

Memoria tesis

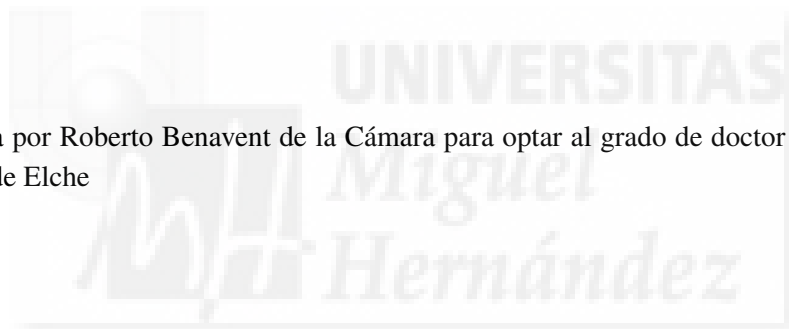
Título: Modelos de área lineales multivariantes
Autor: Roberto Benavent de la Cámara

Director: Domingo Morales González
Departamento de Estadística, Matemáticas e Informática
Universidad Miguel Hernández



Título: Modelos de área lineales multivariantes
Autor: Roberto Benavent de la Cámara

Memoria presentada por Roberto Benavent de la Cámara para optar al grado de doctor por la Universidad Miguel Hernández de Elche

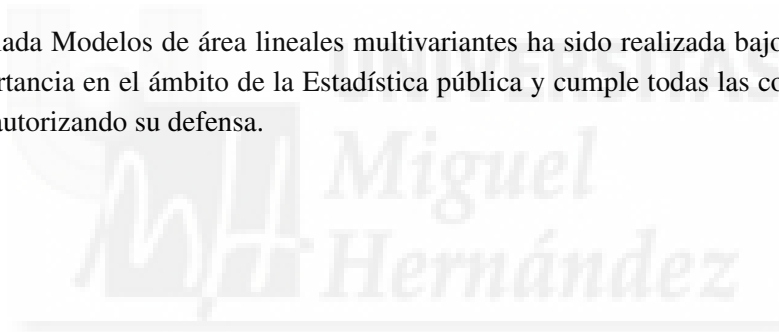


Director: Domingo Morales González

D. Domingo Morales González catedrático de Estadística e Investigación Operativa del departamento de Estadística, Matemáticas e Informática en la Universidad Miguel Hernández de Elche

CERTIFICA

que la memoria titulada Modelos de área lineales multivariantes ha sido realizada bajo mi dirección; trata de un tema de importancia en el ámbito de la Estadística pública y cumple todas las condiciones exigibles para ser defendida, autorizando su defensa.



Para que así conste, firmo el presente certificado en Elche a fecha del día

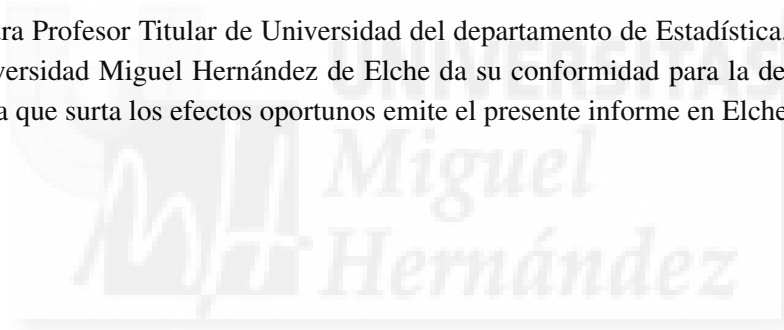
Domingo Morales González

Título: Modelos de área lineales multivariantes

Autor: Roberto Benavent de la Cámara

Dpto.: Estadística, Matemáticas e Informática

D. José Valero Cuadra Profesor Titular de Universidad del departamento de Estadística, Matemáticas e Informática en la Universidad Miguel Hernández de Elche da su conformidad para la defensa pública de la tesis doctoral. Y para que surta los efectos oportunos emite el presente informe en Elche fecha del día



Para que así conste,

José Valero Cuadra

Índice general

Prólogo	5
1. Introducción	7
1.1. Antecedentes	7
1.2. Marco Teórico	8
1.3. Estado del arte	9
1.4. Resumen	11
2. Modelos de área lineales multivariantes	13
2.1. Introducción	13
2.2. Definición del modelo	14
2.3. Máxima verosimilitud residual	17
2.4. Cálculos matriciales	18
2.5. Matriz de errores cuadráticos medios de los EBLUP	19
2.5.1. El caso univariante	26
2.6. Estimación bootstrap del MSE	27
3. Modelo diagonal	31
3.1. Definición del modelo	31
3.2. Experimentos de simulación	32
3.2.1. Experimento de simulación 1	33
3.2.2. Experimento de simulación 2	39
3.2.3. Experimento de simulación 3	44

4. Modelo AR(1)	47
4.1. Definición del modelo	47
4.2. Experimentos de simulación	48
4.2.1. Experimento de simulación 1	50
4.2.2. Experimento de simulación 2	56
4.2.3. Experimento de simulación 3	61
4.2.4. Experimento de simulación 4	64
5. Modelo AR(1) heterocedástico	67
5.1. Definición del modelo	67
5.2. Experimentos de simulación	70
5.2.1. Experimento de simulación 1	71
5.2.2. Experimento de simulación 2	78
5.2.3. Experimento de simulación 3	84
6. Estimación bivalente de indicadores de pobreza	89
6.1. Datos y modelo	89
6.2. Estimación basada en modelos de área (enfoque multivariante)	93
6.2.1. Modelo con varianza de los efectos diagonal	93
6.2.2. Modelo con varianza de los efectos AR(1)	98
6.2.3. Modelo con varianza de los efectos AR(1) heterocedástico	102
6.2.4. Conclusiones	108
6.3. Estimación basada en modelos de área (enfoque temporal)	113
6.3.1. Modelo con varianza de los efectos diagonal	113
6.3.2. Modelo con varianza de los efectos AR(1)	117
6.3.3. Modelo con varianza de los efectos AR(1) heterocedástico	121
6.3.4. Conclusiones	126
7. Conclusiones generales	129
7.1. Conclusiones generales	129
7.2. Comparaciones	130
7.3. Líneas futuras de investigación	132
A. Modelo Fay-Herriot univariante	133
B. Modelo para el experimento de simulación	137
C. Código R para realizar estimaciones	139

Prólogo

Desde los primeros días en los estudios universitarios que cursé considere las distintas asignaturas como una tarea a cumplir que iba más allá de superar la prueba correspondiente (a menudo pienso que al final de la enseñanza media ya practicaba lo anterior); por ello, es normal que acabara planteándome en serio el dedicarme, de alguna de las formas posibles, a la tarea investigadora. Resulta un tanto difícil fijar los motivos concretos que hicieron pensar en una colección de temas posibles para investigar, por otra parte estoy convencido de que los libros que más me influyeron fueron los siguientes:

1. Metodología de la programación, Luis Joyanes Aguilar
2. Investigación operativa, programación lineal y aplicaciones, Sixto Ríos Insúa
3. Probabilidad y Estadística, Morris H. DeGroot
4. Programación lineal y no lineal, David E. Luenberger
5. Análisis de series temporales, Ezequiel Uriel
6. Análisis Matemático 2, Enrique Linés Escardó

Después de observar los textos anteriores, se comprende perfectamente que pensara en temas afines al estudio de algoritmos, a temas referentes a programación entera y a temas que relacionan el análisis funcional con el cálculo de probabilidades. Las circunstancias personales hicieron que comenzara a dedicarme al estudio de las curvas del tipo Peano y también al estudio de la convexidad en espacios de Banach. La distinción entre teoría y práctica nunca he llegado a hacerla de forma precisa; por poner un ejemplo de la imprecisión que acabo de nombrar, comentaré que el estudio de los espacios L_1 y L_2 me hicieron ver el planteamiento de nuevos problemas prácticos de estimación óptima de parámetros, y, por otra parte, algunas

listas de problemas prácticos me hicieron ver conceptos nuevos que motivaron nuevos teoremas. Tras un proceso considerable de tiempo en el que me adapté a los nuevos conocimientos conseguí llegar a algunos resultados y también accedí al certificado de docencia en tercer ciclo titulado: «Localización de dominios».

Las circunstancias cambiaron mucho y me vi obligado a buscar una línea de investigación distinta, llegué al estudio de la estimación en áreas pequeñas gracias al profesor Domingo Morales que me recomendó y facilitó el material necesario para que me pusiera al corriente de los temas que estudian, y, con ello sentí que regresaba a temas estudiados con anterioridad, ya que era fundamental el conocimiento de lenguajes de programación y la implementación correcta de programas, y también los modelos ARIMA que había leído en el texto de Ezequiel Uriel y que he citado anteriormente.

