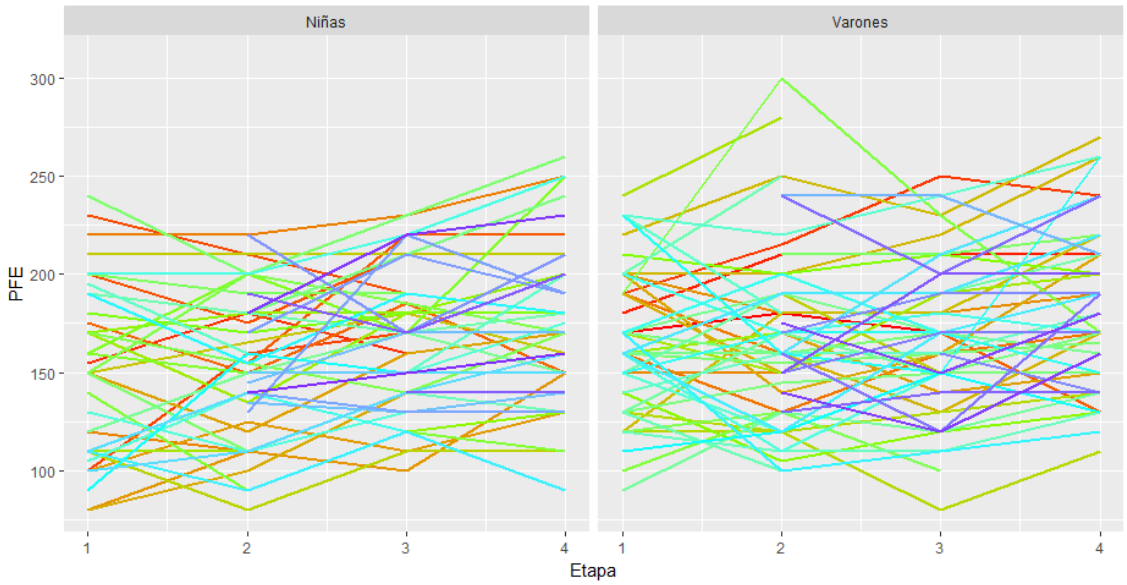


Figura 3.3: Trayectoria individual de PFE por sexo según etapa.

En la Figura 3.3, se presentan las trayectorias de los valores de PFE para un subgrupo de niñas y varones, y en la Figura 3.4, se desagregan estas trayectorias por las edades correspondientes al inicio del estudio.

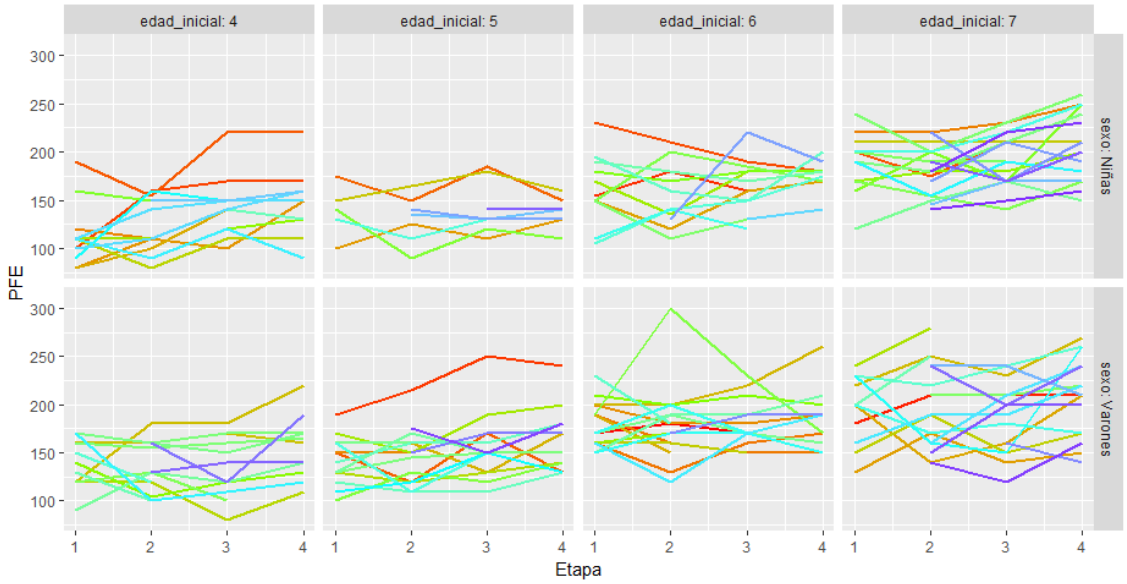


Figura 3.4: Trayectoria individual de PFE por sexo según etapa y edad inicial.

Sobre la base de las trayectorias que se muestran en las Figuras 3.3 y 3.4, se pueden hacer las siguientes observaciones:

- En general, las trayectorias de PFE tienden a aumentar con el tiempo. Esto es de acuerdo a lo esperado, ya que a medida que un participante (sin importar el género) se desarrolla y crece en edad y talla debería también crecer su capacidad pulmonar.

- En particular, hay algunos casos que presentan aumento lineal de la capacidad pulmonar con el tiempo, pero también hay otros casos, para los que se observan crecimientos bruscos que se desvían de una tendencia lineal.

- Las mediciones de PFE desagregadas por edades al inicio del monitoreo, reflejan la tendencia de que a mayor edad el PFE es en promedio mayor. Si se analizan en particular cada uno de los subgráficos se observa que a lo largo de las etapas, cada trayectoria tiende a crecer.

- La edad que presenta mayor dispersión en el subconjunto de trayectorias graficadas, corresponde a niños y niñas de 7 años al inicio del estudio.

Por otra parte, interesa analizar el comportamiento de la variable PFE en las 4 etapas de monitoreo, para ello, se realizan los gráficos de correlación de los valores de PFE entre etapas presentados en la Figura 3.5, donde se puede apreciar que dicha correlación intra-etapa crece con cada intervalo de tiempo.

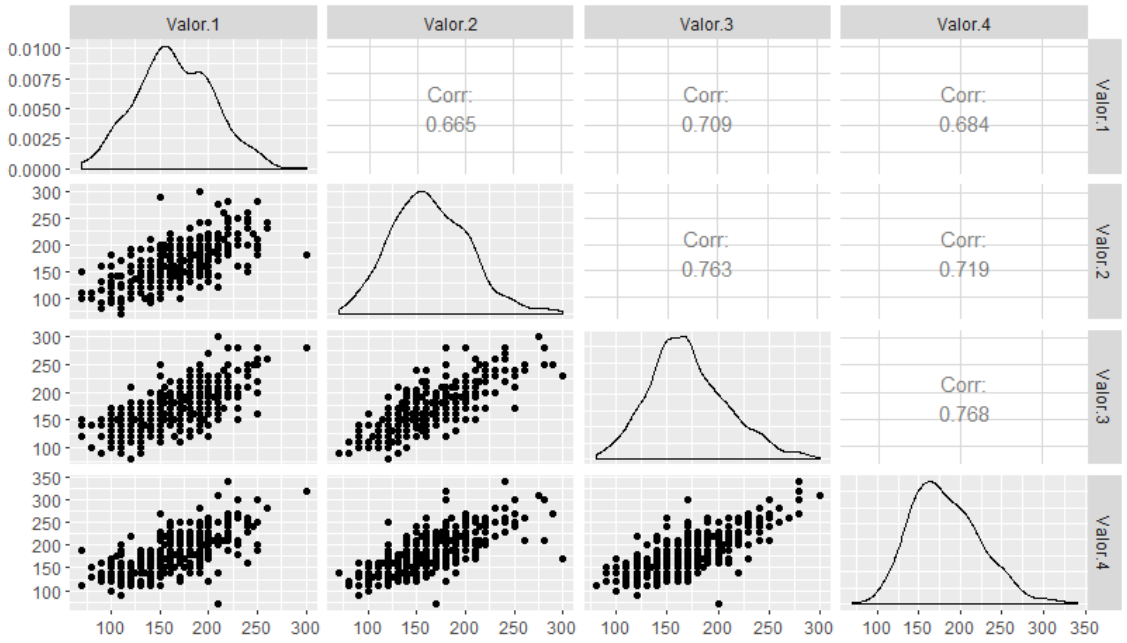


Figura 3.5: Correlación de PFE por etapas.

3.6. Comparación de datos monitoreados con datos a nivel nacional

Para abordar el primer objetivo planteado, se considera como punto de referencia de una población sin patologías y en condiciones ambientales óptimas, el estudio realizado por la Facultad de Medicina del Centro Hospitalario Pereira Rossell(2). En dicho estudio, se determinó el valor percentilar 10, 50 y 90 según sexo, peso, talla y edad y se correlacionó mediante una ecuación de mínimos cuadrados por el polinomio de segundo grado $y = a + bx + cx^2$. Para este estudio se toma el ajuste por talla, la cual presenta los siguientes coeficientes de ajuste:

	Niñas			Varones		
	a	b	c	a	b	c
Percentil 10	-388,2	6,3	-0,01	-123,7	1,4	0,01
Percentil 50	-649,6	10,8	-0,03	445,0	-8,0	0,05
Percentil 90	-1023,0	17,2	-0,05	324,0	-6,2	0.05

Tabla 3.3: Coeficientes para ajuste del polinomio $y = a + bx + cx^2$, siendo y la capacidad pulmonar y x la talla.

Se calcularon los valores percentilares del PFE en función de las tallas monitoreadas en cada una de las etapas del presente estudio, a partir del polinomio de segundo grado $y = a + bx + cx^2$ con los coeficientes presentados en la Tabla 3.3.
En los gráficos de los valores de PFE observados con los valores percentilares según la talla, se espera que el 10 % de los datos se encuentre por debajo de la curva del percentil 10, 40 % entre las curvas del percentil 10 y la del percentil 50, 40 % en el tramo siguiente (entre el percentil 50 y el percentil 90) y 10 % por encima del percentil 90.

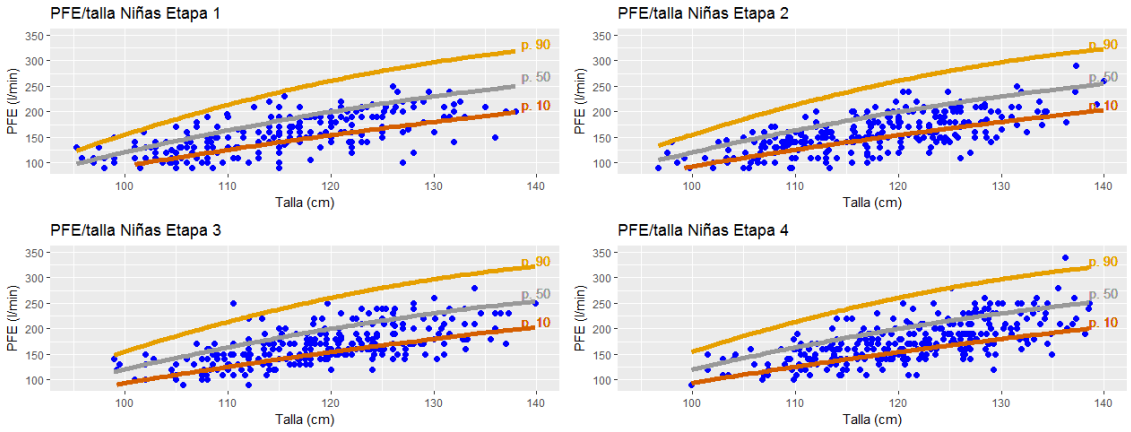


Figura 3.6: PFE observado vs. PFE/talla para niñas según etapa.

Teniendo en cuenta que PFE/talla, hace referencia a los valores poblacionales de PFE ajustados por talla, la Figura 3.6 muestra los valores observados del PFE para

niñas monitoreadas en cada una de las etapas con sus respectivos valores percentilares calculados a partir de la talla como se mencionó anteriormente. Se puede apreciar que los valores observados, no tienen el comportamiento esperado. La mayoría de los valores se encuentran concentrados por debajo de percentil 50 y prácticamente no hay observaciones por encima del percentil 90. Esto puede comprobarse con la Tabla 3.4[6] en donde se muestran la frecuencia de los datos observados en cada una de las etapas, según zona de residencia y grupo percentilar. Tal como se veía en la Figura 3.6, para el caso del Percentil 10, se esperaba contar con el 10 % de las observaciones, sin embargo figuran el 26 %, 32 %, 36 % y 28 % respectivamente para cada etapa. Lo mismo para el caso del percentil 50, dónde se concentran más del 80 % de las observaciones a lo largo de todo el monitoreo y se verifica también que por encima del percentil 90, se registran pocas observaciones lejos del 10 % esperado.

	Etapa 1			Etapa 2			Etapa 3			Etapa 4		
	n	N	%	n	N	%	n	N	%	n	N	%
Cerros	13	28	46 %	13	31	42 %	11	30	37 %	11	35	31 %
Industrial	10	51	20 %	17	55	31 %	19	55	35 %	14	57	25 %
Periferia	29	105	28 %	35	119	29 %	41	122	34 %	29	123	24 %
Centro	6	37	16 %	15	52	29 %	19	44	43 %	18	48	38 %
Microcentro	4	14	29 %	7	15	47 %	7	17	41 %	7	18	39 %
Percentil 10	62	235	26 %	87	272	32 %	97	268	36 %	79	281	28 %

Tabla 3.4: Cantidad y porcentaje de niñas por debajo del percentil 10 por etapa según zona de residencia.

La Tabla 3.4 muestra el total de niñas por debajo del percentil 10, el total de niñas monitoreadas y la correspondiente proporción, para cada una de las etapas según zona de residencia. Por ejemplo, para el caso de los Cerros, en la primera etapa se

[6]Ver tabla completa en anexo Tabla C.1.

monitorearon 28 niñas, de las cuales el 46 % (13 niñas), tuvieron mediciones de PFE por debajo del percentil 10.

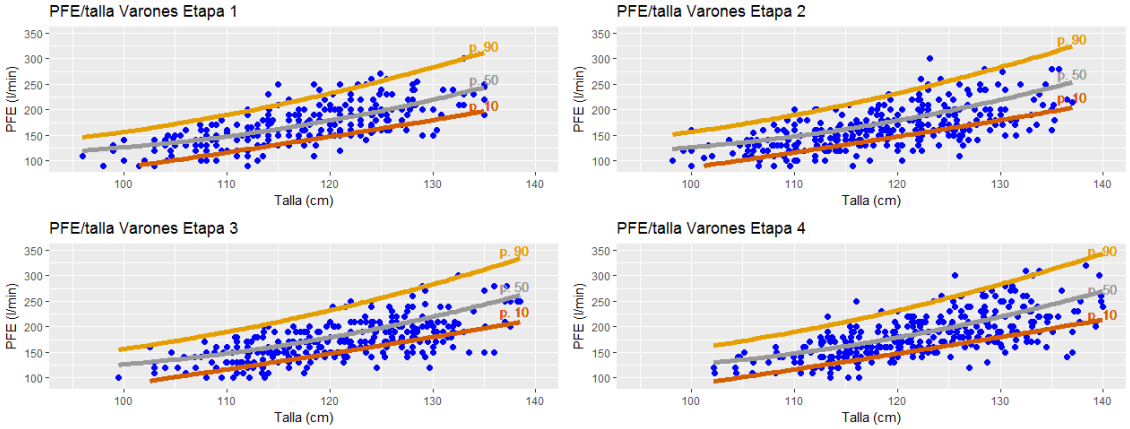


Figura 3.7: PFE observado vs. PFE/talla para varones según etapa.

Análogamente para los varones monitoreados, en la Figura 3.7 se presentan los valores observados de PFE contra los percentiles 10, 50 y 90 según talla para cada etapa. También, para facilitar el análisis se contruye la Tabla 3.5, que contiene las frecuencias de datos observados para cada etapa según zona de residencia y grupo percentilar. En la Tabla 3.5, se observa que los porcentajes acumulados hasta el

	Etapa 1			Etapa 2			Etapa 3			Etapa 4		
	n	N	%	n	N	%	n	N	%	n	N	%
Cerros	7	29	24 %	8	37	22 %	5	37	14 %	7	41	17 %
Industrial	2	52	4 %	13	65	20 %	7	55	13 %	13	66	20 %
Periferia	19	102	19 %	36	121	30 %	30	114	26 %	26	116	22 %
Centro	8	58	14 %	17	66	26 %	18	64	28 %	10	68	15 %
Microcentro	2	15	13 %	2	16	13 %	3	22	14 %	4	24	17 %
Percentil 10	38	256	15 %	76	305	25 %	63	292	22 %	60	315	19 %

Tabla 3.5: Cantidad y porcentaje de varones por debajo del percentil 10 por etapa según zona de residencia.

percentil 10 son de 15 %, 25 %, 22 % y 19 % respectivamente para cada etapa. Están

por encima de los valores esperados, pero tienen mejor desempeño con respecto a las niñas.

Análogamente, la Tabla 3.5 muestra el total de varones por debajo del percentil 10, el total de varones monitoreados y la correspondiente proporción, para cada una de las etapas según zona de residencia. Por ejemplo, para el caso de los Cerros, en la primera etapa se monitorearon 29 varones, de los cuales el 24 % (7 varones), tuvieron mediciones de PFE por debajo del percentil 10[7].