Desde los inicios pude aplicar conocimientos adquiridos con anterioridad como ya he comentado en el párrafo precedente, y, además pude conocer técnicas de estimación más complejas que las que estudié durante la licenciatura de estadística. En todo el periodo transcurrido en la elaboración del presente estudio tuve acceso a los diversos materiales necesarios como artículos de investigación, libros sobre el tema, herramientas informáticas, también tuve la ocasión de asistir al congreso sobre estimación en áreas pequeñas que se celebró en la ciudad de Elche en julio de 2009, también fueron muy importantes las reuniones con el profesor Domingo Morales en las cuales me explicaba detalles sobre la elaboración de la tesis tanto teóricos como prácticos.

Por todo ello, quisiera mostrar mi agradecimiento, desde este prólogo, al profesor Domingo Morales y también a todos los miembros del departamento que trabajan en estimación en áreas pequeñas, recuerdo sobre todo el asesoramiento informático que recibí de Agustín Pérez, María Dolores Esteban y María Chiara Pagliarella.

Roberto Benavent
Junio 2015

1

Introducción

1.1. Antecedentes

Los estados modernos tienen la obligación de poner en marcha estrategias globales con planes a largo plazo para lograr un impacto decisivo en la erradicación de la pobreza en sus territorios. Un tema de gran interés es ahora, por tanto, la estimación y difusión de los indicadores de pobreza, desigualdad y condiciones de vida. En España, tales indicadores pueden ayudar mucho en la vigilancia de las condiciones de vida y en la orientación de la aplicación de políticas encaminadas a mejorarlas en las provincias y comarcas de España. Dados los crecientes problemas sociales, demográficos y económicos, la comunidad de investigadores, responsables políticos y profesionales ponen gran énfasis en la elaboración de indicadores eficientes, eficaces y fiables y en la recogida de datos de alta calidad sobre las condiciones de vida, no sólo a nivel nacional sino también en regional y en los niveles geográficos inferiores.

El Instituto Nacional de Estadística ha diseñado la Encuesta de Condiciones de Vida para obtener estimaciones directas fiables a nivel de comunidad autónoma. Esta encuesta no permite generar estimaciones de ámbitos territoriales inferiores. Por tanto, para generar estimaciones a nivel de comarcas, sin recurrir a operaciones directas de enumeración por muestreo, se hace necesario recurrir a metodologías estadísticas denominadas de estimación en pequeñas áreas.

La demanda de estadísticas oficiales con un gran detalle en la desagregación, tanto en el campo de la estadística económica como en el de la estadística social y laboral, no deja de crecer. En consecuencia, la necesidad de disponer sistemáticamente de datos publicados para dominios pequeños, se ha venido consolidando en los últimos años entre los objetivos de los sistemas de estadística oficiales. Los problemas de la estimación en pequeños dominios surgen por el aumento en los costes y en la complejidad de los diseños muestrales que aspiren a alcanzar cotas aceptables de calidad de las estimaciones en todas las áreas o dominios de interés para los usuarios, lo que indirectamente puede afectar negativamente en la calidad de las estimaciones para dominios superiores.

Los límites, por razones de coste, para la ampliación sin restricciones de los tamaños muestrales en todos los dominios de interés, deben ser interpretados en un sentido amplio: bajo el punto de vista de la recogida

y producción de datos, claro está, pero también bajo el punto de vista de la carga de respuesta a las unidades a contactar en la encuesta.

El aumento de tamaños muestrales para mejorar la eficiencia en dominios pequeños debe tener en cuenta ambos costes, sin olvidar otras pérdidas de calidad debidas a los plazos de obtención de resultados y al impacto de determinados errores ajenos al muestreo (falta de respuesta, errores de medida, efectos entrevistador, etc.) de consecuencias más negativas cuanto mayores son los tamaños muestrales. Las técnicas estadísticas de estimación en áreas pequeñas dan una respuesta adecuada que evita el aumento indiscriminado de los tamaños muestrales. Sin embargo, no tienen por objetivo el bajar una encuesta entera de un nivel de agregación a otro más bajo (por ejemplo, de departamento a municipio), dado que la utilización de modelos se hace imprescindible.

La estimación en áreas pequeñas es una parcela de la estadística que trata el problema de estimar parámetros de subconjuntos de la población (llamados áreas pequeñas o dominios) a partir de muestras e información auxiliar. Debido a la falta de precisión de los estimadores directos de parámetros de áreas pequeñas, se han desarrollado nuevos procedimientos de estimación. El libro de Rao (2003) y los artículos de revisión de Ghosh and Rao (1994), Rao (1999), Pfeffermann (2002), Jiang y Lahiri (2006) y Pfeffermann (2013) dan una descripción detallada de esta teoría.

1.2. Marco Teórico

El muestreo estadístico, en contraposición con los censos, permite obtener información sobre materias muy dispares con un coste reducido. El muestreo se utiliza no solamente para la obtención de estimaciones en la población completa, sino para estimar parámetros en una variedad de subpoblaciones dominios). Los dominios se definen generalmente como áreas geográficas o grupos socioeconómicos. Ejemplos de dominios geográficos (áreas) son las comarcas, islas, municipios, distritos sanitarios, etc. Ejemplos de grupos socioeconómicos son grupos sexo-edad, sectores industriales o empresariales, etc.

En el contexto de la estimación en áreas pequeñas, se dice que un estimador de un parámetro en un dominio dado es directo si está basado solamente en los datos específicos del dominio. Un estimador directo puede usar también información auxiliar, como por ejemplo el total en el dominio de una variable “ x ” relacionada con la variable de interés “ y ”.

Un estimador directo es típicamente un estimador basado en el diseño muestral, aunque en ocasiones su uso se justifique con modelos. Los estimadores basados en el diseño muestral utilizan los pesos muestrales. Las inferencias derivadas de los mismos están basadas en la distribución de probabilidad inducida por el mecanismo aleatorio de extracción de la muestra, bajo el supuesto de que los valores de la variable en los elementos de la población permanecen fijos. Los estimadores asistidos por modelos se introducen a partir de modelos de trabajo, pero optimizando sus propiedades de sesgo y varianza respecto de la distribución del diseño. En la literatura estadística estos estimadores también se consideran basados en el diseño muestral. Un dominio es grande si la muestra específica del dominio es suficientemente grande para obtener estimadores directos con una precisión adecuada. Un dominio es pequeño en caso contrario. En este texto usaremos el

término área pequeña para denotar a los dominios pequeños.

En todas las sociedades hay cada vez una mayor exigencia de información estadística, tanto en cantidad como en calidad. Este fenómeno es consecuencia de la mayor cultura económica y social. La recolección de datos estadísticos y su utilización para estimar parámetros demográficos o socioeconómicos es de vital importancia para el mantenimiento de nuestra sociedad de la información.

El estado de la nación, el de la comunidad autónoma o el de la provincia puede diagnosticarse a partir del análisis de los datos publicados por las Oficinas de Estadística. Parámetros como el Índice de Precios al Consumo, Producto Interior Bruto, Tasa de Paro, Tasa de Natalidad, Ingresos Netos por Hogar, indicadores de pobreza, etc., están siendo utilizados constantemente por los gobernantes para decidir cuándo y cómo invertir el dinero público y, más generalmente, para elaborar políticas sociales y económicas.

En esta memoria se formularán algunos estimadores de áreas pequeñas basados en modelos de área multivariantes, se estudian sus propiedades y se ilustra su eventual utilización en la encuesta de condiciones de vida (ECV).

1.3. Estado del arte

Los modelos de regresión lineal mixta (véase Searle, Casella y McCulloch, 1982) incrementan la eficiencia de la información usada en el proceso de estimación estableciendo nexos o relaciones entre todas las observaciones de la muestra, y al mismo tiempo introduciendo variabilidad entre áreas. Estos modelos se han usado en Estados Unidos para estimar ingresos per cápita en áreas pequeñas (Fay y Herriot, 1979), para estimar totales de dominios pequeños a partir de datos censales (Ericksen y Kadane, 1985, y Dick, 1995 en el censo canadiense), y para estudios de pobreza en población escolar (National Research Council, 2000).

Conviene mencionar que utilizando estimadores de áreas pequeñas, el Departamento de Educación de Estados Unidos asigna más de 7000 millones de dólares en fondos generales a los condados, y luego los estados distribuyen estos fondos entre los distritos escolares (Rao, 2003). El uso de estas técnicas no se restringe a datos socioeconómicos. El trabajo de Battese, Harter y Fuller (1988) es un ejemplo de aplicación al campo de la agricultura. Estos autores usaron un modelo lineal mixto para estimar las extensiones de cultivos de maíz y soja en condados estadounidenses.

Cuando los parámetros son lineales (combinaciones lineales de los valores que la variable objetivo toma en los elementos de la población), los predictores lineales insesgados óptimos (BLUP - Best Linear Unbiased Predictor) dependen de algunos parámetros desconocidos, habitualmente componentes de la varianza o correlaciones. Cuando esos parámetros se reemplazan por estimadores, entonces los correspondientes predictores se denominan "empíricos" (EBLUP - Empirical BLUP).

Los predictores EBLUP presentan el inconveniente de que no existe fórmula explícita exacta para su error cuadrático medio de predicción (MSE - Mean Squared Error). En la literatura científica han aparecido diversas aproximaciones para esta cantidad. La primera simplificación del MSE fue obtenida por Kackar y Harville (1981) asumiendo normalidad para los errores y de los efectos aleatorios del modelo. En un segundo

artículo, Kackar y Harville (1984) obtuvieron una aproximación del MSE y propusieron un estimador.

Prasad y Rao (1990) llegaron a una aproximación asintótica del MSE para modelos con matrices de covarianza diagonales a bloques. Bajo ciertas condiciones de regularidad para los modelos y los estimadores de las componentes de la varianza, demostraron que cuando el número de bloques D tiende a infinito, su aproximación es del orden $1/D$. También propusieron un estimador del MSE y dieron expresiones específicas para tres tipos de modelos, concretamente los modelos Fay-Herriot, con errores anidados y con coeficientes aleatorios.

Los estimadores de las componentes de la varianza, obtenidos por el método del ajuste de constantes, satisfacen las condiciones de regularidad mencionadas; sin embargo, esto no ocurre con los estimadores de máxima verosimilitud. Datta y Lahiri (2000) obtuvieron el estimador análogo del MSE en modelos con matrices de covarianza diagonales por bloques y con componentes de la varianza estimadas por máxima verosimilitud (ML) o por máxima verosimilitud residual (REML). Más recientemente, Das, Jiang y Rao (2004) estudiaron la aproximación del error de predicción en una clase más amplia de modelos, cuando las componentes de la varianza se estiman por los métodos ML o REML. En lo relativo a la estimación del error cuadrático medio mediante bootstrap paramétrico se pueden citar los trabajos de González-Manteiga et al. (2007, 2008a, 2008b, 2010).

En el contexto de la estimación de indicadores de pobreza en áreas pequeñas, usando datos de la encuesta de condiciones de vida de 2006, Molina y Morales (2009) aplican estimadores EBLUP basados en el modelo de Fay-Herriot a la estimación de indicadores de pobreza de provincias. Posteriormente, usando datos de EECV de 2004-2006 y de 2004-2008 respectivamente, Esteban y otros (2012) Marhuenda y otros (2013) aplican estimadores EBLUP, basados en un modelos Fay-Herriot temporales y espacio-temporales, a la estimación de indicadores de pobreza de provincias. Estos trabajos presentan unas aplicaciones con un alto potencial de adaptabilidad al contexto Español.

En los últimos años, muchos investigadores has estudiado la aplicabilidad del modelo Fay-Herriot a problemas de estimación en áreas pequeñas. Sin ser exhaustivo, se pueden citar algunos trabajos relacionados con el modelo Fay-Herriot. Prasad & Rao (1990), Datta & Lahiri P. (2000), Das et al. (2004), González-Manteiga et al. (2010), Jiang et al. (2011), Datta et al. (2011) y Kubokawa (2011) dan herramientas para medir la incertidumbre y estimar errores cuadráticos medios de estimadores de áreas pequeñas basados en modelos. Datta et al. (2011), Bell et al. (2013) y Pfeffermann et al. (2014) estudian el problema de la consistencia con estimaciones directas en dominios con nivel de agregación superior (benchmarking). Ybarra & Lohr (2008) proponen un estimador de áreas pequeñas que tiene en cuenta la variabilidad de la información auxiliar. Slud et al. (2011) se interesan por la estimación en áreas pequeñas con datos muestrales censurados por la izquierda. Herrador et al. (2011) tratan situaciones donde las áreas se dividen en dos grandes grupos y los efectos aleatorios tienen varianzas distintas en ambos grupos.

A los estadísticos se les pide frecuentemente estimar medidas descriptivas correladas, como inidcadores de pobreza o del mercado laboral. Los modelos multivariantes tienen en cuenta la correlación entre variables y se adaptan a este tipo de situaciones. En la literatura de estimación en áreas pequeñas se pueden encontrar algunos artículos donde se emplean modelos lineales mixtos multivariantes. Fay (1987) y Datta et al. (1991) compararon la precisión de estimadores en áreas pequeñas obtenidos de modelos univariantes para

cada variable de respuesta con los obtenidos de modelos multivariantes. Datta et al. (1996) usaron también modelos Fay-Herriot multivariantes para obtener estimadores Bayes jerárquicos de los ingresos medianos de familias de cuatro personas en los estados de EEUU. González-Manteiga et al. (2008) estudiaron una clase de modelos Fay-Herriot multivariantes con un efecto aleatorio común a todas las componentes del vector de variables objetivo. Ellos introdujeron además estimadores bootstrap de los errores de predicción. Esta memoria introduce una clase más general de modelos multivariantes que utiliza distintos efectos aleatorios para las componentes del vector de variables objetivo.

Los datos históricos dan información relevante que puede ser usada para mejorar los estimadores de áreas pequeñas. Varios autores han propuesto extensiones del modelo Fay-Herriot que usan información temporal. Choudry & Rao (1989) introdujeron un modelo que incluye varios instantes temporales y consideraron una estructura de correlación en los errores. Rao & Yu (1994) propusieron un modelo que toma información cruzada entre áreas y a lo largo del tiempo. Ghosh et al. (1996) propusieron un modelo de correlación temporal para estimar los ingresos medianos de familias de cuatro personas en los estados de EEUU. Datta et al. (1999), You & Rao (2000), Datta et al. (2002), Esteban et al. (2011, 2012), Marhuenda et al. (2013) y Morales et al. (2015) estudiaron extensiones del modelo Rao-Yu con aplicaciones a la estimación de indicadores de pobreza y del mercado laboral. Pfeiffermann & Burck (1990) y Singh et al. (2005) consideraron modelos con pendientes que varían con el tiempo y que siguen un proceso autoregresivo. Esta memoria aplica modelos Fay-Herriot multivariantes al tratamiento de datos correlados temporalmente.

1.4. Resumen

La parte concreta de la estadística matemática a la que pertenece el trabajo que se ha realizado es conocida como estimación en áreas pequeñas basada en modelos. De forma muy abreviada se puede decir que su objetivo principal consiste en el estudio de factores adicionales que se incluyen en una relación, en la que intervienen la variable que es objeto de estudio y la información auxiliar a la que se tiene acceso, que pretenden explicar las características de los distintos dominios que forman parte de una muestra dada. En particular el interés está centrado en el caso en el que los tamaños muestrales de los dominios no son lo suficientemente grandes como para ofrecer estimaciones directas por sí mismos.

Con el objeto de abarcar un mayor número de posibles aplicaciones prácticas, se ha tenido en cuenta el caso multivariante, dado que en general el interés suele estar centrado en varias variables respuesta. A partir de lo anterior, se puede adelantar que todo el trabajo se centra en el estudio de un modelo lineal mixto multivariante, en el que intervienen un vector de variables en las que se está interesado, un vector de variables auxiliares, un factor aleatorio que pretende contemplar las áreas pequeñas y un vector aleatorio de errores debidos a la estimación.

En el presente trabajo se pueden considerar, principalmente, tres partes. A continuación, se describen cada una de ellas. La primera parte (capítulo 2) consiste en una introducción a la estimación en áreas pequeñas basadas en modelos y una descripción de la notación principal, así como los métodos de estimación que se han utilizado. La segunda parte (capítulos 3, 4 y 5) está centrada en el estudio de tres modelos muy concretos y una validación de los mismos utilizando, para ello, simulaciones de muestras. Las simulaciones

se han realizado con el programa estadístico R versión 2.13.1. Por último, en la tercera parte (capítulo 5) se expone una aplicación de los modelos estudiados en la parte segunda basada en una muestra de datos reales pertenecientes al ámbito socio-económico.

En el párrafo precedente se ha descrito, de forma muy abreviada, cada una de las partes en las que está dividido el trabajo. Por ello, en lo que sigue, se avanza un poco más por tal vía añadiendo algunos detalles que pueden ser de interés para el lector que quiera obtener, de una forma rápida, una idea bastante aproximada del alcance del estudio que se ha realizado.

En la primera parte se estudia el origen y motivación de la estimación en áreas pequeñas y el interés principal se centra en el caso del modelo de regresión lineal mixto multivariante. En primer lugar, se describe de forma pormenorizada la notación seguida, presentando la forma resumida del modelo y también la forma matricial, insistiendo en cada una de los distintos elementos que forman parte del modelo. El método de estimación que se sigue, tanto de los parámetros como de los efectos, es el conocido como método de la máxima verosimilitud residual. Para conocer una medida de la precisión del estimador obtenido para la variables objetivo se aplican unos métodos de aproximación asintóticos, dado que no existen medios analíticos para obtener una medida exacta.