En la Tabla 3.6, se desagrega la población relevada en las distintas etapas del monitoreo, según zona de residencia, sexo y edad[8], también se muestra para cada apertura de las variables, el porcentaje de respuesta sobre la población total.

[7]Ver tabla completa en apéndice Tabla C.2.

[8]Se refiere a la edad al inicio del monitoreo.

	Variable	Etapa 1		Etapa 2		Etapa 3		Etapa 4		Total
		n	%	n	%	n	%	n	%	
zonas	Cerros	57	66	68	79	67	77	76	88	86
	Industrial	103	71	120	82	110	75	123	84	145
	Periferia	207	72	240	83	236	82	239	83	286
	Centro	95	65	118	80	108	73	116	79	146
	Microcentro	29	56	31	60	39	76	42	82	51
sexo	Niñas	235	68	272	78	268	77	281	81	345
	Varones	256	69	305	82	292	79	315	85	369
edad inicial	4	112	73	122	79	111	72	122	79	153
	5	103	63	116	71	119	73	132	80	163
	6	147	72	168	83	168	83	172	85	202
	7	129	65	171	87	162	82	170	86	196
	Total	491	68	577	80	560	78	596	83	714

Tabla 3.6: Cantidad de respuestas por etapa según zona de residencia, sexo y edad inicial.

En la Tabla 3.6, se puede apreciar que las niñas y varones presentes en cada una de las 4 etapas de monitoreo, representan el 68 %, 80 %, 78 % y 83 % respectivamente, de los niños sorteados al inicio del monitoreo. Se observa además, que el Microcentro, es la zona peor representada en todas las etapas, mientras que la Periferia es en promedio la mejor representada de todo el monitoreo. Desde el punto de vista del sexo, los varones monitoreados representan mejor a los varones de la población en comparación al comportamiento de las niñas. En cuanto a las edades, de la muestra observada, el grupo de niños de 6 años al comienzo del monitoreo son los mejor representados de su clase en comparación al resto de los grupos etarios. Desde una perspectiva general, todas las categorías de las variables analizadas en la Tabla 3.6, salvo la zona Microcentro, representan más del 60 % de la población.

Zona	Etapa 1			Etapa 2			Etapa 3			Etapa 4			Total
residencia	p10	n	%	p10	n	%	p10	n	%	p10	n	%	
Cerros	19	57	33	20	68	29	14	67	20	17	76	22	86
Industrial	12	103	11	32	120	26	24	110	21	26	123	21	145
Periferia	52	207	25	73	240	30	66	236	27	54	239	22	286
Centro	14	95	14	37	118	31	41	108	37	28	116	24	146
Microcentro	6	29	20	6	31	19	10	39	25	10	42	23	51
Total	103	491	20	168	577	29	155	560	27	135	596	22	714

Tabla 3.7: Cantidad de respuestas por debajo del percentil 10 y totales según zona de residencia.

En la Tabla 3.7 se presenta el número de observaciones de PFE por debajo del percentil 10, la cantidad de observaciones totales presentes en todo el monitoreo, y la relación porcentual, por etapa según la zona de residencia. Se observa, que las zonas de residencia con mayor proporción de observaciones por debajo del percentil 10 en promedio en todo el monitoreo, son la Periferia y el Centro, mientras que la zona Industrial es la que tiene menor proporción de observaciones con valores de PFE por debajo del percentil 10. Esto no concuerda con la percepción que se tenía previo al estudio. Si se hace foco en los totales por etapa, se puede apreciar que entre el 20 % y el 29 % de la población analizada, presenta valores de capacidad pulmonar por debajo del percentil 10, cuando esta proporción debería acercarse más al 10 % de la población. Dado que los valores del percentil 10 corresponden a una población sin patologías respiratorias, se puede concluir que la capacidad pulmonar de la población estudiada en la ciudad de Artigas, se encuentra muy por debajo de los valores a nivel nacional. Para analizar a los niños cuya capacidad pulmonar se encuentra por debajo del percentil 10, se estudia el indicador de patología de afecciones respiratorias, calculado a partir de las variables del cuestionario *diagnóstico médico* y *uso de medicamentos*. Los resultados se presentan en la siguiente tabla:

Etapa	Percentil 10	Con patología		Sin patología		Sin dato		Total
		n	%	n	%	n	%	
1	debajo	40	38,8	56	54,4	7	6,8	103
	encima	167	43,0	204	52,6	17	4,38	388
	sin dato	51	22,9	75	33,6	97	43,5	223
2	debajo	78	46,4	78	46,4	12	7,14	168
	encima	150	36,7	227	55,5	32	7,82	409
	sin dato	30	21,9	30	21,9	77	56,2	137
3	debajo	71	45,8	73	47,1	11	7,1	155
	encima	141	34,8	213	52,6	51	12,6	405
	sin dato	46	29,9	49	31,8	59	38,3	154
4	debajo	59	43,7	63	46,7	13	9,63	135
	encima	167	36,2	241	52,3	53	11,5	461
	sin dato	32	27,1	31	26,3	55	46,6	118

Tabla 3.8: Cantidad y porcentaje de niños con y sin patología según etapa y percentil 10.

A partir de la Tabla 3.8 se observa que para los participantes que en cada etapa registraron valores de PFE por debajo del percentil 10 y que respondieron a las preguntas relacionadas a *patología* de afecciones respiratorias, supera el 38,8 % para todas las etapas, toma el mayor valor para la Etapa 2, donde el 46,4 % de los niños con mediciones de capacidad pulmonar por debajo del percentil 10, declara tener diagnóstico médico y/o tomar medicación asociada a afecciones respiratorias.

Etapa	Percentil 10	Con exposición		Sin exposición		Sin dato		Total
		n	%	n	%	n	%	
1	debajo	29	28,2	64	62,1	10	9,71	103
	encima	95	24,5	262	67,5	31	7,99	388
	sin dato	39	17,5	81	36,3	103	46,2	223
2	debajo	37	22,0	110	65,5	21	12,5	168
	encima	105	25,7	259	63,3	45	11,0	409
	sin dato	21	15,3	38	27,7	78	56,9	137
3	debajo	32	20,6	99	63,9	24	15,5	155
	encima	96	23,7	252	62,2	57	14,1	405
	sin dato	35	22,7	56	36,4	63	40,9	154
4	debajo	37	27,4	79	58,5	19	14,1	135
	encima	108	23,4	286	62,0	67	14,5	461
	sin dato	18	15,3	42	35,6	58	49,2	118

Tabla 3.9: Cantidad y porcentaje de niños con y sin exposición según etapa y percentil 10.

Análogamente para el indicador de exposición, en la Tabla 3.9 se observa que para los participantes que en cada etapa registraron valores de PFE por debajo del percentil 10, y que respondieron a las preguntas relacionadas a exposición a humo de tabaco y/o leña, supera el 20 % para todas las etapas, toma el mayor valor para la Etapa 1, donde el 38,2 % de los niños con mediciones de capacidad pulmonar por debajo del percentil 10, declara estar expuesto a humo de tabaco y/o de leña.

Capítulo 4

Aplicación de Modelos Mixtos

En el capítulo anterior, se presenta un análisis descriptivo de los datos obtenidos a lo largo del monitoreo. En este primer análisis, se logran resolver algunos de los objetivos planteados. En el presente capítulo, se decide abordar el tema de investigación mediante la aplicación de un conjunto de modelos. Se parte de modelos lineales clásicos y se van levantando supuestos (homocedasticidad e incorrelación), hasta la aplicación de modelos lineales con efectos mixtos. En todos los casos, se presentan únicamente los modelos que mejor ajustan a los datos.

Se organizan los modelos en distintos escenarios como se listan a continuación:

Escenario I: modelos lineales clásicos, aplicados independientemente en cada una de las cuatro etapas, con el fin de estudiar qué variables explican la capacidad pulmonar en cada momento del tiempo.

Escenario II: modelos lineales clásicos, aplicados desde la segunda a la cuarta etapa, tomando como variable explicativa la medición de la capacidad pulmonar en la primera instancia del monitoreo. El fin de este escenario, es analizar si el valor del PFE relevado en la primera toma, incide en las siguientes, y además analizar qué otras variables lo hacen.

Escenario III: modelo lineal con supuestos clásicos, considera datos provenientes

del estudio longitudinal[1]. A diferencia de los escenarios anteriores en donde se consideraban modelos distintos por etapas, como "fotos" de la realidad en cada momento del tiempo, en este escenario se desea modelizar los valores de la capacidad pulmonar de forma conjunta para las etapas 2, 3 y 4 en función de las variables fijas para todas las etapas como la *escuela*, la *zona de residencia* o el *estrato de la escuela*, y adicionalmente se incorpora el valor del PFE en la etapa inicial como dato basal y los valores de las variables *talla*, *peso*, *edad* y *etapa* de forma conjunta para las etapas 2, 3 y 4.

Escenario IV: modelos lineales con varianza heterogénea. Se considera el modelo que mejor ajusta en el escenario anterior y se levanta el supuesto de homogeneidad de varianzas. Se evalúan distintas funciones de varianza.

Escenario V: modelos lineales con varianza heterogénea y errores correlacionados. Se considera el modelo elegido en el escenario anterior y se levanta el supuesto de incorrelación. Se evalúan distintas funciones de correlación.

Escenario VI: modelos lineales con efectos mixtos. Al modelo que mejor ajusta en el escenario anterior, se le incorporan efectos variables.

A modo de resumen en la Tabla 4.1, se exponen los distintos escenarios y sus características.

[1]Se define un estudio longitudinal, como un estudio que recopila observaciones para un mismo individuo a lo largo de un período de tiempo.

Escenario	Etapa	Nombre	Tipo ML
I	1	ML_I_E1	clásicos
	2	ML_I_E2	
	3	ML_I_E3	
	4	ML_I_E4	
II	2	ML_II_E2	clásicos
	3	ML_II_E3	
	4	ML_II_E4	
III	Todas	ML_III	clásico
IV	Todas	ML_IV	varianza heterogénea
V	Todas	ML_V	varianza heterogénea y correlación
VI	Todas	ML_VI	efectos mixtos

Tabla 4.1: Descripción modelos en cada escenario.

4.1. Escenario I

Como se describió anteriormente, en este escenario se aplican modelos lineales clásicos a cada una de las etapas. Como punto inicial, se considera el modelo completo[2]:

$$\begin{aligned} PFE_{ij} = \beta_0 &+ \beta_1 \text{ sexo}_{ij} + \beta_2 \text{ edad}_{ij} + \beta_3 \text{ peso}_{ij} + \beta_4 \text{ talla}_{ij} \\ &+ \beta_5 \text{ zona residencia}_{ij} + \beta_6 \text{ zona escuela}_{ij} + \beta_7 \text{ patología}_{ij} \\ &+ \beta_8 \text{ exposición}_{ij} + \beta_9 \text{ estrato}_{ij} + \varepsilon_{ij} \end{aligned} \tag{4.1}$$

Se realiza una selección jerárquica descendente (de tipo `stepwise`), usando el criterio de AIC menor AIC (criterio de información de Akaike[3]), los modelos resultantes

[2]Modelo lineal con todas las variables disponibles.

[3]Medida de la calidad relativa de un modelo estadístico, para un conjunto dado de datos.

para cada etapa se presentan en la Tabla 4.2.

ML_I_E1	$PFE_{i1} = -156,7 + 10,5$ sexo_masculino$_{i1} + 8,4$ edad$_{i1} + 2,2$ talla$_{i1}$ $+18,1$ zona_industrial$_{i1} + 4,4$ zona_periferia$_{i1}$ $+12,3$ zona_centro$_{i1} + 14,9$ zona_microcentro$_{i1}$
ML_I_E2	$PFE_{i2} = -115,6 + 7,0$ sexo_masculino$_{i2} + 11,4$ edad$_{i2} + 1,8$ talla$_{i2}$ -5,4 patología$_{i2} - 14,6$ estrato_público$_{i2}$
ML_I_E3	$PFE_{i3} = -155,9 + 4,0$ edad$_{i3} + 2,6$ talla$_{i3} - 10,9$ estrato_público$_{i3}$
ML_I_E4	$PFE_{i4} = -82,3 + 6,5$ sexo_masculino$_{i4} + 10,7$ edad$_{i4} + 1,3$ peso$_{i4}$ $+1,4$ talla$_{i4} - 7,2$ patología$_{i4} - 13,1$ estrato_público$_{i4}$

Tabla 4.2: Escenario I - Modelos estimados en cada etapa.

En la Tabla 4.2, se puede observar que las variables explicativas $peso_{ij}$, $talla_{ij}$ y $edad_{ij}$, hacen referencia al valor que toma dicha variable para el individuo i en la etapa j ($j = 1, 2, 3, 4$), mientras que el resto de las variables no presentan variaciones entre etapas (por ejemplo, *estrato*, *patología*, *exposición*).

Al analizar la significación de las variables en cada modelo, se destaca que la *talla* y la *edad* son variables significativas para explicar el PFE, en todas las etapas del monitoreo, aspecto a tener en cuenta en los siguientes análisis. Por otra parte, se observa que el *sexo* es relevante para explicar el PFE para todos los modelos excepto en la tercera etapa. La *zona de residencia* es importante en la primera etapa. El *estrato* de la escuela ("pública" o "privada"), resulta significativa al 0,1 % de la segunda a la cuarta etapa. En cuanto a los indicadores creados a partir de los datos del cuestionario, el *indicador de patología* resulta significativo al 1 % para explicar la capacidad pulmonar en la segunda y última etapa, y el *indicador de exposición* no es considerado por ningún modelo seleccionado.

En base a los resultados, los modelos especificados se puede observar que el estadísti-

co F, presenta mayores valores que su respectivo valor teórico para los grados de libertad dados en cada etapa. Se comprueba así, que los modelos son significativos y se rechaza la hipótesis de que todos los parámetros sean nulos.

Diagnóstico de los modelos: Escenario I

Para validar la calidad de los modelos con respecto a los supuestos clásicos se enfoca la atención en el análisis de los errores. Se realizan los histogramas de los residuos a efectos de verificar que el comportamiento sea de una distribución Normal.

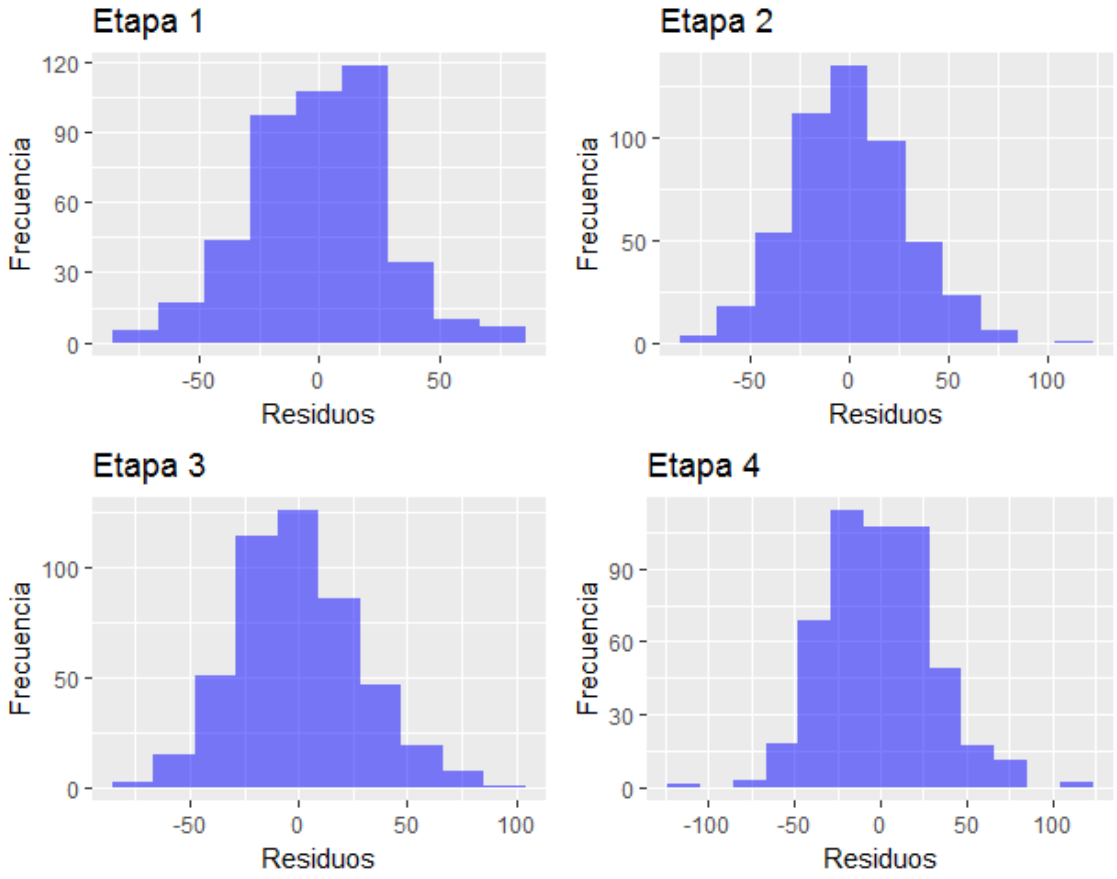


Figura 4.1: Escenario I - Histogramas residuos.