En los tres capítulos que forman la segunda parte se presentan cada uno de los tres modelos que se han considerado, según la complejidad de la variabilidad de los efectos incluidos en el modelo general. Para referirse a ellos se ha optado por denominarlos de la forma siguiente: modelo de varianzas diagonal, modelo AR(1) y modelo AR(1) heterocedástico. En los capítulos se expone de forma breve la metodología que se ha seguido para introducir cada uno de los modelos. En primer lugar, se describe como quedan los principales elementos que intervienen en el proceso de estimación por máxima verosimilitud residual. En segundo lugar, se realizan simulaciones para contrastar la validez del modelo. Se han considerado tres simulaciones que, de forma resumida, consisten en lo siguiente:

1. Cálculo por simulación tipo Monte Carlo de los sesgos y errores cuadráticos medios empíricos de las estimaciones.
2. Cálculo por simulación tipo Monte Carlo de los sesgos y errores cuadráticos medios empíricos de los estimadores analíticos en la estimación de los errores cuadráticos medios del estimador de la variable objetivo.
3. Comprobación del funcionamiento del bootstrap paramétrico en la estimación de los errores cuadráticos medios de las estimaciones de la variable objetivo.

La tercera y última parte del trabajo presenta una aplicación práctica de todo lo estudiado. Para ello, se han usado muestras de datos reales procedentes de la encuesta de condiciones de vida de los años 2005 y 2006 con el objetivo de estimar indicadores de pobreza. El papel de áreas pequeñas lo han realizado la combinación de las distintas provincias con las dos categorías de la variable sexo. Se han ajustado los tres modelos estudiados y se han comparado los resultados con el objeto de averiguar cuál de ellos es el que se comporta mejor para los datos de la muestra.

2

Modelos de área lineales multivariantes

2.1. Introducción

En muchas tareas de investigación se requiere el análisis de una colección de datos que pueden ser de naturaleza muy diversa y que se reúnen en una muestra. A veces resulta conveniente para el estudio de los datos considerar una diferenciación adecuada entre ellos y clasificarlos en dominios, más o menos numerosos, que tengan una o varias características en común. La ventaja de tener en cuenta lo anterior es que se pueden plantear estimaciones diversas para cada uno de los dominios; sin embargo, un problema que surge con frecuencia es la insuficiencia del tamaño de la muestra en alguno de los dominios. Eso tiene como consecuencia que la precisión de las estimaciones no sea la adecuada.

Para tratar de corregir lo que se acaba de describir, surge la estimación en áreas pequeñas. De forma breve se puede decir que un área pequeña consiste en una parte de la muestra total de datos que por sí misma no puede producir estimaciones directas con una precisión adecuada. Las innovaciones en procedimientos de estimación se han sucedido, se pueden citar como ejemplos los trabajos Ghosh y Rao (1994), Rao (2003) o Jiang y Lahiri (2006), que ofrecen una descripción detallada de este novedoso enfoque estadístico.

Uno de los métodos de predicción que más protagonismo ha adquirido, dentro de todo lo desarrollado en áreas pequeñas, está basado en los modelos de regresión lineal mixta (véase Searle, Casella y McCulloch, 1992). Estos modelos incrementan la eficiencia de la información usada en el proceso de estimación estableciendo nexos o relaciones entre todas las observaciones de la muestra. Al mismo tiempo introducen la variabilidad que existe entre las distintas áreas. Los modelos de este estilo se han usado en Estados Unidos para estimar ingresos per cápita en áreas pequeñas (Fay y Herriot, 1979), para estimar conteos no incluidos en el censo (Ericksen y Kadane, 1985, y Dick, 1995 en el censo canadiense), y para estudios de pobreza en población escolar (National Research Council, 2000). Conviene mencionar que utilizando estos estimadores, el Departamento de Educación de Estados Unidos asigna más de 7000 millones de dólares en fondos generales a los condados, y luego los estados distribuyen estos fondos entre los distritos escolares (Rao, 2003). El trabajo de Battese, Harter y Fuller (1988) es un ejemplo de aplicación al campo de la agricultura. Estos autores usaron un modelo lineal mixto para estimar extensiones de determinados cultivos.

En este trabajo se desarrolla el caso de un modelo lineal mixto multivariante; motivado, entre otras cosas, por el hecho de que en la práctica es frecuente que exista más de una variable que es objeto de estudio. En los apartados que se ofrecen a continuación se describe la notación de los modelos estudiados, así como los diversos métodos de estimación de los parámetros y predictores que intervienen en los mismos.

2.2. Definición del modelo

Sea P un población finita particionada en D subpoblaciones o áreas pequeñas que se denotan por P_d , de tamaños N_d , $d = 1, \dots, D$; es decir, $P = P_1 \cup \dots \cup P_D$, donde $P_{d_1} \cap P_{d_2} = \emptyset$, $d_1 \neq d_2$. Sean

$$(y_{dj1}, \dots, y_{djR})', \quad j = 1, \dots, N_d, \quad d = 1, \dots, D,$$

los valores que toma el vector R -variante objeto de estudio en las unidades del área d . Las medias poblacionales

$$\mu_{dr} = \frac{1}{N_d} \sum_{j=1}^{N_d} y_{djr}, \quad r = 1, \dots, R,$$

son los parámetros de interés. Sean

$$\mu_d = (\mu_{d1}, \dots, \mu_{dR})', \quad y_d = (y_{d1}, \dots, y_{dR})', \quad d = 1, \dots, D,$$

los vectores de medias de las áreas y de estimadores directos, de modo que y_{dr} es el estimador directo de la media poblacional μ_{dr} .

El modelo muestral indica que los estimadores directos son centrados y se expresa de la forma

$$y_d = \mu_d + e_d, \quad d = 1, \dots, D,$$

donde los vectores $e_d \sim N(0, V_{ed})$ son independientes y las matrices de covarianzas V_{ed} , de dimensión $R \times R$, son conocidas.

Se supone además que las medias μ_{dr} están linealmente relacionadas con p_r variables explicativas asociadas a la r -ésima variable en el área d . Sea β_r un vector columna de tamaño p_r . Sea $x_{dr} = (x_{dr1}, \dots, x_{drp_r})$ el vector fila de variables explicativas para el estimador directo y_{dr} y sea $p = \sum_{r=1}^R p_r$. Se considera la matriz $X_d = \text{diag}(x_{d1}, \dots, x_{dR})_{R \times p}$ y el vector de coeficientes de regresión $\beta = (\beta'_1, \dots, \beta'_r)'_{p \times 1}$. Sea 1_r un vector $r \times 1$ de unos. Sea I_D la matriz identidad $D \times D$. Finalmente, se considera el vector $u_d = (u_{d1}, \dots, u_{dR})'_{R \times 1}$ de efectos aleatorios asociados al área d , y que recoge las variaciones entre áreas no explicadas por X_1, \dots, X_D . El modelo de regresión establece la relación lineal entre el vector de medias y las variables explicativas. El modelo es

$$\mu_d = X_d \beta + u_d, \quad u_d \sim N(0, V_{ud}), \quad d = 1, \dots, D,$$

donde los vectores $u_d \sim N(0, V_{ud})$ son independientes e independientes de los e_d y las matrices de covarianzas V_{ud} dependen de m parámetros desconocidos que se denotan por $\theta_1, \dots, \theta_m$. El número de parámetros m que intervienen en la matriz V_{ud} está limitado por el número de elementos distintos de la matriz; es decir, $1 \leq m \leq \frac{R(R-1)}{2} + R$.

Se definen los vectores y matrices

$$y = \underset{1 \leq d \leq D}{\text{col}}(y_d), \quad u = \underset{1 \leq d \leq D}{\text{col}}(u_d), \quad e = \underset{1 \leq d \leq D}{\text{col}}(e_d), \quad u_d = \underset{1 \leq r \leq R}{\text{col}}(u_{dr}), \quad e_d = \underset{1 \leq r \leq R}{\text{col}}(e_{dr}),$$

$$Z_d = \underset{1 \leq \ell \leq D}{\text{col}}(\delta_{\ell d} I_R), \quad Z = \underset{1 \leq d \leq D}{\text{col}}'(Z_d) = I_{DR}.$$

El modelo completo admite la siguiente representación lineal

$$y = X\beta + Zu + e = X\beta + Z_1 u_1 + \cdots + Z_D u_D + e, \quad (2.1)$$

donde e, u_1, \dots, u_D son independientes con distribuciones

$$e \sim N(0, V_e), \quad u \sim N(0, V_u) \quad \text{y} \quad u_d \sim N(0, V_{ud}), \quad d = 1, \dots, D,$$

donde $V_u = \underset{1 \leq d \leq D}{\text{diag}}(V_{ud})$. La estructura matricial del modelo es

$$\begin{pmatrix} y_{11} \\ \vdots \\ y_{1R} \\ \vdots \\ y_{D1} \\ \vdots \\ y_{DR} \end{pmatrix} = \begin{pmatrix} x_{11} & & & \\ & \ddots & & \\ & & x_{1R} & \\ & & \vdots & \\ x_{D1} & & & \\ & \ddots & & \\ & & & x_{DR} \end{pmatrix} \begin{pmatrix} \beta_1 \\ \vdots \\ \beta_R \end{pmatrix} + \begin{pmatrix} I_R \\ \vdots \\ 0 \end{pmatrix} \begin{pmatrix} u_{11} \\ \vdots \\ u_{1R} \end{pmatrix} + \cdots + \begin{pmatrix} 0 \\ \vdots \\ I_R \end{pmatrix} \begin{pmatrix} u_{D1} \\ \vdots \\ u_{DR} \end{pmatrix} + \begin{pmatrix} e_{11} \\ \vdots \\ e_{1R} \\ \vdots \\ e_{D1} \\ \vdots \\ e_{DR} \end{pmatrix}.$$

La parte del modelo correspondiente al área d es

$$\begin{cases} y_{d1} = x_{d1}\beta_1 + u_{d1} + e_{d1}, \\ y_{d2} = x_{d2}\beta_2 + u_{d2} + e_{d2}, \\ \vdots \\ y_{dR} = x_{dR}\beta_R + u_{dR} + e_{dR}. \end{cases}$$

Se observa que al cambiar de variable dentro de la misma área va cambiando el vector de parámetros de regresión.

Los vectores y matrices

$$\delta_r = \left(0, \dots, 0, 1^{(r)}, 0, \dots, 0\right)'_{R \times 1}, \quad Z_{.r} = \text{diag}(\delta_r, \dots, \delta_r)_{RD \times D} \quad \text{y} \quad u_{.r} = \text{col}(u_{1r}, \dots, u_{Dr})_{D \times 1} \quad (2.2)$$

permiten dar la representación alternativa del modelo

$$y = X\beta + Z_{.1}u_{.1} + \cdots + Z_{.R}u_{.R} + e, \quad (2.3)$$

donde los $u_{.r} \sim N(0, V_{urr})$ son dependientes e independientes de $e \sim N(0, V_e)$. En este caso el vector $\underline{u} = \underset{1 \leq r \leq R}{\text{col}}(u_{.r})$ es normal multivariante con vector de medias 0 y matriz de covarianzas $V_{\underline{u}} = (V_{ur_1 r_2})_{r_1, r_2=1, \dots, R}$,

$V_{ur_1r_2} = \text{diag}(\text{cov}(u_{dr_1}, u_{dr_2}))$. Bajo esta representación, la estructura matricial del modelo es

$$\begin{pmatrix} y_{11} \\ \vdots \\ y_{1R} \\ \vdots \\ y_{D1} \\ \vdots \\ y_{DR} \end{pmatrix} = \begin{pmatrix} x_{11} & & & & \\ & \ddots & & & \\ & & x_{1R} & & \\ & & \vdots & & \\ x_{D1} & & & & \\ & & & \ddots & \\ & & & & x_{DR} \end{pmatrix} \begin{pmatrix} \beta_1 \\ \vdots \\ \beta_R \end{pmatrix} + \begin{pmatrix} e_{11} \\ \vdots \\ e_{1R} \\ \vdots \\ e_{D1} \\ \vdots \\ e_{DR} \end{pmatrix} \\ + \begin{pmatrix} \delta_1 & \dots & 0 \\ & \ddots & \\ 0 & \dots & \delta_1 \end{pmatrix} \begin{pmatrix} u_{11} \\ \vdots \\ u_{D1} \end{pmatrix} + \dots + \begin{pmatrix} \delta_R & \dots & 0 \\ & \ddots & \\ 0 & \dots & \delta_R \end{pmatrix} \begin{pmatrix} u_{1R} \\ \vdots \\ u_{DR} \end{pmatrix}.$$

La parte del modelo correspondiente a la variable r -ésima es

$$\begin{cases} y_{1r} = x_{1r}\beta_r + u_{1r} + e_{1r}, \\ y_{2r} = x_{2r}\beta_r + u_{2r} + e_{2r}, \\ \vdots \\ y_{Dr} = x_{Dr}\beta_r + u_{Dr} + e_{Dr}. \end{cases}$$

Al contrario que en el caso anterior, ahora el vector de parámetros de regresión es siempre el mismo pues no se cambia de variable sino de área. Esta representación es menos útil para la realización de cálculos.

El vector de medias y la matriz de covarianzas de y son

$$E(y) = X\beta \quad y \quad V = \text{var}(y) = Z'V_uZ + V_e = V_u + V_e = \text{diag}(V_d),_{1 \leq d \leq D}$$

donde

$$V_d = V_{ud} + V_{ed}, \quad d = 1, \dots, D.$$

En el vector a predecir,

$$\mu_d = X_d\beta + u_d,$$

intervienen p efectos fijos que se corresponden con los parámetros que forman parte del vector β y R efectos aleatorios que son los que se corresponden a cada una de las variables en el área d . En este trabajo se estudian varios modelos (2.1) resultantes de considerar distintas matrices de correlaciones V_u .

Los estimadores y predictores insesgados lineales óptimos (BLUE y BLUP) de β y u son

$$\hat{\beta}_B = (X'V^{-1}X)^{-1}X'V^{-1}y, \quad \hat{u}_B = V_uZ'V^{-1}(y - X\hat{\beta}_B). \quad (2.4)$$

Para calcular $\hat{\beta}$ y \hat{u} se aplican las fórmulas

$$\hat{\beta}_B = \left(\sum_{d=1}^D X_d'V_d^{-1}X_d \right)^{-1} \left(\sum_{d=1}^D X_d'V_d^{-1}y_d \right), \quad \hat{u}_B = \text{col}_{1 \leq d \leq D} \left(V_{ud}Z_d'V_d^{-1}(y_d - X_d\hat{\beta}_B) \right).$$

Puesto que los parámetros de la matriz V_u son desconocidos, el BLUE y el BLUP no son calculables. En la práctica, se sustituye V_u por un estimador adecuado y de esa forma se obtienen los BLUE y BLUP empíricos (EBLUE y EBLUP). En la siguiente sección se introduce el método de la máxima verosimilitud residual para estimar las componentes de la la matriz V_u .

2.3. Máxima verosimilitud residual

El método de la máxima verosimilitud residual (REML) maximiza la función de densidad conjunta de un vector de $DR - p$ contrastes independientes $\omega = W'y$, donde W es una matriz $DR \times (DR - p)$ con columnas linealmente independientes y tal que $W'W = I_{DR-p}$ y $W'X = 0_{DR-p}$. Es fácil comprobar que ω es independiente del BLUE $\hat{\beta}_B$ dado en (2.4). La función de densidad conjunta de ω es la verosimilitud REML y su logaritmo neperiano es la logverosimilitud REML del método de la máxima verosimilitud residual. Para el modelo (2.1), la logverosimilitud REML es

$$l_{reml}(\theta) = -\frac{DR-p}{2} \log 2\pi + \frac{1}{2} \log |X'X| - \frac{1}{2} \log |V| - \frac{1}{2} \log |X'V^{-1}X| - \frac{1}{2} y'Py,$$

donde $\theta = (\theta_1, \dots, \theta_m)$,

$$P = V^{-1} - V^{-1}X(X'V^{-1}X)^{-1}X'V^{-1}, \quad PVP = P, \quad PX = 0.$$

Sean $V_{dl} = \frac{\partial V_d}{\partial \theta_l}$. Se verifica que

$$V_l = \frac{\partial V}{\partial \theta_l} = \text{diag}_{1 \leq d \leq D} (V_{dl}), \quad P_l = \frac{\partial P}{\partial \theta_l} = -P \frac{\partial V}{\partial \theta_l} P = -PV_l P, \quad l = 1, \dots, m.$$

Si se deriva l_{reml} con respecto de θ_l , entonces se obtiene lo que sigue

$$S_l = \frac{\partial l_{reml}}{\partial \theta_l} = -\frac{1}{2} \text{tr}(PV_l) + \frac{1}{2} y'PV_l Py, \quad l = 1, \dots, m.$$

Volviendo a derivar las expresiones S_l respecto de θ_a y θ_b , tomando esperanzas y cambiando de signo, se obtiene lo siguiente:

$$F_{ab} = \frac{1}{2} \text{tr}(PV_a PV_b), \quad a, b = 1, \dots, m.$$

Para el cálculo de los estimadores REML el método Fisher-scoring usa la fórmula de actualización

$$\theta^{k+1} = \theta^k + F^{-1}(\theta^k) S(\theta^k).$$

Para la primera iteración del algoritmo se propondrán unos valores iniciales para los parámetros. Estos valores se denominan semillas. En este capítulo no se propondran semillas de aplicabilidad general. Las semillas son el punto de arranque de los algoritmos de ajustes y deben de darse en función del modelo concreto que se utilice. Este problema se aborda en los capítulos posteriores.