De la Figura 4.1, se observa que los histogramas de los residuos que mejor se ajustan a una distribución normal, son los correspondientes a los modelos seleccionados en las etapas 2 y 3. Éstos, presentan simetría con respecto al cero, que se asemeja a

una campana de Gauss.

Se continúa la validación, mediante los gráficos de los residuos. A modo ilustrativo se muestran los referentes a la primera etapa en la Figura 4.2, para las siguientes etapas pueden verse las Figuras D.1, D.2 y D.3 en el apéndice D.

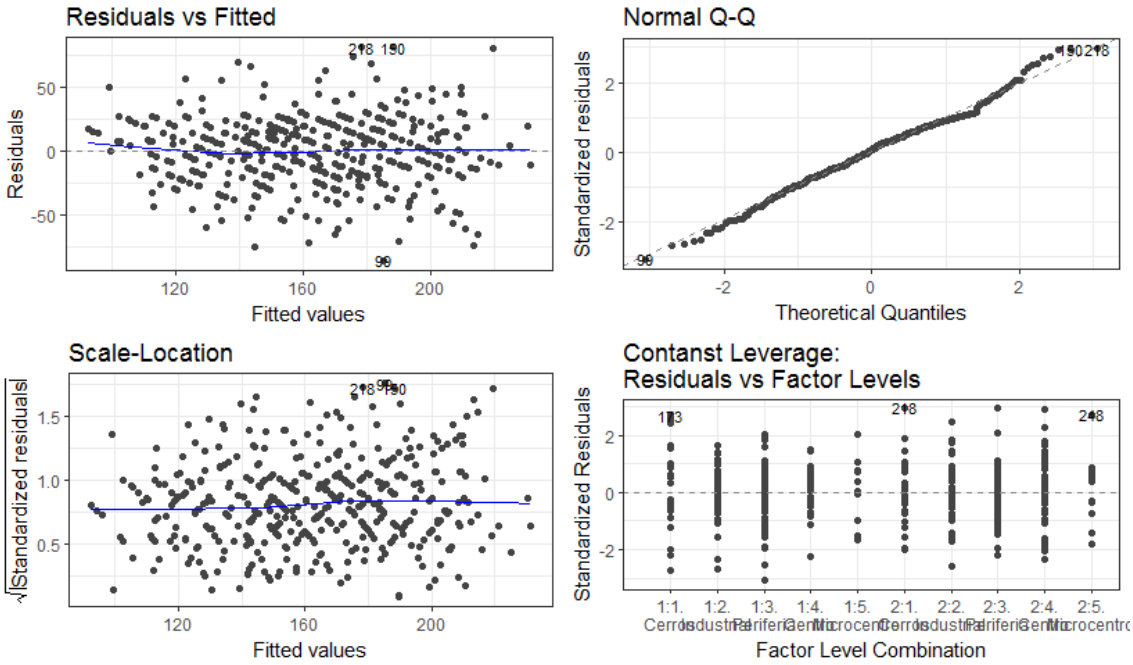


Figura 4.2: Escenario I - Residuos Etapa 1.

El primer gráfico "residuos vs. valores ajustados" es una herramienta clara para reconocer posibles patrones de heterocedasticidad, en este caso, se puede observar un leve aumento en la dispersión con respecto al incremento de valores ajustados, también se reconoce la presencia de posibles outliers.

El segundo gráfico "QQ plot" al igual que los histogramas presentados en la Figura 4.1, se busca chequear la normalidad de los residuos donde se espera que los valores se aproximen a la recta $y = x$, en este caso se observa que se cumple para todos los modelos con una pequeña distorsión en las colas.

El tercer gráfico "Localización-Escala" también busca verificar el supuesto de homocedasticidad, muestra si la dispersión de los residuos es uniforme en función de los predictores. Un buen indicio de homocedasticidad, es cuando se visualiza una

tendencia lineal horizontal con pendiente cero. En este caso vemos que la dispersión no es equitativa en ningún modelo, así como la línea azul no se muestra constante. También se reconoce la presencia de puntos raros.
El cuarto gráfico "Residuos vs. Niveles de Factor", presenta los residuos frente a los factores de cierta variable, busca comprobar si la variación no contabilizada por el modelo es diferente para los diferentes niveles de un factor. En este caso, el gráfico varía en cada modelo ya que las variables categóricas elegidas no son las mismas. Para algún nivel de ellas se observa cierta concentración en los puntos y en otros casos la dispersión es similar.

4.2. Escenario II

El segundo escenario de modelos, busca ajustar la capacidad pulmonar medida en las últimas tres etapas de monitoreo, se considera dentro del conjunto de las variables explicativas, al valor del PFE en la etapa inicial. Análogamente al escenario I, se considera cada modelo como instancias independientes entre sí.
Del conjunto de modelos analizados en cada caso, se seleccionan por criterio de menor AIC, los modelos presentados en la Tabla 4.3, donde se destaca que las variables *PFE inicial* y *estrato*, son consideradas en todos los modelos para explicar el valor de la capacidad pulmonar. De los tests de significación de las variables, se comprueba que *PFE inicial*, es significativo al 0 % para explicar el PFE.

ML_II_E2	$PFE_{i2} = -68,0 + 8,9$ edad$_{i2} + 1,0$ talla$_{i2} + 0,4$ PFE$_{i1}$ -12,7 estrato_público$_{i2}$
ML_II_E3	$PFE_{i3} = -89,3 + 1,6$ talla$_{i3} + 0,5$ PFE$_{i1}$ +2,3 zona_industrial$_{i1} - 4,7$ zona_periferia$_{i1}$ -14,4 zona_centro$_{i1} - 5,9$ zona_microcentro$_{i1}$ -7,1 estrato_público$_{i3}$
ML_II_E4	$PFE_{i4} = 22,0 + 5,9$ edad$_{i4} + 1,6$ peso$_{i4} + 0,5$ PFE$_{i1}$ -4,7 patología$_{i4} - 9,3$ estrato_público$_{i4}$

Tabla 4.3: Escenario II - Modelos estimados en cada etapa.

En cuanto a la significación de los modelos, se cumple en todos los casos que el estadístico F (ver apéndice Tabla D.2), presenta mayores valores que su correspondiente valor teórico, descartando la hipótesis planteada de que todos los coeficientes estimados sean nulos.

Diagnóstico de los modelos: Escenario II

Para validar los modelos seleccionados, se analizan los supuestos sobre los errores. En primer lugar, al igual que en el escenario I, se grafican los histogramas de los residuos con el fin de testear si correspondan a una distribución normal.

De la Figura 4.3, se puede apreciar que los histogramas presentan distribución con media 0.

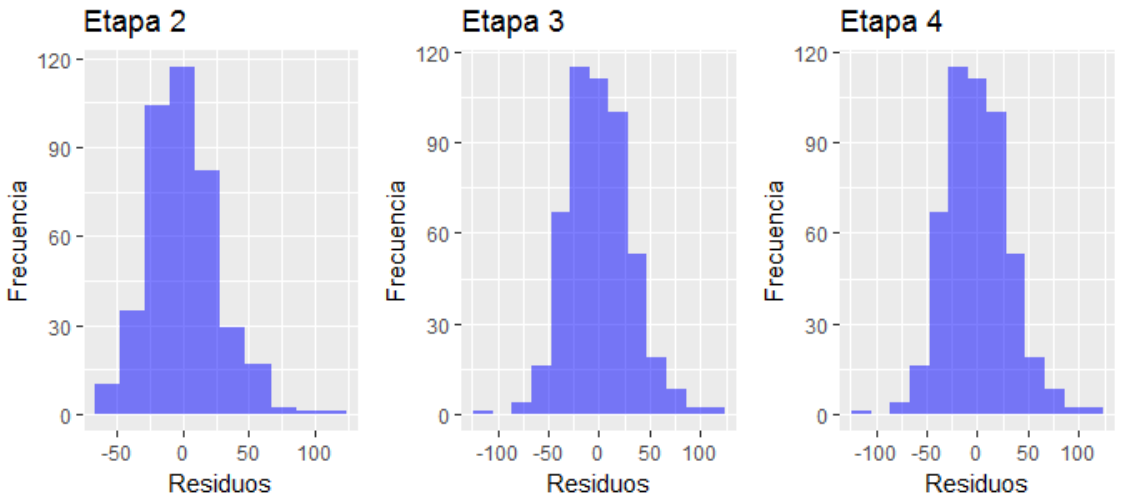


Figura 4.3: Escenario II - Histogramas residuos.

Se continúa el análisis, con los gráficos de los errores:

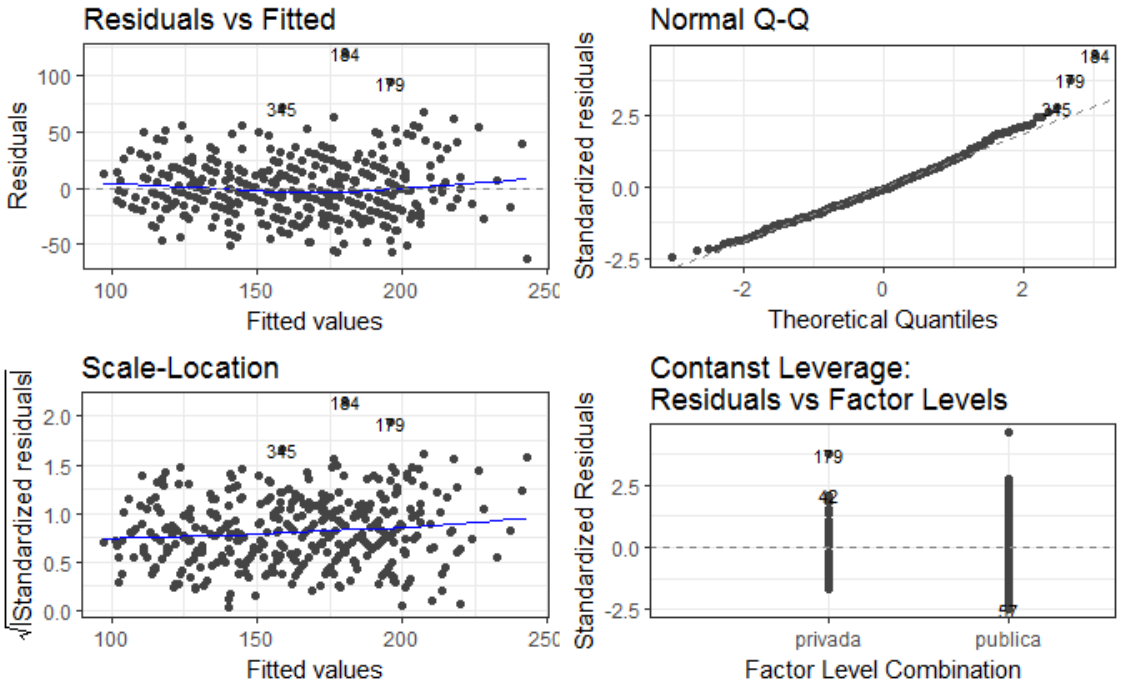


Figura 4.4: Escenario II - Residuos Etapa 2.

A partir de los gráficos presentados en la Figura 4.4 y en las Figuras D.4 y D.5 del apéndice D, similar a lo observado para el escenario I, se puede apreciar en el gráfico "residuos vs. valores ajustados", que los tres modelos presentan un leve aumento de

la dispersión de los errores, que puede estar asociado a heterogeneidad de varianzas. La interpretación es análoga para los restantes gráficos.

4.3. Escenario III

Hasta ahora, se ajustaron modelos como fotos independientes en el tiempo. En este escenario y los siguientes, se desea ajustar los datos relevados como observaciones independientes entre etapas por un único modelo. Se desea explicar la variabilidad de la capacidad pulmonar a partir de las variables explicativas antes mencionadas y además se añade la variable "etapa de monitoreo" como variable relevante para explicar el comportamiento de la capacidad pulmonar. Por lo que un participante que haya estado en los cuatro monitoreos se encontrará 4 veces especificándose sus datos recolectados del monitoreo así como los datos obtenidas en la primer etapa resultado del cuestionario.

En este escenario, se desea encontrar un modelo con los supuestos clásicos. El cometido principal, es estudiar qué variables resultan significativas para explicar el PFE cuando se analizan todas las etapas en conjunto. Del análisis realizado en el capítulo anterior y en el escenario I, se comprueba la importancia de la *talla* para explicar el PFE, a partir del escenario II, parece determinante considerar el PFE de la etapa inicial para explicar la capacidad pulmonar de la etapa siguiente.

De todos los modelos evaluados bajo este escenario, se selecciona a partir de los criterios de menor AIC y mayor R^2 el siguiente:

$$\begin{aligned} PFE_i = &-59,7 + 0,5\text{PFE}_{i1} + 4,7\ \text{edad}_i + 0,6\ \text{peso}_i + 0,9\ \text{talla}_i \\ &- 1,2\ \text{zona\_industrial}_i - 4,8\ \text{zona\_periferia}_i \\ &- 9,1\ \text{zona\_centro}_i + 0,001\ \text{zona\_microcentro}_i \\ &+ 2,5\ \text{etapa\_3} + 10,2\ \text{etapa\_4}_i \end{aligned} \tag{4.2}$$

Para el modelo del escenario III especificado en (4.2), todas las variables son significativas al 5 %. En cuánto a la significación del modelo, el estadístico F (ver apéndice

Tabla D.3) indica que el modelo resulta significativo al 5 % con R^2 de 0,59 y con el menor AIC, de los modelos realizados bajo este escenario.

Parece oportuno mencionar que dentro de los modelos planteados, se decidió también incorporar interacción entre la zona de residencia y la etapa, lo que no resultó significativo para explicar el PFE.

Diagnóstico del modelo: Escenario III

En este escenario la validación de los supuestos es un paso importante, ya que se desea verificar los supuestos realizados desde el comienzo.

En primer lugar, se presenta el histograma de los residuos (Figura 4.5).

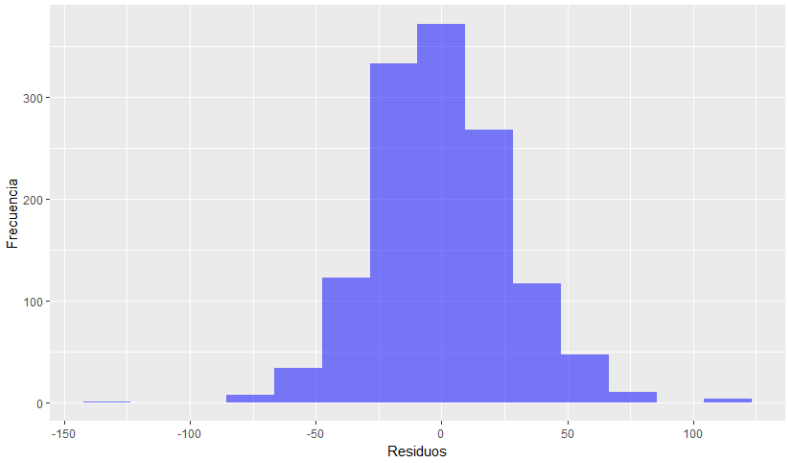


Figura 4.5: Escenario III - Histograma residuos.

De la Figura 4.5, se puede apreciar simetría y se reconocen posibles outliers sobre ambos extremos.

En segundo lugar, se desea analizar los gráficos de los errores para verificar su comportamiento y descartar ausencia de patrones.

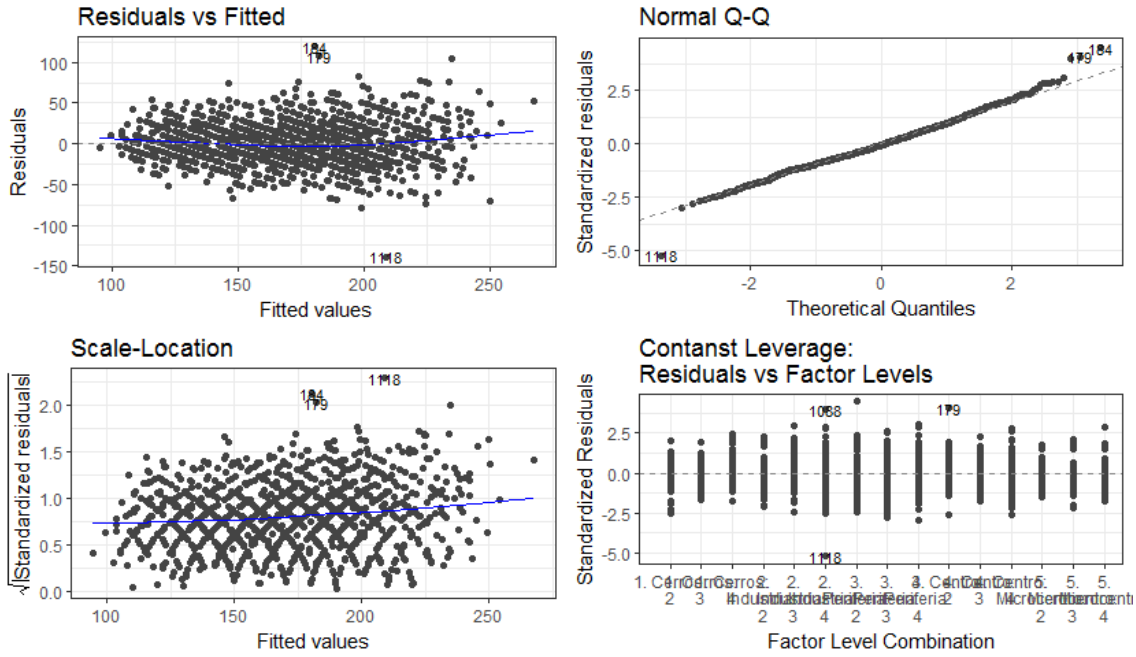


Figura 4.6: Escenario III - Residuos.

En el primer gráfico de la Figura 4.6, se observa un patrón de mayor dispersión de los residuos a medida que aumentan los valores ajustados. Al igual que en los escenarios anteriores, podría reconocerse la presencia de algún dato atípico. Del gráfico "QQ plot", se desprende que los residuos estandarizados se comportan a una distribución Normal, los puntos se ajustan a la recta $y = x$, con pequeña distorción en las colas que pueden asociarse también, a presencia de puntos atípicos. De la gráfica de localización y escala se denota un patrón que sugiere "no aleatoriedad" de los residuos, además la recta no toma valores constantes.

Para testear el supuesto de homocedasticidad ser realizó la siguiente prueba de hipótesis[4]:

H0) Varianza Constante

H1) No H0

Dicha prueba con $p - valor < 0,0001$, es menor a un nivel de significación del 5 %, por lo tanto, podemos rechazar la hipótesis nula y concluir la existencia de heterocedasticidad, lo que confirma lo observado gráficamente.

[4]NCV: Non-constant Variance Score Test.

4.4. Escenario IV

A partir de los análisis anteriores, se decide levantar el supuesto de homocedasticidad, contemplando el modelo antes elegido para el escenario III, especificado en (4.2). En esta sección, se evaluaron distintas funciones de varianzas, entre ellas, se seleccionó la función de varianza del conjunto $< \delta >$, que implicó mejor ajuste de los datos. Así, se obtienen las siguientes estimaciones:

Efectos Fijos	*Parámetros*	*coef (D.E)*
Intercepto	β_0	-65,1 (15,8)
PFE.1	β_1	0,42 (0,03)
edad	β_2	4,54 (1,07)
peso	β_3	0,54 (0,19)
talla	β_4	1,03 (0,20)
Zona Industrial	$\beta_{5,2}$	-1,38 (2,64)
Zona Periferia	$\beta_{5,3}$	-4,70 (2,25)
Zona Centro	$\beta_{5,4}$	-9,52 (2,57)
Zona Microcentro	$\beta_{5,5}$	-1,61 (3,57)
Etapa 3	$\beta_{6,3}$	2,72 (1,7)
Etapa 4	$\beta_{6,4}$	9,5 (1,79)
	Parámetros	*coef (I.C)*
Función de varianza	δ	0,80 (0,60 - 0,99)
Escala	σ	0,44 (0,16 - 1,20)
Log-MVR		-6140
AIC		12305
BIC		12373

Tabla 4.4: Escenario IV - Modelo estimado.

Al igual que en los escenarios anteriores, se analizan los residuos para validar el

modelo seleccionado.

Diagnóstico de los modelos: Escenario IV

Aunque el criterio AIC, sugiere que el modelo con la función de varianza elegida es el que mejor se ajusta, puede ignorar la correlación de las observaciones. Por lo cual, se evalúa el ajuste del modelo utilizando los gráficos de los residuos. Se crean diagramas de dispersión de los valores residuales frente a los valores ajustados en función de la etapa.

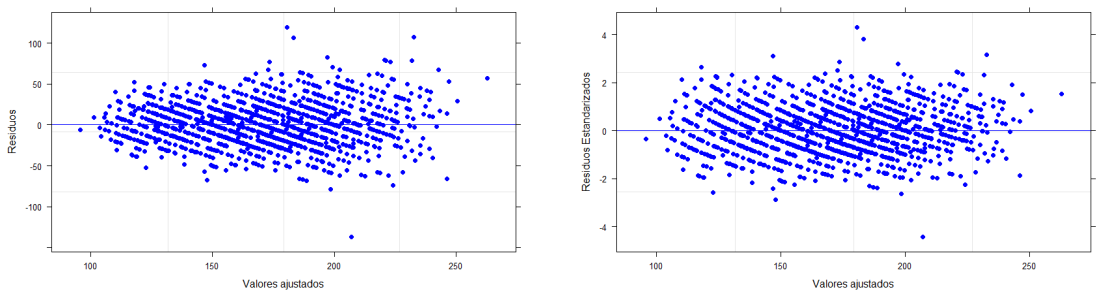


Figura 4.7: Escenario IV - Dispersión: (a) residuos vs. valores ajustados (b) residuos de Pearson vs. valores ajustados.

La Figura 4.7a muestra un patrón asimétrico, donde los residuos presentan mayor dispersión a medida que aumentan los valores ajustados.
En el diagrama de dispersión de los residuos de Pearson frente a los valores ajustados (Figura 4.7b) , se atenúa el patrón observado.

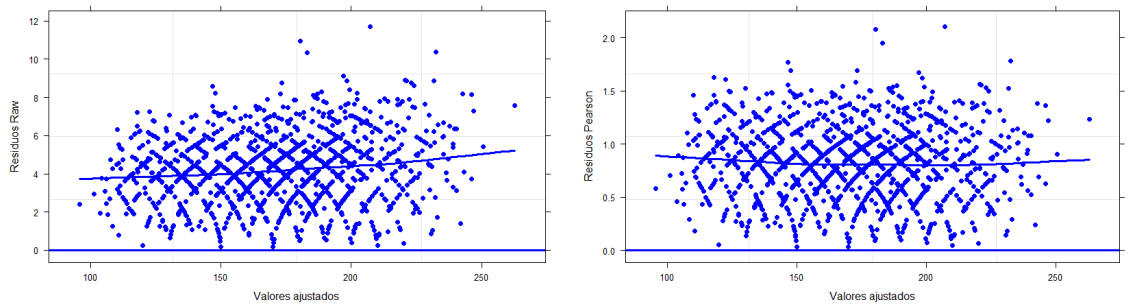


Figura 4.8: Escenario IV - Localización y escala: (a) residuos vs. valores ajustados (b) residuos de Pearson vs. valores ajustados.

Los gráficos presentados en la Figura 4.8 muestran un patrón que puede estar asociado a la no aleatoriedad de los residuos.

A continuación se presenta el gráfico de correlación del PFE respecto a las etapas.

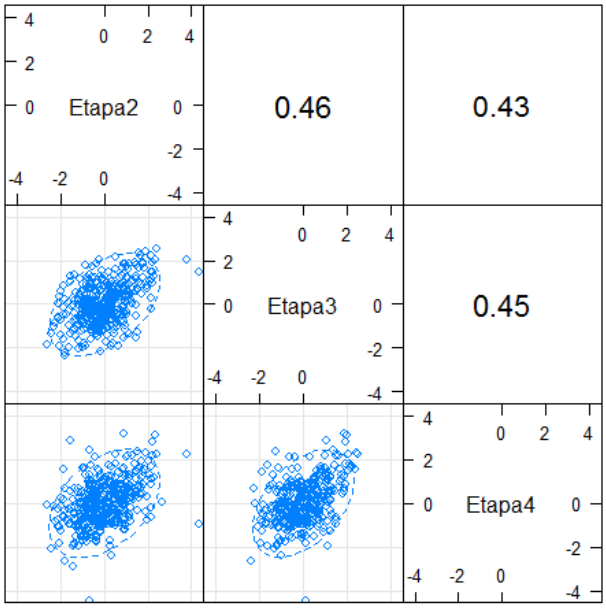


Figura 4.9: Escenario IV - Matriz de dispersión: residuos de Pearson.

La Figura 4.9 presenta una matriz de dispersión de los residuos de Pearson para las 3 ocasiones de medición. Los diagramas de dispersión muestran una correlación

de los residuos entre las etapas lo cual puede implicar una violación del supuesto de independencia entre las observaciones. Se realiza el test de hipótesis de Durbin-Watson (7), como resultado se obtiene un valor de estadístico $d = 2{,}003$ por lo que se confirma la correlación de los errores, dado que cuando el estadístico toma valores mayores a 1,96, con nivel de significación al 5 %, se rechaza la hipótesis nula de incorrelación.

4.5. Escenario V

En esta oportunidad, se levanta el supuesto de incorrelación. El modelo seleccionado en el escenario anterior, ahora es ajustado utilizando distintas funciones de estructuras de correlación. El modelo que mejor ajusta los datos, considera una estructura de correlación de "simetría compuesta" (C-S), que utiliza el mismo coeficiente de correlación para diferentes observaciones. Es decir, se permite correlación constante de las mediciones de la capacidad pulmonar relevadas en diferentes puntos de tiempo para el mismo individuo.

Las estimaciones de los parámetros del modelo elegido para el escenario V son los descritos en la Tabla 4.5.

Efectos Fijos	*Parámetros*	*coef (D.E)*
Intercepto	β_0	-53,74 (20,40)
PFE.1	β_1	0,41 (0,04)
edad	β_2	5,16 (1,42)
peso	β_3	0,78 (0,29)
talla	β_4	0,86 (0,27)
Zona Industrial	$\beta_{5,2}$	-2,40 (3,51)
Zona Periferia	$\beta_{5,3}$	-4,30 (3,02)
Zona Centro	$\beta_{5,4}$	-9,68 (3,43)
Zona Microcentro	$\beta_{5,5}$	-1,73 (4,81)
Etapa 3	$\beta_{6,3}$	2,98 (1,32)
Etapa 4	$\beta_{6,4}$	9,50 (1,47)
	Parámetros	*coef (I.C)*
Funciones de varianza	δ	0,86 (0,67 - 1,06)
Estructura de Correlación	ϱ_{CS}	0,45 (0,39 - 0,50)
Escala	σ	0,32 (0,12 - 0,85)
Log-MVR		-6028
AIC		12084
BIC		12156

Tabla 4.5: Escenario V - Modelo estimado.

La Tabla 4.5 presenta la salida de las estimaciones puntuales con sus correspondientes desvíos estándar y los límites inferior y superior de los intervalos de confianza, para el coeficiente de potencia δ, el coeficiente de estructura de correlación ϱ, y el coeficiente de escala σ. Los resultados indican que, el coeficiente de correlación de cualquiera de las dos mediciones de capacidad pulmonar obtenidas para el mismo ϱ del niño es igual a 0,45 lo que confirma que existe una correlación significativa entre las mediciones de la capacidad pulmonar. El coeficiente de potencia estimado de la

función de varianza, 0,86, es mayor al valor de 0.80 obtenido para el modelo ML_IV, lo que indica una variabilidad creciente de las mediciones a lo largo del tiempo.

	Modelo	df	AIC	BIC	Log-MVR	Test	Razón	p-valor
ML_IV	1	13	12305	12373	-6.140			
ML_V	2	14	12084	12156	-6.028	1 vs. 2	223,30	$<$ 0,0001

Tabla 4.6: Prueba de independencia vs. estructura de correlación.

En la Tabla 4.6 se presenta el test LR para testear la importancia de incluir la correlación en el ajuste del modelo. El resultado de la prueba LR es estadísticamente significativo, lo que indica la importancia del ajuste para la correlación en el modelado de los datos.

Diagnóstico del modelo: Escenario V

Para este escenario, se analizan en detalle los residuos de Pearson, los cuales se obtienen dividiendo los residuos por sus correspondientes desvíos estándar, por lo que sus gráficos de dispersión podrían ser más fáciles de interpretar. Sin embargo, debido a que los residuos están correlacionados dentro de las etapas, se requiere cierto grado de precaución al interpretar los gráficos. Además, se puede construir un gráfico más informativo, si se considera que los residuos para cada momento del tiempo no están correlacionados. Por lo tanto, es más apropiado presentarlos por separado para cada etapa, ver Figuras 4.10 y 4.11.

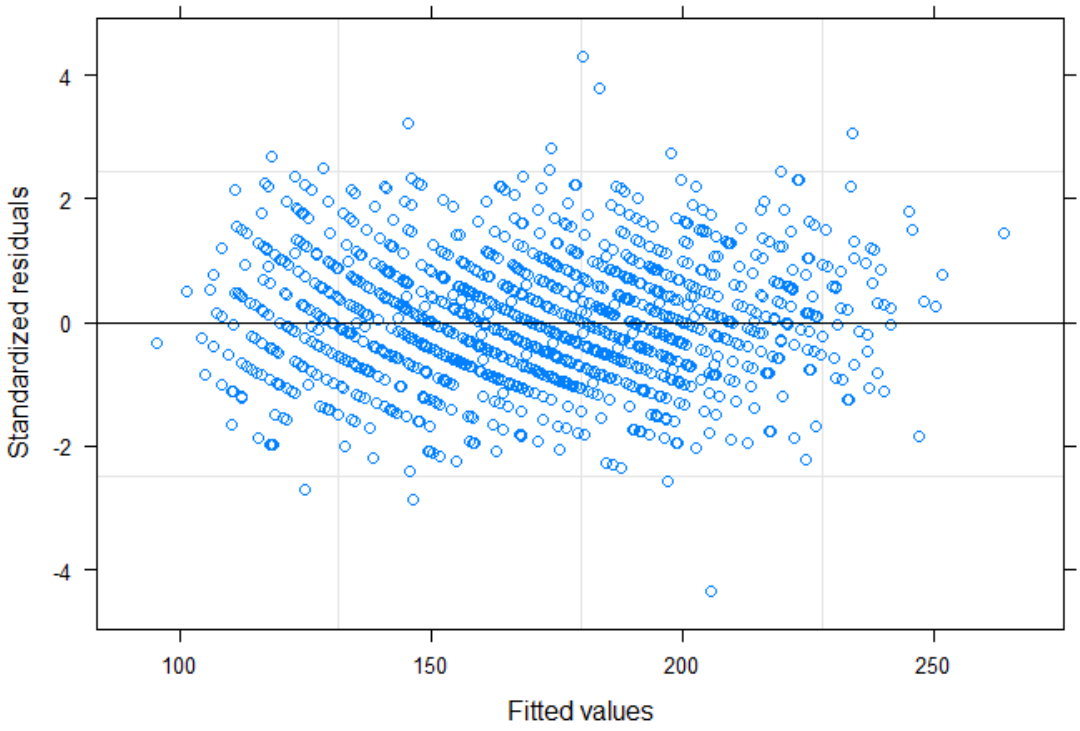


Figura 4.10: Escenario V - Dispersión: residuos de Pearson vs. valores ajustados.

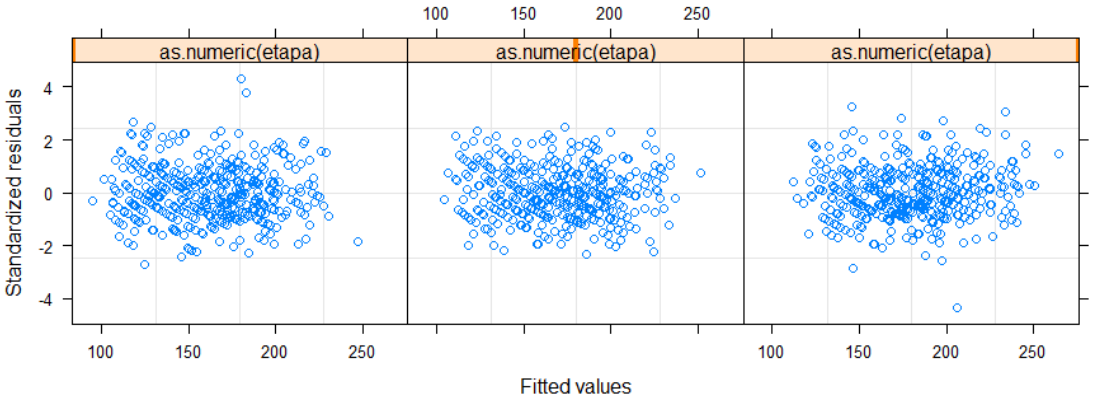


Figura 4.11: Escenario V - Dispersión: residuos de Pearson vs. valores ajustados según etapa.

En las Figuras 4.10 y 4.11 se muestran los diagrama de dispersión de los residuos de Pearson frente a los valores ajustados. En el diagrama de dispersión, los residuos de las observaciones a lo largo del tiempo se grafican conjuntamente. Como resultado,

debido a la correlación de los residuos correspondientes a las mediciones obtenidas para el mismo individuo en diferentes etapas, la Figura 4.10 sigue revelando cierto patrón que puede contemplar la no aleatoriedad en los residuos.

El gráfico resultante por etapa se muestra en la Figura 4.11. Por el contrario, los tres diagramas de dispersión no muestran un patrón.

En la Figura 4.12 se presentan los diagramas de dispersión de los residuos de Pearson agrupados por etapa y zona de residencia. El principal problema en la interpretación de los residuos de Pearson es el hecho de que los errores están correlacionados.

Para eliminar la correlación entre los residuos de Pearson, podemos usar los residuos normalizados, que se obtienen de una transformación de los residuos basados en la descomposición de Cholesky de la matriz de varianza-covarianza de los residuos.