Los estimadores EBLUE-REML de β y EBLUP-REML de u son

$$\hat{\beta}_E = (X'\hat{V}^{-1}X)^{-1}X'\hat{V}^{-1}y, \quad \hat{u}_E = \hat{V}_u Z'V^{-1}(y - X\hat{\beta}_E), \quad \hat{V}_u = V_u(\hat{\theta}).$$

Las distribuciones asintóticas de los estimadores REML de θ y β son

$$\hat{\theta} \sim N_m(\theta, F^{-1}(\theta)), \quad \hat{\beta} \sim N_p(\beta, (X'V^{-1}X)^{-1}).$$

A partir de lo anterior se tiene que los intervalos de confianza asintóticos a nivel $1 - \alpha$ para θ_l y β_j son

$$\hat{\theta}_l \pm z_{\alpha/2} v_{ll}^{1/2}, \quad l = 1, \dots, m, \quad \hat{\beta}_j \pm z_{\alpha/2} q_{jj}^{1/2}, \quad j = 1, \dots, p,$$

donde $\hat{\theta} = \theta^\kappa$, $F^{-1}(\theta^\kappa) = (v_{ab})_{a,b=1,\dots,m}$, $(X'V^{-1}(\theta^\kappa)X)^{-1} = (q_{ij})_{i,j=1,\dots,p}$, κ es la iteración final del algoritmo Fisher-scoring y z_α es el α -cuantil de la distribución $N(0, 1)$. Observado $\hat{\beta}_j = \beta_0$, el p -valor del contraste de la hipótesis $H_0 : \beta_j = 0$ es

$$p = 2P_{H_0}(\hat{\beta}_j > |\beta_0|) = 2P(N(0, 1) > |\beta_0|/\sqrt{q_{jj}}).$$

Sea el vector $c' = (c_1, \dots, c_m)$ y la combinación lineal $c'\theta$. A partir de la distribución asintótica de $\hat{\theta}$ se deduce que $c'\hat{\theta} \sim N_m(c'\theta, c'F^{-1}c)$, y así se puede hacer el contraste $H_0 : c'\theta = 0$. Se rechaza H_0 si

$$\left| \frac{c'\hat{\theta}}{\sqrt{c'F^{-1}c}} \right| > z_{\alpha/2}.$$

2.4. Cálculos matriciales

En este apartado se muestra cómo realizar los cálculos matriciales del algoritmo Fisher-scoring sin construir matrices de dimensión RD .

$$\begin{aligned} Q &= (X'V^{-1}X)^{-1} = \left(\sum_{d=1}^D X'_d V_d^{-1} X_d \right)^{-1}, \\ P &= \text{diag}_{1 \leq d \leq D} (V_d^{-1}) - \text{col}_{1 \leq d \leq D} (V_d^{-1} X_d) Q \text{col}'_{1 \leq d \leq D} (X'_d V_d^{-1}), \\ PV_l &= \text{diag}_{1 \leq d \leq D} (V_d^{-1} V_{dl}) - \text{col}_{1 \leq d \leq D} (V_d^{-1} X_d) Q \text{col}'_{1 \leq d \leq D} (X'_d V_d^{-1} V_{dl}), \\ \text{tr}(PV_r) &= \sum_{d=1}^D \text{tr}(V_d^{-1} V_{dl}) - \sum_{d=1}^D \text{tr}(X'_d V_d^{-1} V_{dl} V_d^{-1} X_d Q), \\ \text{tr}(PV_a PV_b) &= \sum_{d=1}^D \text{tr}(V_d^{-1} V_{da} V_d^{-1} V_{db}) - 2 \sum_{d=1}^D \text{tr}(X'_d V_d^{-1} V_{da} V_d^{-1} V_{db} V_d^{-1} X_d Q) \\ &\quad + \text{tr} \left\{ \left(\sum_{d=1}^D X'_d V_d^{-1} V_{da} V_d^{-1} X_d \right) Q \left(\sum_{d=1}^D X'_d V_d^{-1} V_{db} V_d^{-1} X_d \right) Q \right\}. \\ y' PV_l P y &= \sum_{d=1}^D y'_d V_d^{-1} V_{dl} V_d^{-1} y_d - \left(\sum_{d=1}^D y'_d V_d^{-1} V_{dl} V_d^{-1} X_d \right) Q \left(\sum_{d=1}^D y'_d V_d^{-1} X_d \right)' \\ &\quad - \left(\sum_{d=1}^D y'_d V_d^{-1} X_d \right) Q \left(\sum_{d=1}^D X'_d V_d^{-1} V_{dl} V_d^{-1} y_d \right) \\ &\quad + \left(\sum_{d=1}^D y'_d V_d^{-1} X_d \right) Q \left(\sum_{d=1}^D X'_d V_d^{-1} V_{dl} V_d^{-1} X_d \right) Q \left(\sum_{d=1}^D y'_d V_d^{-1} X_d \right)'. \end{aligned}$$

2.5. Matriz de errores cuadráticos medios de los EBLUP

El predictor insesgado lineal óptimo y empírico (EBLUP) de la media $\mu = X\beta + Zu$ se construye sustituyendo β y u por el EBLUE $\hat{\beta}_E$ y el EBLUP \hat{u}_E respectivamente. Por tanto, el EBLUP de $\mu = X\beta + Zu$ es

$$\hat{\mu}_E = X\hat{\beta}_E + Z\hat{u}_E.$$

El objetivo de este apartado es calcular la matriz de errores cuadráticos medios cruzados

$$MSE(\hat{\mu}_E) = E((\hat{\mu}_E - \mu)(\hat{\mu}_E - \mu)').$$

Mediante adición y sustracción del término $\hat{\mu}_B$, se tiene que

$$\hat{\mu}_E - \mu = \hat{\mu}_B - \mu + \hat{\mu}_E - \hat{\mu}_B.$$

Por tanto,

$$\begin{aligned} (\hat{\mu}_E - \mu)(\hat{\mu}_E - \mu)' &= (\hat{\mu}_B - \mu)(\hat{\mu}_B - \mu)' + (\hat{\mu}_B - \mu)(\hat{\mu}_E - \hat{\mu}_B)' + (\hat{\mu}_E - \hat{\mu}_B)(\hat{\mu}_B - \mu)' \\ &+ (\hat{\mu}_E - \hat{\mu}_B)(\hat{\mu}_E - \hat{\mu}_B)'. \end{aligned} \quad (2.5)$$

Bajo las hipótesis de normalidad sobre u y e , Kackar and Harville (1984) demostraron que para cualquier estimador de θ insesgado e invariante mediante traslaciones, el valor esperado de los dos últimos términos de la ecuación (2.5) son nulos. Por tanto, tomando esperanzas se obtiene

$$MSE(\hat{\mu}_E) = MSE(\hat{\mu}_B) + E[(\hat{\mu}_E - \hat{\mu}_B)(\hat{\mu}_E - \hat{\mu}_B)']. \quad (2.6)$$

Por el teorema general de predicción, se tiene que

$$MSE(\hat{\mu}_B) = G_1(\theta) + G_2(\theta),$$

donde, para $T = V_u - V_u Z' V^{-1} Z V_u$, G_1 y G_2 son

$$\begin{aligned} G_1(\theta) &= TZT', \\ G_2(\theta) &= (X - TZT'V_e^{-1}X)Q(X' - X'V_e^{-1}TZT'). \end{aligned}$$

Dado que $Z = I_{RD}$, se obtiene $G_1(\theta) = T = \text{diag}(T_d)$, donde $T_d = V_{ud} - V_{ud}V_d^{-1}V_{ud} = G_{1d}(\theta)$. Además, $G_2(\theta) = (G_{2d_1d_2}(\theta))_{d_1, d_2=1, \dots, D}$, donde

$$G_{2d_1d_2} = (X_{d_1} - T_{d_1}V_{ed_1}^{-1}X_{d_1})Q(X_{d_1} - T_{d_1}V_{ed_1}^{-1}X_{d_1})'.$$

El interés principal se centra en el caso $d_1 = d_2$; es decir, en las matrices $R \times R$

$$G_{1d} = T_d, \quad G_{2d} = (X_d - T_dV_{ed}^{-1}X_d)Q(X_d - T_dV_{ed}^{-1}X_d)',$$

y en el error cuadrático medio $MSE(\hat{\mu}_{Bd}) = G_{1d}(\theta) + G_{2d}(\theta)$.