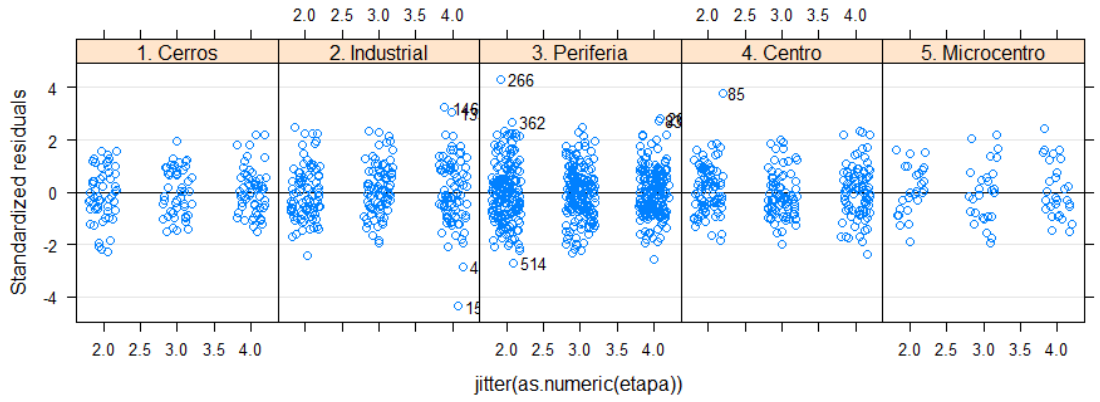


Figura 4.12: Escenario V - Dispersión: residuos de Pearson según etapa y zona de residencia.

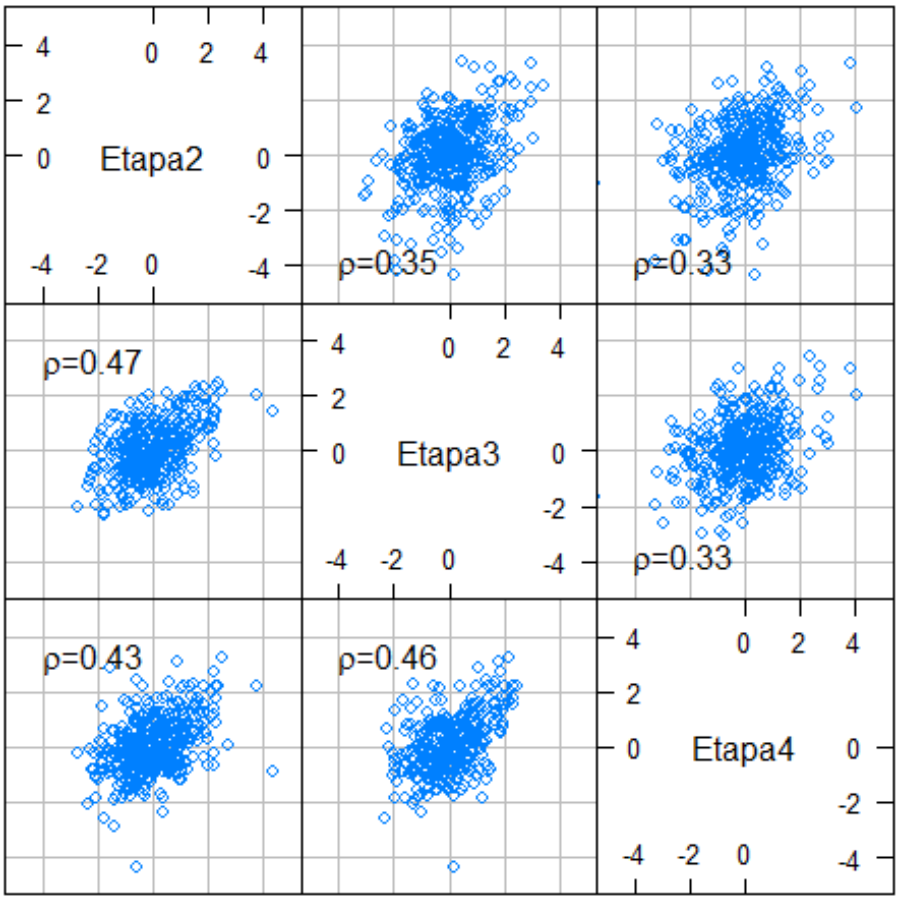


Figura 4.13: Escenario V - Matriz de dispersión: residuos de Pearson y residuos normalizados.

La Figura 4.13 muestra los diagramas de dispersión de los residuos de Pearson (debajo de la diagonal) y los residuos normalizados (encima de la diagonal) para todos los pares de puntos de cada etapa para el modelo ML_V. Los diagramas de dispersión de los residuos de Pearson muestran una correlación entre los residuos correspondientes a diferentes puntos de tiempo. Por otro lado, los gráficos de los residuos normalizados ilustran una menor correlación.

Los patrones mostrados en la Figura 4.14 se aproximan a la recta $y = x$, por lo que, el supuesto de normalidad parece cumplirse.

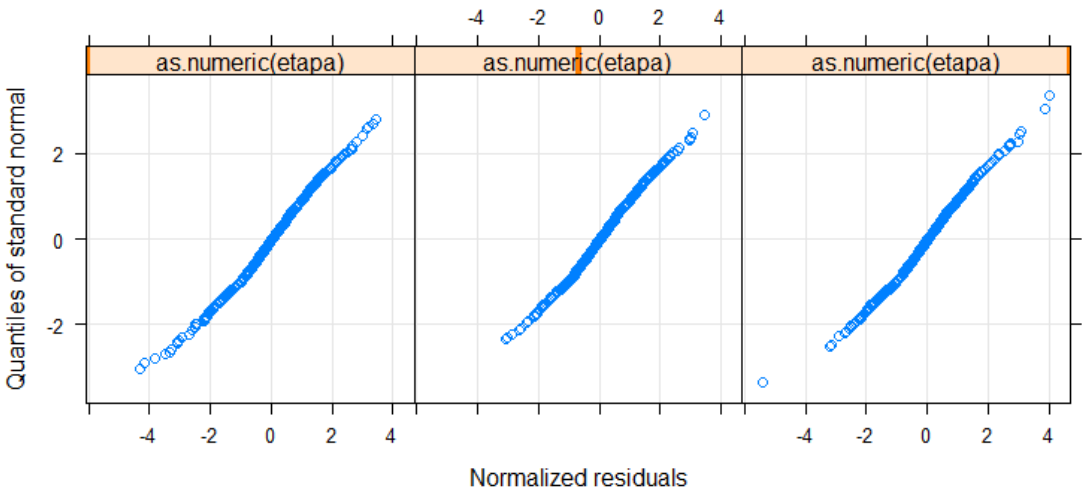


Figura 4.14: Escenario V - QQ plot: residuos normalizados.

4.6. Escenario VI: Modelos Mixtos

Finalmente se introducen los efectos aleatorios a las variables explicativas. A partir de la función de varianza seleccionada en el escenario IV se obtiene:

$$\mathcal{R}_i = \sigma^2 \begin{pmatrix} (\text{etapa}_{i2})^{2\delta} & 0 & 0 \\ 0 & (\text{etapa}_{i3})^{2\delta} & 0 \\ 0 & 0 & (\text{etapa}_{i4})^{2\delta} \end{pmatrix} \tag{4.3}$$

Donde $\mathcal{R}_i$, puede descomponerse como $\mathcal{R}_i = \sigma_i^2 \mathbf{\Lambda}_i \mathbf{C}_i \mathbf{\Lambda}_i$ con $\mathbf{\Lambda}_i$ definida en (2.48) y con $\mathbf{C}_i = \mathbf{I}_4$. Debe destacarse aquí que el parámetro σ^2 solo puede interpretarse como un parámetro de escala (desconocido).

La matriz $\mathcal{R}_i$, definida en (4.3), es diagonal con elementos desiguales definidos por la función de varianza. La matriz de varianza-covarianza marginal resulta:

$$\mathcal{V}_i = \begin{pmatrix} \sigma_1^2 + d_{11} & d_{11} & d_{11} \\ d_{11} & \sigma_2^2 + d_{11} & d_{11} \\ d_{11} & d_{11} & \sigma_3^2 + d_{11} \end{pmatrix} \tag{4.4}$$

donde,

$$\sigma_t^2 = \sigma^2(\text{etapa}_{it})^{2\delta}$$

Vale la pena observar que, debido a que la varianza cambia con el tiempo, los coeficientes de correlación marginal entre las observaciones realizadas en diferentes momentos ya no son iguales.

La Tabla 4.7 presenta un resumen de las estimaciones de los parámetros del modelo.

Efectos Fijos	*Parámetros*	coef	Error Std.	p-valor
(Intercepto)	β_0	-49,36	21,27	0,02
PFE.1	β_1	0,45	0,04	0,00
edad	β_2	4,84	1,46	0,00
peso	β_3	0,81	0,27	0,00
talla	β_4	0,78	0,27	0,00
Zona Industrial	$\beta_{5,2}$	-2,00	3,59	0,58
Zona Periferia	$\beta_{5,3}$	-4,46	3,19	0,16
Zona Centro	$\beta_{5,4}$	-9,49	3,61	0,01
Zona Microcentro	$\beta_{5,5}$	-0,36	4,90	0,94
Etapa 3	$\beta_{6,2}$	3,04	1,36	0,03
Etapa 4	$\beta_{6,3}$	10,66	1,59	0,00

Tabla 4.7: Escenario VI - Modelo estimado.

Los resultados presentados en las Tablas 4.7 y 4.8 indican que el parámetro de escala σ estimado es 20,4. Los coeficientes de potencia δ de la función de varianza para las etapas 3 y 4, toman los valores 0,88 y 1,07 respectivamente. La estimación de la

	Parámetros	*coef (I.C)*
Reestructura: SD (b_{0i})	$\sqrt{d_{1,1}}$	17,54 (15,92 19,33)
Función varianza: ($etapa_3{}^{\delta}$)	δ	0,88 (0,76 1,01)
Función varianza: ($etapa_4{}^{\delta}$)	δ	1,07 (0,95 1,22)
Escala	σ	20,37 (18,67 22,23)
Log-MVR		-6069
AIC		12168
BIC		12246

Tabla 4.8: Escenario VI - Modelo estimado.

desviación estándar de las intersecciones aleatorias es de 17,5.

La varianza estimada de los interceptos aleatorios es igual a 308. Se debe tener presente que es el menor valor obtenido para todos los modelos realizados. Esto se espera, ya que, al permitir errores aleatorios residuales heterocedásticos, una gran parte de la variabilidad total se explica por las variaciones residuales.

La matriz de varianza-covarianza estimada $\mathcal{R}_i$ obtenida para el modelo elegido es:

$$\mathcal{R}_i = \begin{pmatrix} 479 & 0 & 0 \\ 0 & 479 & 0 \\ 0 & 0 & 479 \end{pmatrix}$$

La matriz de varianza, covarianza marginal ($\mathcal{V}_i$) estimada correspondiente indica una correlación decreciente entre las mediciones de capacidad pulmonar de los niños realizadas en puntos de tiempo más distantes.

$$\mathcal{V}_i = \begin{pmatrix} 787 & 308 & 308 \\ 308 & 787 & 308 \\ 308 & 308 & 787 \end{pmatrix}$$

Hay que tener presente que la comparación directa de las matrices de covarianza de varianza marginal de los modelos de los Escenarios V y VI no es apropiada. Esto

se debe a que la matriz de varianza-covarianza marginal del modelo del Escenario VI, que se muestra, es mucho más estructurada que la del modelo elegido en el Escenario V. Por otro lado, ambos modelos permiten los coeficientes de correlación marginal, que dependen del tiempo "distancias", o "posiciones", de las mediciones de la capacidad pulmonar.

Diagnóstico del modelo: Escenario VI

En la Figura 4.15 se muestran los residuos de Pearson frente a los valores ajustados, al graficar a todos los residuos en conjunto no resulta relevante. Sin embargo, puede servir para detectar otros comportamientos, como por ejemplo la presencia de valores atípicos.

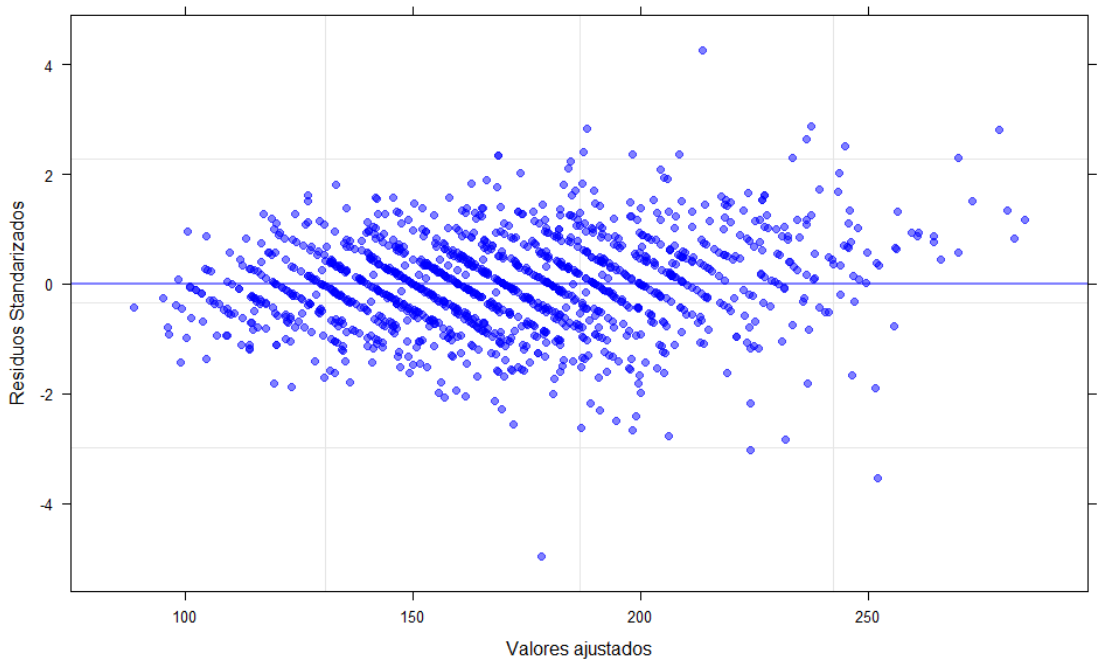


Figura 4.15: Escenario VI - Dispersión: residuos de Pearson condicionales vs. valores ajustados.

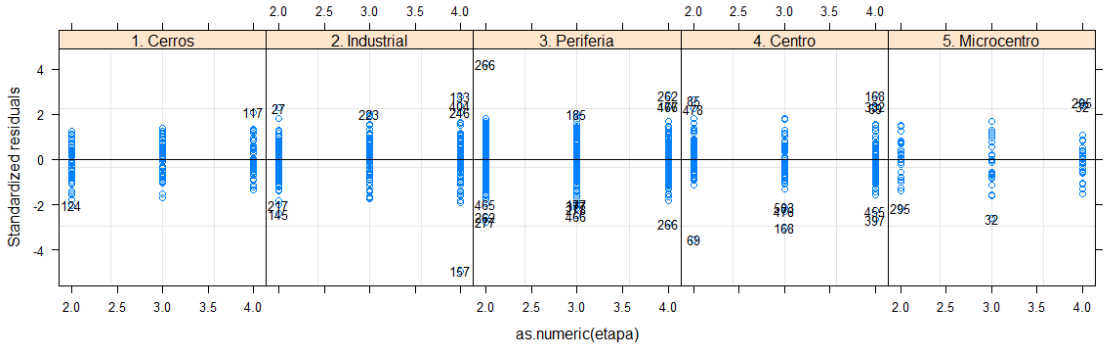


Figura 4.16: Escenario VI - Dispersión: residuos de Pearson condicionales según etapa y zona de residencia.

El gráfico de la Figura 4.16 permite una evaluación de la distribución de los residuos de Pearson condicionales para cada fase de monitoreo y zona de residencia. A pesar de la estandarización, la variabilidad de los residuos parece ser aleatoria. También se reconoce la presencia de posibles valores atípicos. Salvo para la etapa 3 en las zonas Cerros e Industrial, se reconocen puntos raros en todos las *zonas de residencia* y *etapas* de monitoreo.

La Figura 4.17 muestra la gráfica "QQ plot" de los residuos condicionados de residuos de Pearson por etapa. Los patrones muestran algunas desviaciones de una tendencia lineal.

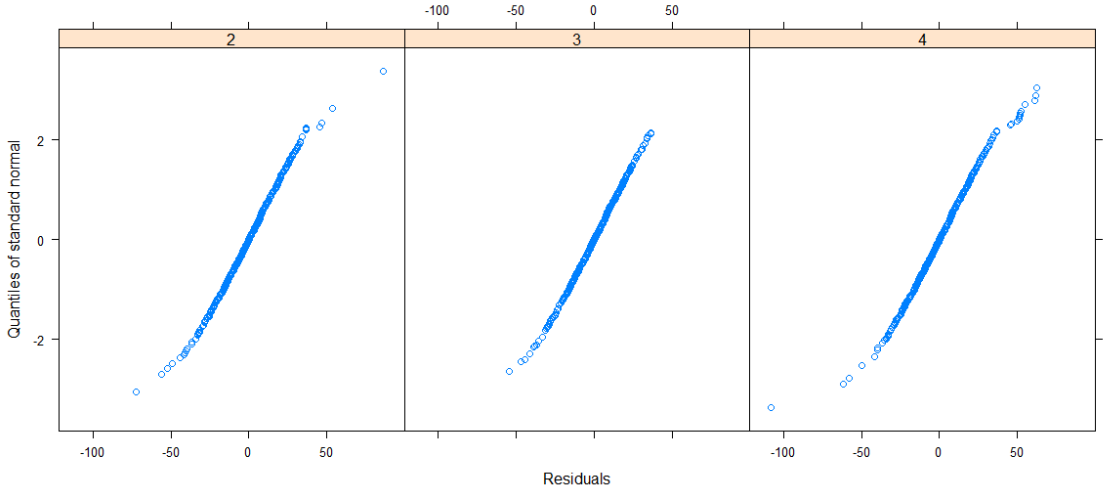


Figura 4.17: Escenario VI - QQ plot: residuos de Pearson condicionales según etapa.

En la Figura 4.17 se grafican los valores QQ plot de los efectos aleatorios estimados (interceptos aleatorios). Se puede apreciar que es ligeramente curvilínea, que podría tomarse como una indicación de no normalidad de los efectos aleatorios.

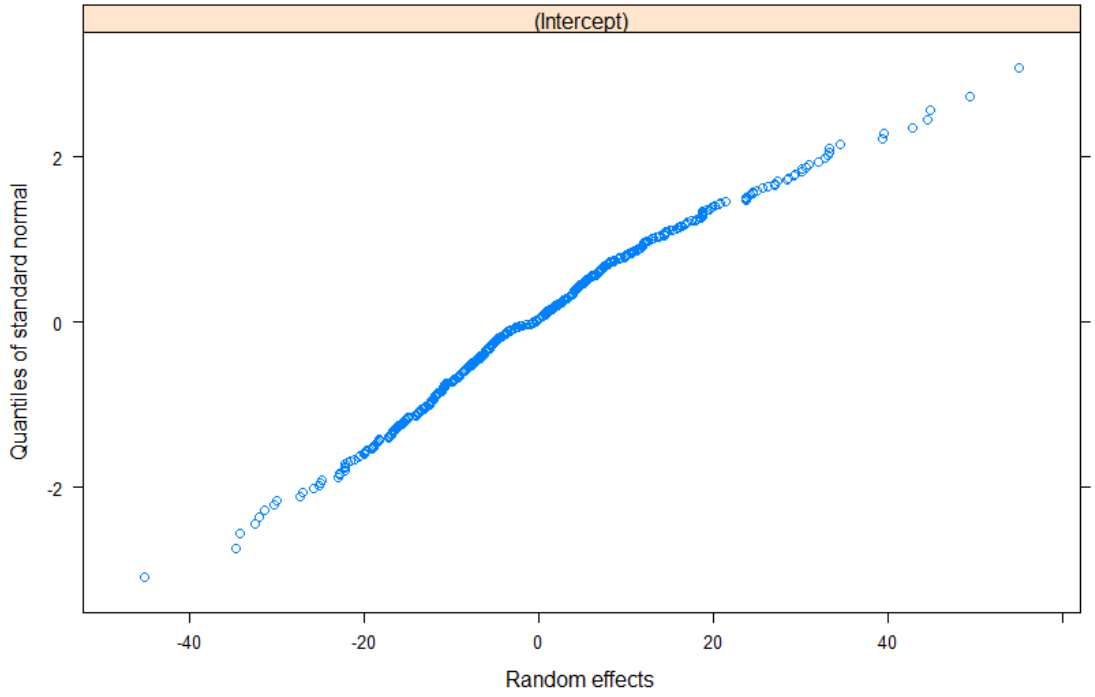


Figura 4.18: Escenario VI - QQ plot: Interceptos aleatorios estimados.

4.7. Interpretación de Resultados

Los escenarios I, II y III muestran varias dificultades para ajustar los datos, ya que no contemplan la correlación de las observaciones y la heterocedasticidad de los errores como fue demostrado.

El modelo del escenario IV, aunque contemple la heterocedasticidad sigue sin tener en cuenta la correlación de los datos.

Por lo tanto, para las resultados que se desean obtener, los modelos de los escenarios V y VI son los que mejor ajustan a los datos de estudio.

Por la complejidad del modelo lineal de efectos mixtos que incorpora variables a

estimar en los efectos aleatorios, la herramienta para selección de modelos que se utilizó en los escenarios anteriores, el criterio de información de Akaike, ya no resulta adecuada ya que penaliza la incorporación de nuevas variables a la estimación. A continuación, se muestra una tabla comparativa de los modelos estimados para los escenarios V y VI:

Efectos fijos	*Parámetros*	*coef (D.E)*	
		Modelo V	*Modelo VI*
Intercepto	β_0	-53,54 (20,40)	-49,36 (21,27)
PFE.1	β_1	0,41 (0,04)	0,45 (0,04)
edad	β_2	5,16 (1,42)	4,84 (1,46)
peso	β_3	0,78 (0,29)	0,81 (0,27)
talla	β_4	0,86 (0,27)	0,78 (0,27)
Zona Industrial	$\beta_{5,2}$	-2,49 (3,51)	-2,00 (3,59)
Zona Periferia	$\beta_{5,3}$	-4,30 (3,02)	-4,46 (3,19)
Zona Centro	$\beta_{5,4}$	-9,68 (3,43)	-9,49 (3,61)
Zona Microcentro	$\beta_{5,5}$	-1,73 (4,81)	-0,36 (4,90)
Etapa 3	$\beta_{6,3}$	2,98 (1,32)	3,04 (1,36)
Etapa 4	$\beta_{6,4}$	9,50 (1,47)	10,66 (1,59)
	Parámetros	*coef (I.C)*	
Funciones de varianza	δ	0,86 (0,67 1,96)	0,88 (0,76 1,01)
			1,07 (0,95 1,22)
Estr. Corr (C-S)	ϱ_{CS}	0,45 (0,39 0,50)	
Reestruc.(SD (b_{i0}))	$\sqrt{d_{11}}$		17,54 (15,92 19,33)
Escala	σ	0,32 (0,12 0,85)	20,37 (18,67 22,23)
Log-MVR		-6028	-6069
AIC		12084	12168
BIC		12156	12246

Tabla 4.9: Escenario V vs. Escenario VI - Comparación modelos estimados.

Al comparar los modelos estimados en cada escenario, se observa que el modelo del escenario VI, presenta mayor AIC, lo que es resultado de que el modelo con efectos mixtos presenta mayor cantidad de parámetros a estimar. Pero al analizar el error cuadrático medio de este con respecto al modelo obtenido en el escenario V se obtienen mejores resultados en el escenario VI. Entonces, se opta por considerar el modelo lineal de efectos mixtos como el modelo que mejor ajusta la capacidad pulmonar medida longitudinalmente.

Para este modelo, se puede interpretar que la zona de residencia del niño infiere en la capacidad pulmonar, dado que dejando las demás variables constantes, los niños de todas las zonas constatan PFE por debajo de la zona cerros, siendo que en promedio tienen una capacidad respiratoria 9,49 l/min menor si reside en el centro, 4,46 l/min menor si reside en la periferia, 2,00 l/min por debajo si reside en la zona industrial y 0,36 l/min menor si reside en el microcentro de la ciudad. Además, dejando las demás variables constantes, en la etapa 3 los niños registraron una capacidad pulmonar de 3,04 l/min por encima que en la etapa 2 y en la cuarta etapa incrementan su capacidad en 10,66 l/min respecto a la etapa 2.

Las variables como la *edad*, el *peso* y la *talla*, confirman los resultados mostrados en la investigación (2), en cuanto a la relación de estas variables con la capacidad pulmonar.

Capítulo 5

Conclusiones

En este trabajo se desea determinar qué variables inciden en la capacidad pulmonar de los escolares de la ciudad de Artigas. Particularmente, evaluar las variables asociadas al proceso de industrialización del arroz, a los antecedentes clínicos y contaminación intradomiciliaria del escolar.

Las conclusiones de este trabajo se presentan a continuación, teniendo en consideración los resultados obtenidos a partir de cada uno de los objetivos propuestos.

Respecto al primer objetivo: **"Comparar el valor de Pico Flujo Espiratorio de niñas y varones de la ciudad de Artigas con respecto a valores percentilares de las curvas de referencia",** se contrastan las mediciones de PFE obtenidas en cada etapa del monitoreo con los valores percentilares de una población sin patologías y en condiciones ambientales óptimas y se ajusta por sexo y talla.
La cantidad de niñas por debajo de los percentiles 10, 50 y 90 son mayores a las esperadas según su talla, la mayoría de los valores se encuentran concentrados por debajo de percentil 50 y prácticamente no hay observaciones por encima del percentil 90. Para el caso de las mediciones por debajo del percentil 10, se esperaba contar con el 10 % de las observaciones, sin embargo se registran valores porcentuales del

26, 32, 36 y 28 respectivamente para cada etapa. Además se constata que por encima del percentil 90, no se registra el 10 % esperado.

Para el caso de los varones, se observa que las mediciones de PFE hasta el percentil 10 según la talla toman valores porcentuales de 15, 26, 21 y 19, respectivamente para cada etapa. Esta situación es trasladada al resto de las franjas percentilares (50 y 90), donde la cantidad de varones con mediciones de PFE también está por encima de los esperado, sin embargo registraron mejor desempeño con respecto a las niñas.

Si se analiza desde el punto de vista geográfico, se puede concluir que las zonas de residencia con mayor proporción de mediciones de capacidad pulmonar por debajo del percentil 10, son en promedio para las 4 etapas de monitoreo, el Centro y Microcentro, mientras que la zona Industrial es la que tiene mejores mediciones en promedio para todo el monitoreo. Por lo tanto, se concluye que residir próximo al molino, no es factor que incida en los niveles de capacidad pulmonar. Al tener en cuenta los totales por etapa, sin desagregar por género, se puede apreciar que más del 26 % de la población analizada, presenta valores de capacidad pulmonar por debajo del percentil 10.

Se pudo concluir que la capacidad pulmonar de la población monitoreada en la ciudad de Artigas se encuentra por debajo a los valores esperados con respecto a una población uruguaya sin patologías respiratorias, por lo cual indica una tendencia a presencia de patologías respiratorias.

Para dar respuesta al segundo objetivo: **"Analizar la exposición a contaminación intradomiciliaria y patologías a cuadros respiratorios de los individuos monitoreados",** se estudió el indicador de patologías respiratorias, calculado a partir de las variables del cuestionario diagnóstico médico y uso de medicamentos. Los participantes con patologías respiratorias que en cada etapa registraron valores de PFE por debajo del percentil 10 superaron el 38 % en todas las etapas, registrando el valor más alto en la tercera etapa, donde el 46 % de las niñas y varones con mediciones de capacidad pulmonar por debajo del percentil 10, declararon presentar

patologías respiratorias.

En cuanto al indicador de exposición, se concluye que los participantes que en cada etapa registraron valores de PFE por debajo del percentil 10 y presentaban exposición a contaminación intradomiciliaria (exposición a humo de tabaco o leña), supera el 20 % para todas las etapas. Este indicador alcanza su máximo en la primera etapa, donde el 30 % de los niños con mediciones de capacidad pulmonar por debajo del percentil 10, declara estar expuesto a humo de tabaco y/o de leña.

Con respecto al tercer objetivo: **"Estudiar la asociación que existe entre la capacidad pulmonar y las variables: zona de residencia y etapa",** se aborda la problemática de los pobladores de la zona industrial de la ciudad de Artigas, quienes declaran sentir molestias respiratorias a raíz de los procesos industriales desarrollados por las plantas arroceras, especialmente en momentos de zafra. Para ello, se propusieron y estimaron un conjunto de modelos, con el fin de explicar la variabilidad del PFE.

En primer lugar, se analizaron cada una de las etapas de monitoreo de forma independiente. Como resultado, se obtuvo que tanto la talla como la edad, fueron significativas para explicar la variabilidad del PFE en todas las etapas de análisis. La zona de residencia, resultó significativa en la primera etapa, momento de post-zafra arrocera. También se destaca, que el sexo resultó significativo para explicar la capacidad pulmonar en todas las etapas a excepción de la tercera. A partir del diagnóstico de los residuos de cada modelo estimado, se reconoce la posible presencia de heterocedasticidad.

En segundo lugar, se plantearon una serie de modelos lineales (bajo supuestos clásicos) para estimar el PFE en las etapas 2, 3 y 4, tomando la primera medición como variable explicativa. El motivo de realizar este análisis se debió a la alta correlación entre los valores de PFE en las distintas etapas con respecto a la primera. Los modelos seleccionados en cada caso, a partir del criterio de información de Akaike, indicaron que el PFE en la etapa inicial resultó significativo para explicar la varia-

bilidad del PFE en las etapas siguientes. Además, el estrato de la escuela resultó significativo en los tres modelos analizados para este escenario. La zona de residencia, fue considerada en el escenario evaluado para la tercera etapa de monitoreo, en la que no había zafra arrocera. El diagnóstico de los residuos, al igual que en el escenario anterior, indica posible presencia de heterocedasticidad en los errores.

Para el tercer escenario de análisis, se consideró el hecho de que los datos obtenidos del monitoreo provienen de una muestra longitudinal. Por lo tanto, se evalúan en conjunto las observaciones de PFE para las etapas 2, 3 y 4 del monitoreo en función del PFE en la etapa 1 y las demás variables de análisis. El modelo resultante, considera las variables PFE en la etapa inicial, la edad, la talla, la zona de residencia y la etapa, todas ellas significativas al 5 %. Al analizar el diagnóstico sobre los residuos para evaluar los supuestos clásicos, se reconoce nuevamente a partir de los gráficos posible heterocedasticidad, la cual se confirma mediante el test de hipótesis NCV.

En el cuarto escenario de análisis, se buscó modelar la heterocedasticidad mediante la aplicación de funciones de varianza. El modelo obtenido, considera las variables PFE inicial, edad, peso, talla, zona de residencia y etapa. La matriz de dispersión de los residuos de Pearson para las tres fases de medición, indican correlación de los residuos entre las etapas, que se confirma mediante el test de hipótesis de Durbin Watson.

En el quinto escenario de análisis, se levanta el supuesto de incorrelación y se añaden, al modelo antes considerado, funciones de estructuras de correlación. En este caso, se implementan estructuras de correlación en serie, adecuadas para modelar datos longitudinales, eligiendo la estructura de correlación de "simetría compuesta" que considera igual coeficiente de correlación para el mismo individuo en diferentes momentos del tiempo. Se realizó el test de hipótesis LR, para comparar los datos obtenidos para el escenario anterior donde se asumía independencia y este escenario donde se considera la correlación. El test resultó significativo, lo que indica la importancia del ajuste de correlación en el modelado de los datos. Además, se observa que los valores de AIC se reducen al considerar la estructura de correlación. Se realizó el

diagnóstico del modelo resultante. Primero, al realizar el gráfico de dispersión de los residuos de Pearson contra los valores ajustados desagregado por etapa de análisis, no se observa patrón que indique posible aleatoriedad de los residuos. Se comprueba que los errores se ajustan a una distribución normal.

Por último se analizó el escenario de modelos lineales con efectos mixtos, donde se incorpora la aleatoriedad a las variables, además de considerar la función de varianzas seleccionada en el cuarto escenario. No fue necesario incorporar la estructura de correlación seleccionada en el quinto escenario, dado que los modelos lineales de efectos mixtos tienen implícita la correlación entre los individuos. El modelo estimado, presenta mayor AIC en comparación con el modelo anterior, lo que es razonable dado que este modelo estima mayor cantidad de parámetros, en cuanto al error cuadrático medio se obtienen mejores resultados en el escenario VI. Por lo tanto, se opta por considerar el modelo lineal de efectos mixtos como el modelo que mejor ajusta la capacidad pulmonar medida longitudinalmente.

En general:

Se puede confirmar que los niños de la ciudad de Artigas tienen mediciones de capacidad pulmonar con respecto a la talla por debajo de los valores registrados a nivel nacional para niños sin patologías.

En cuanto a las etapas de monitoreo comprendidas en períodos de zafra arrocera se concluye que los niños no presentan valores menores en comparación a períodos fuera de zafra. En la etapa 3 las niñas y niños registraron en promedio mediciones de su capacidad pulmonar 3,04 l/min por encima con respecto a la etapa 2 y en la etapa 4 incrementaron sus mediciones promedio de capacidad pulmonar en 10,66 l/min respecto a la etapa 2, considerando las restantes variables constantes (talla, peso, edad, zona de residencia).

Lo mismo sucede con las zonas de residencia, los niños residentes en la zona indus-

trial, registraron mejores mediciones que los niños residentes en otras zonas, por lo que tampoco puede verificarse que la zona de residencia próxima a los molinos impacte en los valores de PFE.

Cuando se realiza el análisis desde el punto de vista de un estudio longitudinal, los indicadores de patologías y exposición, no resultan significativos para explicar el PFE. De todas formas, el análisis de los niños que presentan valores de PFE por debajo del percentil 10 y además registran patologías respiratorias supera los valores esperados.

Bibliografía

[1] Bates, D., Mächler, M., Bolker, B., y Walker, S. (2014). Fitting Linear Mixed-Effects Models using lme4. *arXiv:1406.5823 [stat]*. arXiv: 1406.5823.

[2] Capano, A., Saráchaga, M. J., Estol, P., Orsi, S., Lapides, C., y Ferreira, N. (2007). Pico de flujo espiratorio en niños uruguayos sin enfermedad, de 3 a 13 años. *Revista chilena de pediatría*, 78(4).