Para calcular la esperanza matemática del segundo miembro de la ecuación (2.6), se considera a $\hat{\mu}$ como una función vectorial de la variable θ ; es decir, $\hat{\mu}_B = \hat{\mu}(\theta)$ y $\hat{\mu}_E = \hat{\mu}(\hat{\theta})$. Utilizando la aproximación de Taylor de primer orden, se tiene que

$$(\hat{\mu}_E - \hat{\mu}_B)(\hat{\mu}_E - \hat{\mu}_B)' \approx S(\hat{\theta} - \theta)(\hat{\theta} - \theta)'S',$$

donde S es una matriz $DR \times m$

$$S = \begin{pmatrix} \frac{\partial \hat{\mu}_{11}}{\partial \hat{\theta}_1} & \dots & \frac{\partial \hat{\mu}_{11}}{\partial \hat{\theta}_m} \\ \vdots & & \vdots \\ \frac{\partial \hat{\mu}_{DR}}{\partial \hat{\theta}_1} & \dots & \frac{\partial \hat{\mu}_{DR}}{\partial \hat{\theta}_m} \end{pmatrix},$$

o alternativamente

$$S = \frac{\partial \hat{\mu}}{\partial \hat{\theta}} = \text{col}'_{1 \leq l \leq m} (s^{(l)}), \quad s^{(l)} = \frac{\partial \hat{\mu}}{\partial \theta_l} = \text{col}_{1 \leq d \leq D} (\text{col}_{1 \leq r \leq R} (s_{dr}^{(l)})), \quad s_{dr}^{(l)} = \frac{\partial \hat{\mu}_{dr}}{\partial \theta_l}.$$

En la nueva notación, se tiene que

$$\begin{aligned} S(\hat{\theta} - \theta)(\hat{\theta} - \theta)'S' &= \text{col}'_{1 \leq l \leq m} (s^{(l)}) \text{col}_{1 \leq l \leq m} (\hat{\theta}_l - \theta_l) \text{col}'_{1 \leq l \leq m} (\hat{\theta}_l - \theta_l) \text{col}_{1 \leq l \leq m} (s^{(l)'}) \\ &= \left(\sum_{l=1}^m s^{(l)} (\hat{\theta}_l - \theta_l) \right) \left(\sum_{l=1}^m (\hat{\theta}_l - \theta_l) s^{(l)'} \right) = \sum_{i=1}^m \sum_{j=1}^m (\hat{\theta}_i - \theta_i) (\hat{\theta}_j - \theta_j) s^{(i)} s^{(j)'}. \end{aligned}$$

Tomando esperanzas, se obtiene

$$E[S(\hat{\theta} - \theta)(\hat{\theta} - \theta)'S'] = \sum_{i=1}^m \sum_{j=1}^m E \left[(\hat{\theta}_i - \theta_i) (\hat{\theta}_j - \theta_j) s^{(i)} s^{(j)'} \right].$$

En lo que sigue es necesario introducir algo de notación para poder introducir las hipótesis de regularidad que permiten deducir una aproximación de la matriz de errores cuadráticos medios del EBLUP. Usamos la notación $f(D) = O(g(D))$ para dos funciones $f(D)$ y $g(D)$ que verifican la relación $\lim_{D \rightarrow \infty} |f(D)/g(D)| < \infty$. La notación $f(D) = \underline{O}(g(D))$ se usa para la relación más precisa $\lim_{D \rightarrow \infty} |f(D)/g(D)| \in (0, \infty)$, y $f(D) = o(g(D))$ se usa cuando el mismo límite es cero. Además, $f(D) = O_p(g(D))$ y $f(D) = o_p(g(D))$ denotan respectivamente acotación y convergencia a cero en probabilidad de $f(D)/g(D)$. Cuando $f(D)$ es una matriz $m \times n$ cuyos elementos son $O(g(D))$, se escribe $f(D) = [O(g(D))]_{m \times n}$, y la misma notación de corchetes se usa para el resto de símbolos de orden asintótico.

Se consideran las siguientes hipótesis de regularidad:

H1 $0 < p < \infty$ y $0 < r < \infty$

H2 $|x_{drj}| \leq x < \infty$

H3 Las matrices de varianzas V_{ud} , $d = 1, \dots, D$ son definidas positivas y sus elementos son uniformemente acotados

H4 $X'X = [\underline{Q}(D)]_{pr \times pr}$

$$H5 \quad X'V_e^{-1}X = [O(D)]_{pr \times pr}$$

$$H6 \quad \sum_{d=1}^D 1_r' V_{ed} 1_r = O(D)$$

$$H7 \quad (X'V^{-1}X)^{-1} = [O(D)]_{pr \times pr}$$

H8 $\hat{\sigma}_u^2 = k + y'Cy$ es un estimador de σ_u^2 insesgado, consistente e invariante por traslaciones, donde $k = O(1)$ y $C = \text{diag} \{ [O(D^{-1})]_{R \times R}, \dots, [O(D^{-1})]_{R \times R} \} + [O(D^{-2})]_{DR \times DR}$

Lema 1. Sea $v = Zu + e$ el vector que contiene la parte aleatoria del modelo (2.1). Entonces

$$s^{(i)} = \frac{\partial \hat{\mu}}{\partial \theta_i} = (F^{(i)} + L^{(i)})v,$$

donde

$$\begin{aligned} F^{(i)} &= -(I-R)XQX' \frac{\partial V^{-1}}{\partial \theta_i} A - \frac{\partial R}{\partial \theta_i} XQX'V^{-1}, \quad L^{(i)} = \frac{\partial R}{\partial \theta_i}, \\ A &= I - XQX'V^{-1}, \quad R = V_u V^{-1}, \quad Q = (X'V^{-1}X)^{-1}. \end{aligned}$$

Demostración. El BLUP de μ se puede escribir de la forma

$$\hat{\mu}_B = X\hat{\beta}_B + R(y - X\hat{\beta}_B) = X\hat{\beta}_B - RX\hat{\beta}_B + Ry = XQX'V^{-1}y - RXQX'V^{-1}y + Ry.$$

Sustituyendo y por su valor $y = X\beta + v$, se obtiene

$$\begin{aligned} \hat{\mu}_B &= XQX'V^{-1}X\beta - RXQX'V^{-1}X\beta + RX\beta + XQX'V^{-1}v - RXQX'V^{-1}v + Rv \\ &= X\beta + XQX'V^{-1}v + RA v. \end{aligned}$$

Derivando parcialmente respecto de θ_i se obtiene

$$\begin{aligned} s^{(i)} &= \frac{\partial \hat{\mu}_B}{\partial \theta_i} = -XQX' \frac{\partial V^{-1}}{\partial \theta_i} XQX'v + XQX' \frac{\partial V^{-1}}{\partial \theta_i} v + \frac{\partial R}{\partial \theta_i} Av + R \frac{\partial A}{\partial \theta_i} v \\ &= XQX' \frac{\partial V^{-1}}{\partial \theta_i} (I - XQX'V^{-1})v + \frac{\partial R}{\partial \theta_i} Av + R \frac{\partial A}{\partial \theta_i} v \\ &= XQX' \frac{\partial V^{-1}}{\partial \theta_i} Av + \frac{\partial R}{\partial \theta_i} Av + R \frac{\partial A}{\partial \theta_i} v. \end{aligned} \tag{2.7}$$

La derivada parcial de A respecto de θ_i es

$$\frac{\partial A}{\partial \theta_i} = XQX' \frac{\partial V^{-1}}{\partial \theta_i} XQX'V^{-1} - XQX' \frac{\partial V^{-1}}{\partial \theta_i} = -XQX' \frac{\partial V^{-1}}{\partial \theta_i} A.$$

Por tanto

$$\begin{aligned} s^{(i)} &= XQX' \frac{\partial V^{-1}}{\partial \theta_i} Av - RXQX' \frac{\partial V^{-1}}{\partial \theta_i} Av + \frac{\partial R}{\partial \theta_i} Av = \left[(I-R)XQX' \frac{\partial V^{-1}}{\partial \theta_i} A + \frac{\partial R}{\partial \theta_i} A \right] v \\ &= \left[(I-R)XQX' \frac{\partial V^{-1}}{\partial \theta_i} A - \frac{\partial R}{\partial \theta_i} XQX'V^{-1} + \frac{\partial R}{\partial \theta_i} \right] v = [F^{(i)} + L^{(i)}]v. \end{aligned}$$

donde

$$F^{(i)} = (I - R)XQX' \frac{\partial V^{-1}}{\partial \theta_i} A - \frac{\partial R}{\partial \theta_i} XQX'V^{-1}, \quad L^{(i)} = \frac{\partial R}{\partial \theta_i}.$$

Finalmente, se recuerda que

$$Z = I, \quad R = V_u V^{-1}, \quad V = V_u + V_e = \text{diag}(V_d), \quad V_d = V_{ud} + V_{ed},$$

$$1 \leq d \leq D$$

Las derivadas parciales de V^{-1} y R son

$$\frac{\partial V^{-1}}{\partial \theta_i} = -V^{-1} \frac{\partial V}{\partial \theta_i} V^{-1} = -V^{-1} W_i V^{-1},$$

$$\frac{\partial R}{\partial \theta_i} = \frac{\partial V_u}{\partial \theta_i} V^{-1} + V_u \frac{\partial V}{\partial \theta_i} = W_i V^{-1} - V_u V^{-1} W_i V^{-1} = (I - R) W_i V^{-1}.$$

Por tanto

$$F^{(i)} = -(I - R)XQX'V^{-1}W_iV^{-1}A - (I - R)W_iV^{-1}XQX'V^{-1}, \quad L^{(i)} = (I - R)W_iV^{-1}.$$

Lema 2. Las matrices $F^{(i)}$ y $L^{(i)}$ son tales que

(i) $L^{(i)} = \text{diag}(L_d^{(i)})$, con $L_d^{(i)} = [O(1)]_{R \times R}$, $d = 1, \dots, D$.