[3] Casarone (2010). Informe Ambiental Resúmen. Technical report, Uruguay.

[4] Cavalleri, F. (2018). Modelos mixtos ¿mlne o ml4?

[5] Gałecki, A. y Burzykowski, T. (2013). *Linear Mixed-Effects Models Using R*. Springer Texts in Statistics. Springer New York, New York, NY.

[Grolemund y Wickham] Grolemund, G. y Wickham, H. *R for Data Science*.

[7] Gujarati, D. y Porter, D. (2009). *Econometria*. USA, 5ta edicin.

[8] Henry, L., Wickham, H., y RStudio (2018). purrr: Functional Programming Tools.

[9] INE (2004). Unidades Geoestadísticas (UGeo) - Uruguay.

[10] Lauritsen, Christiansen, J., y T. (1999). Epidata manager intro.pdf.

[11] Marroig Baldini, M. A. (2017). Trayectoria nutricional y desempeño escolar. p. 135.

[12] Massa Mandagaran, F. F. (2015). Efecto de valores faltantes en estudios longitudinales en adultos mayores. p. 92.

[13] MSP (2010). Informe final de la intervención ambiental en salud realizada en la ciudad de Tacuarembó, durante los meses de abril y noviembre 2009. Technical report, Montevideo.

[14] Muller, K. y Wickham, H. (2018). tibble: Simple Data Frames.

[15] OMS (2005). Guías de calidad de aire.

[16] Pinheiro, J., Bates, D., DebRoy, S., Sarkar, D., y R Core Team (2018). nlme: Linear and Nonlinear Mixed Effects Models.

[17] R Core Team (2013). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria.

[18] RStudio Team (2015). *RStudio: Integrated Development Environment for R*. RStudio, Inc., Boston, MA.

[19] Sarkar, D. (2008). *Lattice: Multivariate Data Visualization with R*. Springer, New York.

[20] Sarndal, Swensson, B., y Wretman, J. (1992). *Model Assisted Survey Sampling*. Springer New York.

[21] Uruguay_XXI (2015). Informe Anual Comercio Exterior. Technical report, Uruguay_XXI, Uruguay.

[22] Venables, W. N. y Ripley, B. D. (2002). *Modern Applied Statistics with S*. Springer, New York, fourth edicin.

[23] Wickham, H. (2007). Reshaping data with the reshape package.

[24] Wickham, H. (2016). *ggplot2: Elegant Graphics for Data Analysis*. Springer-Verlag New York.

[25] Wickham, H. (2017). tidyverse: Easily Install and Load the 'Tidyverse'.

[26] Wickham, H. (2018a). forcats: Tools for Working with Categorical Variables (Factors).

[27] Wickham, H. (2018b). stringr: Simple, Consistent Wrappers for Common String Operations.

[28] Wickham, H. y Bryan, J. (2018). readxl: Read Excel Files.

[29] Wickham, H., Franois, R., Henry, L., y Mller, K. (2018). dplyr: A Grammar of Data Manipulation.

[Wickham y Henry] Wickham, H. y Henry, L. tidyr: Easily Tidy Data with 'spread()' and 'gather()' Functions.

[31] Wickham, H., Hester, J., y Francois, R. (2017). readr: Read Rectangular Text Data.

Apéndice A

Apéndice de consideraciones generales

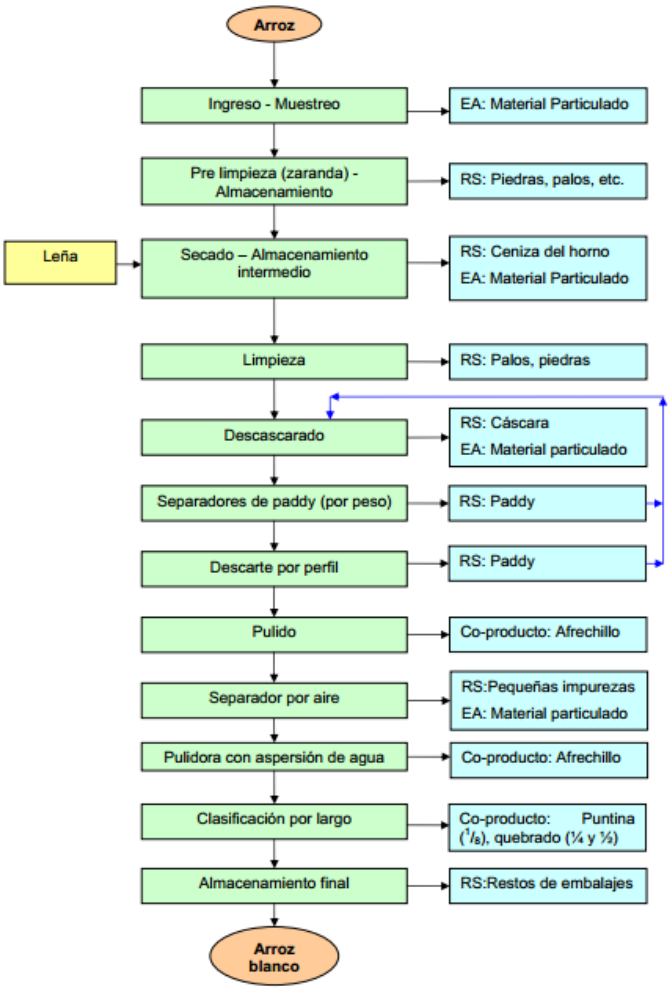


Figura A.1: Proceso del arroz

Mayo 2015

MONITOREO DE LA SALUD RESPIRATORIA EN POBLACIÓN INFANTIL DE LA CIUDAD DE ARTIGAS 2015-2016

CONSENTIMIENTO INFORMADO

Sres. Padres:

Les informamos que el Ministerio de Salud Pública realizará durante el período 2015-2016, estudios de la funcionalidad respiratoria en niños de la ciudad de Artigas.
Para ello se propone la participación de la población infantil que incluye los preescolares de los niveles 4 y 5 y los escolares que estén cursando 1º y 2º año de Primaria.

Este estudio cuenta con el auspicio de la Administración Nacional de Educación Pública.

El estudio que se implementará se repetirá cada 6 meses en 3 oportunidades durante el período 2015-2016 y se llevará a cabo en los centros de enseñanza de los niños

El mismo consiste en soplar a través de un tubo de cartón descartable e individual, y se registran los valores obtenidos, los cuales se relacionan con la talla y el peso de cada niño.

Por este motivo previo al estudio se medirá el peso y la talla de los niños y las niñas, actividad que se realiza sin zapatos y en posición de pie.

Es necesario que previo al estudio ustedes completen un formulario sobre los antecedentes personales y familiares. El mismo tiene que presentarse el día del estudio.
Si al momento de la realización del estudio se constatara alguna alteración del mismo, el paciente será derivado en el momento a su medico tratante.
No se administrará ningún medicamento, siendo un estudio carente de riesgos.

El personal de salud responsable de los estudios son: la Dra. Adriana Sosa pediatra, la Dra. Susana Rodríguez y la Licenciada Graciana Barboza de la División Salud Ambiental y Ocupacional del Ministerio de Salud Pública.

Tengan la garantía de que en todo momento se respetarán los derechos e identidad de los niños y niñas. No se realizará ningún estudio sin el consentimiento de los padres, tutores o en los casos que los propios niños se nieguen.

El presente estudio no tiene costo para los participantes, ni tampoco ningún beneficio económico.

Los resultados finales del monitoreo serán compartidos con los padres en una instancia pública al final del estudio.

Si uds. están de acuerdo con la realización del mismo, es requisito indispensable que lo autorice:

Autorizo a los citados Drs. a realizarle a mi hijo/hija…..……

los estudios de funcionalidad respiratoria en las condiciones explicadas anteriormente.

Lugar……..Fecha……Escuela……...

Firma:……Aclaración: ……..Cl:……..

NOMBRE DEL NIÑO/NIÑA.....

LOCALIDAD: .....FECHA de Nacimiento: / /

PESO .....TALLA..... CEDULA.....
(Datos a completar por el médico que examine a su hija/ hijo el día del estudio)

PROTOCOLO A COMPLETAR POR LOS PADRES o TUTORES
Es necesario llenar todos los items. Complete los espacios de las preguntas con una cruz. En caso de duda consulte a un familiar o a su Médico o Pediatra.
Toda la información es de carácter confidencial.

ESCUELA No..... TURNO..... Año que cursa.....

Domicilio:..... ..... Barrio:..... Tel. o cel. .....

ANTECEDENTES SOBRE AFECCIONES PADECIDAS POR EL niño / niña

1. El Médico le diagnosticó Asma, Bronquitis asmática, Broncoespasmo?: SI NO
2. Cuando ?.....
3. **En los últimos 12 meses** Tos seca por las noches:..... SI NO
4. Cuántas veces en los últimos 12 meses.....
5. En qué meses.....
6. **En los últimos 12 meses** el ejercicio le da tos, le "cierra el pecho" o produce broncoespasmo? SI NO
7. Ha presentado Congestión pulmonar o Neumopatía Aguda en los últimos 12 meses? ..... SI NO
8. **En los últimos 12 meses** ha presentado Bronquitis?: ..... SI NO
9. Toma alguna medicación en este momento?..... ..... SI NO
10. Cuál?.....
11. **En los últimos 12 meses** ha presentado síntomas de irritación de ojos, nariz o boca?..... SI NO
12. Cuántas veces?.....
13. En qué mes?.....

ANTECEDENTES AMBIENTALES:
Alguna fuma dentro de la casa?

		SI	NO
14	Madre		
15	Padre		
16	Otro		

SIGUE AL DORSO


Que utiliza para cocinar y calentar la vivienda?

	Cocina	Calefacción
17. Supergas		
18. Electricidad		
19. Leña		
20. Otros		

A su juicio: existe contaminación atmosférica provocada por humo, gases o polvillo de automotores, fábricas, industrias, caminos de tierra?

	SI	NO
21 Automotores		
22 Fábricas		
23 Industrias		
24 Caminería		
25 Otros		

26. En qué se basa?.....
.....
.....
.....
.....

27.**OBSERVACIONES:**.....
.....
.....

28. Nombre de la persona que llenó el protocolo.....

29. Podríamos comunicarnos con ud. por teléfono si tuviéramos alguna duda acerca de los datos de este Protocolo?

SI	NO

30. Número de teléfono o celular para contactarlo: .....

Fecha de realización de pico flujo / /

Apéndice B

Apéndice de marco teórico

Estrategia de muestra y diseño muestral

El esquema de muestreo es una combinación específica del tipo y la modalidad de muestreo y del número de etapas de selección. El tipo de muestreo, determina su característica pobrabilística. El muestreo probabilístico, asigna a cada elemento de la población de estudio una probabilidad conocida y diferente de cero de ser seleccionado en la muestra.

Para la aplicación de un diseño de muestreo es esperable contar con un marco muestral que permita identificar todos los elementos de la población, seleccionar una muestra y localizar sus unidades en campo.

El factor de expansión es un concepto relacionado con la probabilidad de selección y se interpreta como la cantidad de unidades en la población que representa una unidad en la muestra, ya sean personas, viviendas, áreas económicas o agrícolas, etcétera.

Tanto la población a estudiar como la amplitud del área geográfica a cubrir, se consideran condicionantes para el *diseño muestral*. Asimismo, la cantidad de variables a estudiar, la frecuencia con la que la característica se presenta en la población y la amplitud de los valores, son factores que inciden en la determinación del *tamaño*

de muestra y permiten obtener la representación de la muestra en las estimaciones. Existen aspectos asociados con el presupuesto y duración del evento que influyen también en el cálculo del tamaño de muestra. Por otra parte, el *esquema de selección* de la muestra de varias etapas, puede determinarse por las características del marco muestral, los recursos humanos y herramientas disponibles.

Si se opta por elegir un muestreo probabilístico en la perspectiva del costo-beneficio, el presupuesto aumenta pero permite conocer el grado de precisión de las estimaciones y obtener conclusiones que se generalicen hacia toda la población.

El diseño de muestra que se utiliza en la presente investigación es el Muestreo Probabilístico Sistemático, que se desarrolla a continuación.

Muestreo Sistemático

El muestreo sistemático, es una técnica comprendida dentro de la categoría de muestreos probabilísticos. Consiste en primer lugar, en escoger un individuo inicial de forma aleatoria entre los elementos de la población, luego se selecciona cada enésimo individuo disponible en el marco muestral, hasta obtener el tamaño de muestra deseado. Los resultados obtenidos son representativos de la población, de igual manera que con el muestreo aleatorio simple.

Forma básica

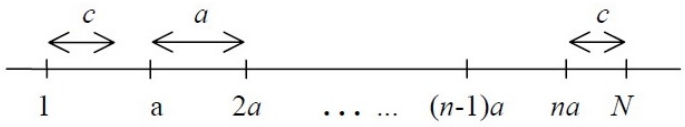
Dada la población $\mathbf{U} = \{1, 2, \ldots, N\}$ se considera $\mathbf{a} \in \mathbf{N}$, fijo, llamado intervalo de muestreo y sea $n = \left[\frac{N}{a}\right] \Rightarrow N = na + c$, donde $0 \leq c < a$ (donde [.] significa parte entera).

Sea $\mathbf{r}$ una variable aleatoria uniforme discreta en $\{1, 2, \ldots, a\}, r \sim Uni(1, 2, \ldots, a)$, llamada "arranque aleatorio". Una vez que se observa un valor de $\mathbf{r}$, la muestra queda conformada por $S = \{k : k = r + (j-1)a \leq N,\ j = 1, 2, \ldots, n_s\}$, donde $\mathbf{n}_s$

es el tamaño de muestra, aleatorio.

$$n_s = \begin{cases} n+1 & si\ 0 < r \leq c \\ n & si\ c < r \leq a \end{cases} \tag{B.1}$$

El conjunto de las muestras posibles viene dado por $S_Y\{S_1, S_2, \ldots, S_r, \ldots, S_a\}$ donde $S_r = \{k : k = r + (j-1)a \leq N\}$ con $r = 1, 2, \ldots, a$. Se cumple que, $S_i \cap S_j = \emptyset,\ \forall i \neq j,\ \bigcup_{i=1}^{a}[S_i = U]$.



$$P(n_S = n+1) = P(r \leq c) = \frac{c}{a}$$

$$P(n_S = n) = P(r > c) = 1 - \frac{c}{a}$$

Existen a muestras posibles, con probabilidad $\frac{1}{a}$ de ser seleccionada. Así,

$$p(s) = \begin{cases} a^{-1} & \text{si } s \in S_{SY} \\ 0 & \text{en otro caso} \end{cases}$$

Esto implica:

$$\pi_k = P(k \in S) = a^{-1}\ \forall k \in U$$

$$\pi_{kl} = P(k\ \text{y}l \in S) = \begin{cases} a^{-1} & \forall k \in U \\ 0 & \text{en otro caso} \end{cases}$$

Es útil representar la población ordenada según las distintas muestras posibles de acuerdo a la siguiente Figura:

Muestra	S_1	.....	S_r	.....	S_a
U	y_1	.....	y_r	.....	y_a
	y_{1+a}	.....	y_{r+a}	.....	y_{2a}
	.....	.....	.....	.....	.....
	$y_{1+(n-1)a}$	.....	$y_{r+(n-1)a}$	.....	y_{na}
Total	t_{S_1}	.....	t_{Sr}	.....	t_{Sa}
Media	$\bar{y}_{S_1}$	.....	$\bar{y}_{Sr}$	.....	$\bar{y}_{Sa}$

Si se supone $N = na$ implica $n_s = n\ \forall s \in S_{SY}$ y el tamaño de muestra es fijo.

Procedimiento simple:

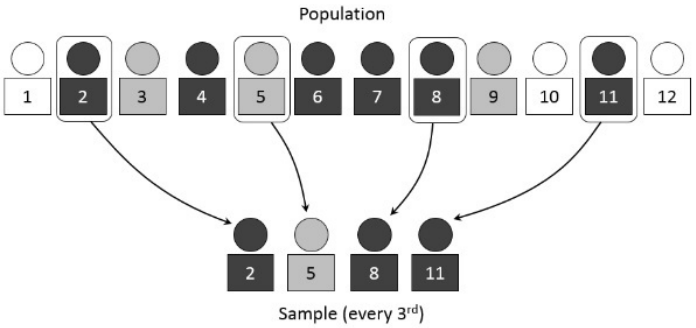
1. Se elabora una lista ordenada de los N individuos de la población U, de esta forma, se obtiene el marco muestral.

2. Se divide el marco muestral en n fragmentos, con n tamaño de muestra y a=N/n tamaño del intervalo.

3. Se escoge el individuo inicial, a partir de un número aleatorio r menor o igual al tamaño del intervalo. Este individuo, es entonces, el primer elemento de la muestra.

4. Se seleccionan los *n-1* elementos restantes, mediante una sucesión aritmética a partir del primer elemento antes seleccionado:

$$r,\ r+a,\ r+2a,\ r+3a, ...,\ r+(n-1)a \tag{B.2}$$



Apéndice C

Apéndice de trabajo de campo y caracterización de la muestra

	Etapa 1		Etapa 2		Etapa 3		Etapa 4	
	n	%	n	%	n	%	n	%
1. Cerros	13	46 %	13	42 %	11	37 %	11	31 %
2. Industrial	10	20 %	17	31 %	19	35 %	14	25 %
3. Periferia	29	28 %	35	29 %	41	34 %	29	24 %
4. Centro	6	16 %	15	29 %	19	43 %	18	38 %
5. Microcentro	4	29 %	7	47 %	7	41 %	7	39 %
Percentil 10	62	26 %	87	32 %	97	36 %	79	28 %
1. Cerros	22	79 %	28	90 %	25	83 %	26	74 %
2. Industrial	38	75 %	50	91 %	46	84 %	45	79 %
3. Periferia	94	90 %	102	86 %	103	84 %	98	80 %
4. Centro	30	81 %	44	85 %	35	80 %	40	83 %
5. Microcentro	11	79 %	13	87 %	12	71 %	14	78 %
Percentil 50	195	83 %	237	87 %	221	82 %	223	79 %
1. Cerros	28	100 %	31	100 %	30	100 %	35	100 %
2. Industrial	51	100 %	55	100 %	55	100 %	56	98 %
3. Periferia	103	98 %	119	100 %	122	100 %	123	100 %
4. Centro	37	100 %	52	100 %	44	100 %	48	100 %
5. Microcentro	14	100 %	15	100 %	16	94 %	18	100 %
Percentil 90	233	99 %	272	100 %	267	100 %	280	100 %

Tabla C.1: Cantidad y porcentaje de niñas por etapa según zona de residencia y valores acumulados hasta percentil 10, 50 y 90.

	Etapa 1		Etapa 2		Etapa 3		Etapa 4	
	n	%	n	%	n	%	n	%
1. Cerros	7	24 %	8	22 %	5	14 %	7	17 %
2. Industrial	2	4 %	13	20 %	7	13 %	13	20 %
3. Periferia	19	19 %	36	30 %	30	26 %	26	22 %
4. Centro	8	14 %	17	26 %	18	28 %	10	15 %
5. Microcentro	2	13 %	2	13 %	3	14 %	4	17 %
Percentil 10	38	15 %	76	25 %	63	22 %	60	19 %
1. Cerros	18	62 %	21	57 %	25	68 %	22	54 %
2. Industrial	19	37 %	41	63 %	33	60 %	38	58 %
3. Periferia	61	60 %	90	74 %	85	75 %	83	72 %
4. Centro	33	57 %	44	67 %	48	75 %	44	65 %
5. Microcentro	6	40 %	6	38 %	13	59 %	12	50 %
Percentil 50	137	54 %	202	66 %	204	70 %	199	63 %
1. Cerros	27	93 %	35	95 %	37	100 %	39	95 %
2. Industrial	47	90 %	61	94 %	53	96 %	63	95 %
3. Periferia	101	99 %	118	98 %	113	99 %	114	98 %
4. Centro	54	93 %	65	98 %	63	98 %	68	100 %
5. Microcentro	14	93 %	16	100 %	21	95 %	23	96 %
Percentil 90	243	95 %	295	97 %	287	98 %	307	97 %

Tabla C.2: Cantidad y porcentaje de varones por etapa según zona de residencia y valores acumulados hasta percentil 10, 50 y 90.

APÉNDICE C. APÉNDICE DE TRABAJO DE CAMPO Y CARACTERIZACIÓN DE LA MUESTRA

Apéndice D

Apéndice de aplicación de Modelos Mixtos

	ML_I_E1	*ML_I_E2*	*ML_I_E3*	ML_I_E4
Error estándar residual	28	29	29	31
Grados de libertad	431	492	465	491
R^2	0,53	0,50	0,45	0,48
R^2 ajustado	0,52	0,49	0,45	0,47
Estadístico F	69	97	127	77
p-valor	$2{,}20e^{-16}$	$2{,}20e^{-16}$	$2{,}20e^{-16}$	$2{,}20e^{-16}$
qf(0.95)	3,02	3,01	3,02	3,01
AIC	4176	4762	4494	4851

Tabla D.1: Escenario I - Resultado de los modelos estimados.

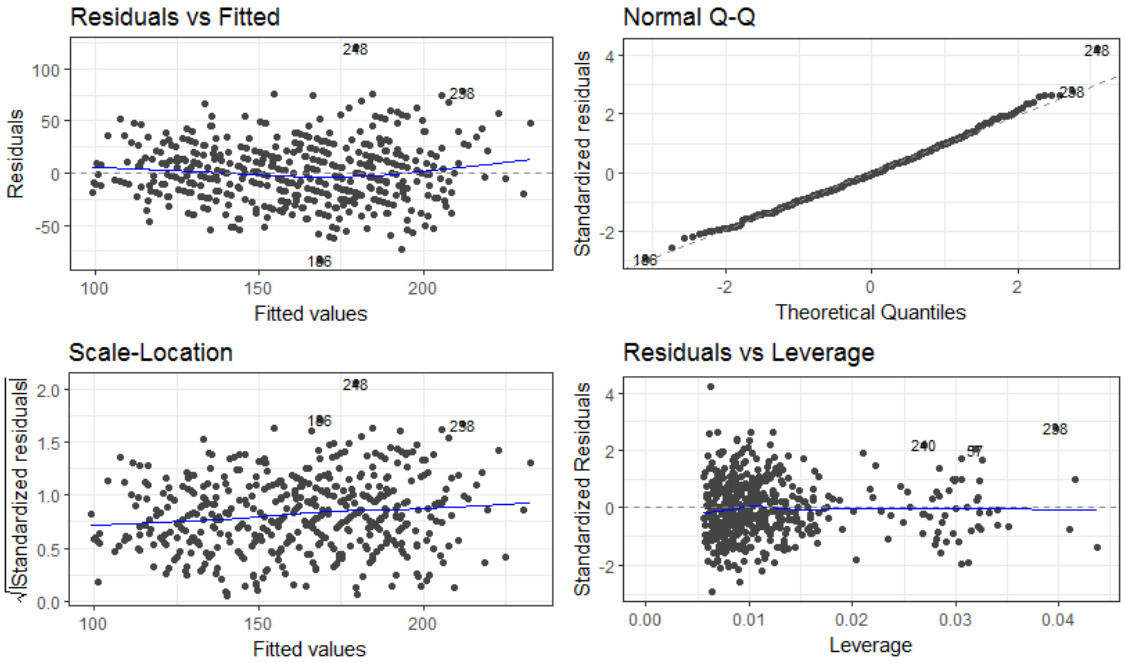


Figura D.1: Escenario I - Residuos Etapa 2.

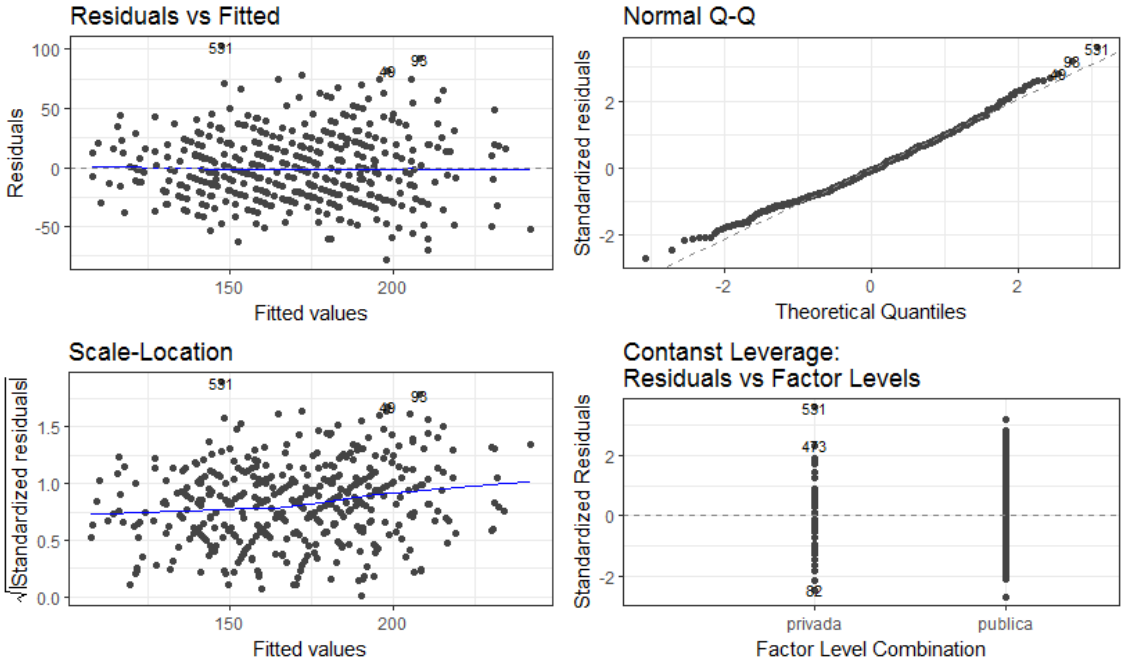


Figura D.2: Escenario I - Residuos Etapa 3.

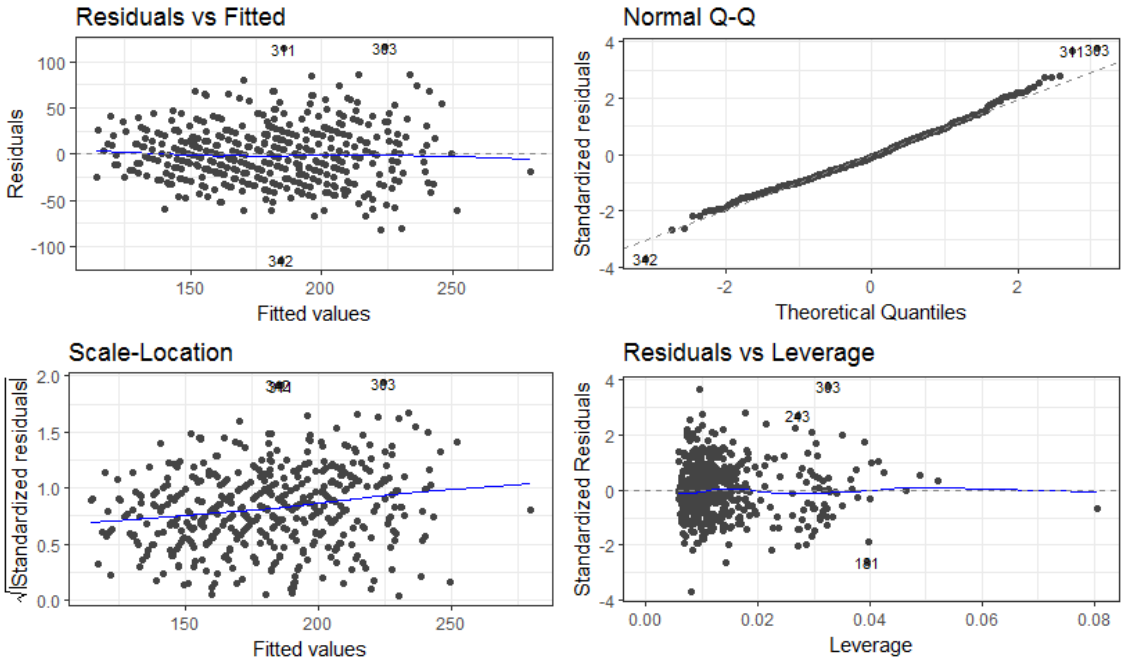


Figura D.3: Escenario I - Residuos Etapa 4.

	ML_II_E2	*ML_II_E3*	*ML_II_E4*
Error residual estándar	26	25	28
Grados de libertad	393	370	394
R^2	0,60	0,61	0,59
R^2 ajustado	0,59	0,60	0,58
Estadístico F	145	83	113
p-valor	$2{,}2e^{-16}$	$2{,}2e^{-16}$	$2{,}2e^{-16}$
qf(0.95)	3,02	3,02	3,02
AIC	3730	3503	3816

Tabla D.2: Escenario II - Resultado de los modelos estimados.

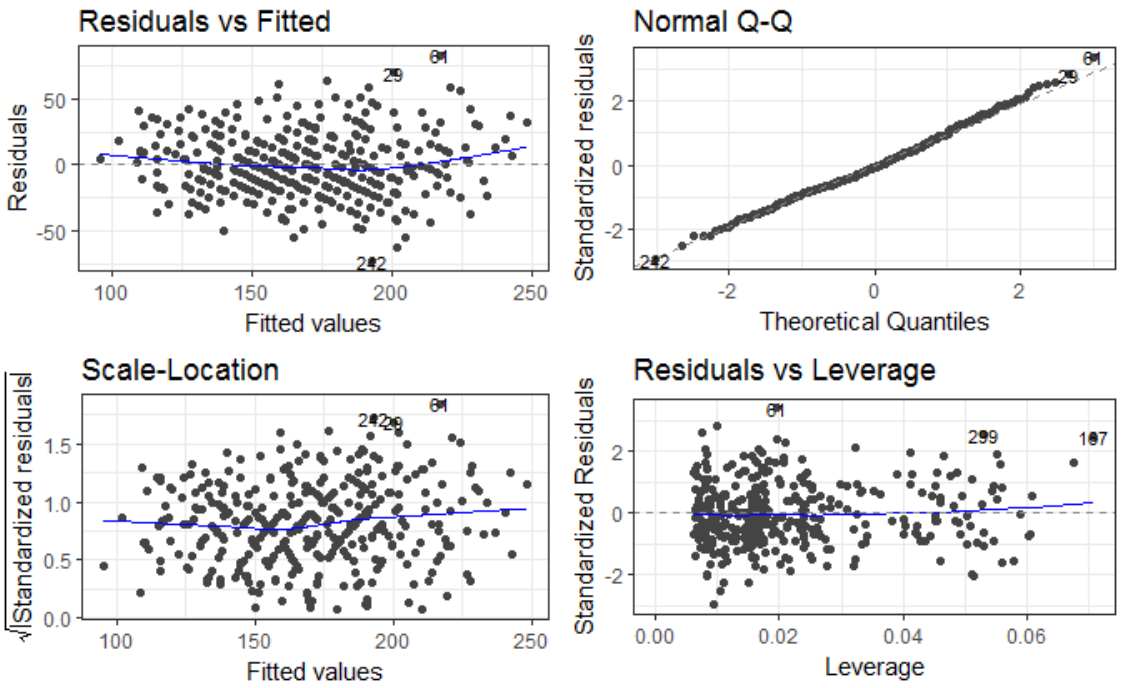


Figura D.4: Escenario II - Residuos Etapa 3.

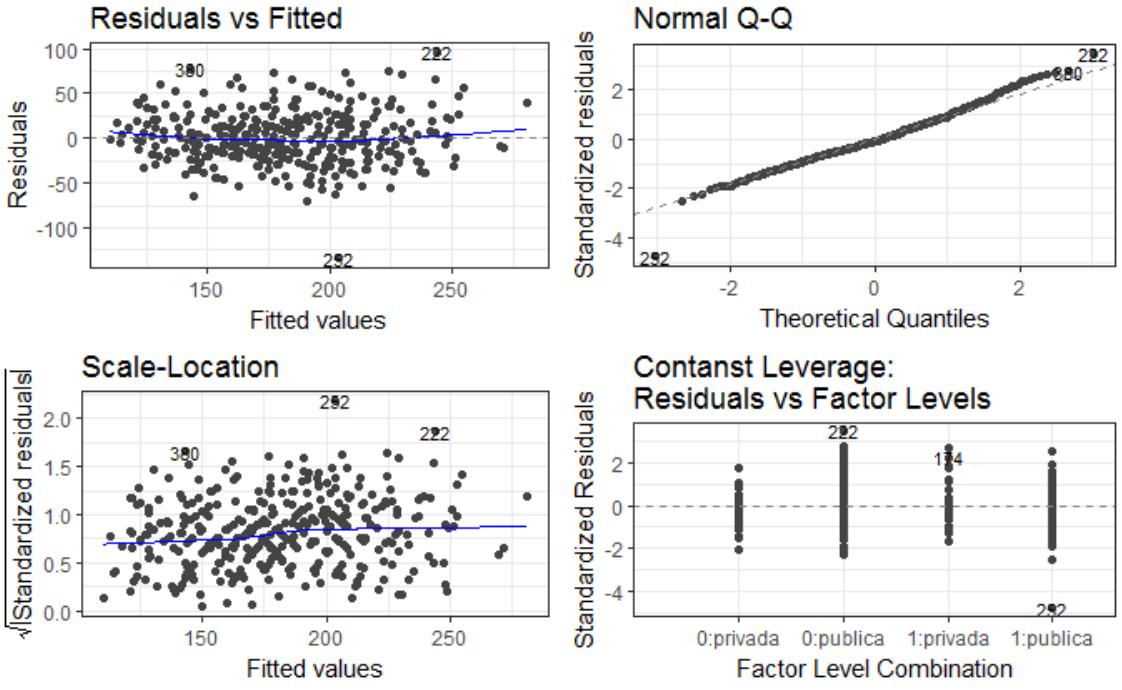


Figura D.5: Escenario II - Residuos Etapa 4.

	ML_III
Error residual estándar	27
Grados de libertad	1302
R^2	0,59
R^2 ajustado	0,59
Estadístico F	188
p-valor	$2{,}2e^{-16}$
qf(0.95)	3,00
AIC	12374

Tabla D.3: Escenario III - Resultado del modelo estimado.