(ii) $F^{(i)} = [O(D^{-1})]_{DR \times DR}$.

Demostración.

La matriz $L_d^{(i)}$ admite la expresión $L_d^{(i)} = W_{di}V_d^{-1} - V_{ud}V_d^{-1}W_{di}V_d^{-1}$. Aplicando las hipótesis H1-H6 se obtiene (i).

La matriz $F^{(i)}$ admite la expresión

$$F^{(i)} = (I - R)XQX' \frac{\partial V^{-1}}{\partial \theta_i} A - L^{(i)}XQX'V^{-1}.$$

Aplicando (H6) y (H2) se tiene que $Q = (X'V^{-1}X)^{-1} = [O(D^{-1})]_{p \times p}$ y $XQX' = [O(D^{-1})]_{DR \times DR}$. Puesto que $L^{(i)} = [O(1)]_{DR \times DR}$ y $V^{-1} = [O(1)]_{DR \times DR}$, se comprueba que

$$L^{(i)}XQX'V^{-1} = [O(D^{-1})]_{DR \times DR}.$$

Con respecto al primer sumando se tiene que

$$I - R = [O(1)]_{DR \times DR}, \quad \frac{\partial V^{-1}}{\partial \theta_i} = [O(1)]_{DR \times DR} \quad \text{y} \quad (I - R)XQX' \frac{\partial V^{-1}}{\partial \theta_i} = [O(1)]_{DR \times DR}.$$

Al multiplicar por $A = I - XQX'V^{-1}$, se obtiene

$$(I - R)XQX' \frac{\partial V^{-1}}{\partial \theta_i} A = (I - R)XQX' \frac{\partial V^{-1}}{\partial \theta_i} - (I - R)XQX' \frac{\partial V^{-1}}{\partial \theta_i} XQX'V^{-1}.$$

Analizando el segundo sumando, se observa que

$$\begin{aligned} XQX' \frac{\partial V^{-1}}{\partial \theta_i} XQX' &= XQ(X' \frac{\partial V^{-1}}{\partial \theta_i} X)QX' = [O(D^{-1})]_{DR \times p} [O(D)]_{p \times p} [O(D^{-1})]_{p \times DR} \\ &= [O(D^{-1})]_{DR \times DR}, \end{aligned}$$

y consecuentemente

$$(I - R)XQX' \frac{\partial V^{-1}}{\partial \theta_i} XQX'V^{-1} = [O(D^{-1})]_{DR \times DR}.$$

Lema A.1 (Prasad-Rao, 1990). Sean A_1 y A_2 matrices $n \times n$ y sea $y \sim N_n(0, V)$, donde V es definida positiva. Entonces

- (a) $E[y(y'A_s y)y'] = \text{tr}(A_s V)V + 2VA_s V, \quad s = 1, 2,$
- (b) $E[(y'A_1 y)(y'A_2 y)] = 2\text{tr}(A_1 VA_2 V) + \text{tr}(A_1 V)\text{tr}(A_2 V),$
- (c) $E[y(y'A_1 y)(y'A_2 y)y'] = \text{tr}(A_1 V)\text{tr}(A_2 V)V + 2\text{tr}(A_1 V)VA_2 V + 2\text{tr}(A_2 V)VA_1 V$
 $+ 2\text{tr}(A_1 VA_2 V)V + 4VA_1 VA_2 V + 4VA_2 VA_1 V.$

Lema A.2 (Prasad-Rao, 1990). Sean $y \sim N_n(0, V)$, $z_j = \lambda'_j y$ y $q_j = y'A_j y$, $j = 1, \dots, p$, donde λ_j es $n \times 1$ y A_j es $n \times n$. Sean $z = (z_1, \dots, z_p)'$, $q = (q_1, \dots, q_p)'$ con matrices de covarianzas V_z y V_q respectivamente. Entonces

$$\begin{aligned} E[(z'(q - E[q]))^2] &= \text{tr}(V_z V_q) + 4 \sum_{j=1}^p \sum_{i=1}^p \{\lambda'_j VA_j VA_i V \lambda_i + \lambda'_j VA_i VA_j V \lambda_i\}, \\ E[z_i z_j (q_i - E[q_i])(q_j - E[q_j])] &= \lambda'_i E[y(y'A_i y)(y'A_j y)y'] \lambda_j - E[q_i] \lambda'_i E[y(y'A_j y)y'] \lambda_j \\ &\quad - E[q_i] \lambda'_i E[y(y'A_i y)y'] \lambda_j + E[q_i] E[q_j] \lambda'_i V \lambda_j. \end{aligned}$$

Lema A.3 (Prasad-Rao, 1990). Sea

- (a) $V = \text{diag} (V_d),$
 $1 \leq d \leq D$
- (b) $C = \text{diag} [O(D^{-1})]_R + [O(D^{-2})]_{DR},$
 $1 \leq d \leq D$
- (c) $r = \text{col}_{1 \leq d \leq D} \text{col}_{1 \leq j \leq R} [O(D^{-1})],$
- (d) $s_i = \text{col}_{1 \leq d \leq D} \text{col}_{1 \leq j \leq R} \delta_{id} [O(1)],$

donde V_d es una matriz $R \times R$ formada por elementos acotados. Entonces se verifica que

- (e) $VCVCV = [O(D^{-2})],$
- (f) $s'_i \sum s_i = O(1),$

$$(g) (r + s_i)'VCV(r + s_i) = [O(D^{-2})].$$

Lema 3. Sea v un vector aleatorio tal que $v \sim N(0, V)$. Sean $s_1 = \lambda_1'v$ y $s_2 = \lambda_2'v$ dos combinaciones lineales de v . Sean $q_1 = v'A_1v$ y $q_2 = v'A_2v$ dos formas cuadráticas. Entonces

$$E[s_1s_2(q_1 - E[q_1])(q_2 - E[q_2])] = \text{cov}(q_1, q_2)\text{cov}(s_1, s_2) + 8\lambda_1'VA_1VA_2\lambda_2.$$

Demostración. Aplicando el Lema A.2 de Prasad-Rao (1990), se tiene

$$\begin{aligned} E &= E[s_1s_2(q_1 - E[q_1])(q_2 - E[q_2])] = \lambda_1'E[v(v'A_1v)v'A_2v]v'\lambda_2 \\ &- E[q_1]\lambda_1'E[v(v'A_2v)v']\lambda_2 - E[q_2]\lambda_1'E[v(v'A_1v)v']\lambda_2 + E[q_1]E[q_2]\lambda_1'V\lambda_2. \end{aligned}$$

Aplicando el Lema A.1(c) de Prasad-Rao (1990), se obtiene

$$\begin{aligned} E[v(v'A_1v)v'A_2v]v' &= \text{tr}(A_1V)\text{tr}(A_2V)V + 2\text{tr}(A_1V)VA_2V + 2\text{tr}(A_2V)VA_1V \\ &+ 2\text{tr}(A_1VA_2V)V + 4VA_1VA_2V + 4VA_2VA_1V. \end{aligned}$$

Aplicando el Lema A.1(a) de Prasad-Rao (1990), se obtiene

$$E[v(v'A_iv)v'] = \text{tr}(A_iV)V + 2VA_iV, \quad i = 1, 2.$$

Además $E[q_i] = \text{tr}(A_iV)$, $\text{cov}(q_1, q_2) = 2\text{tr}(A_1VA_2V)$ y $\text{cov}(s_1, s_2) = \lambda_1'V\lambda_2$. Sustituyendo se obtiene

$$\begin{aligned} E &= E[q_1]E[q_2]\lambda_1'V\lambda_2 + 2E[q_1]\lambda_1'VA_2V\lambda_2 + 2E[q_2]\lambda_1'VA_1V\lambda_2 + 8\lambda_1'VA_1VA_2V\lambda_2 \\ &+ 2\text{tr}(A_1VA_2V)\lambda_1'V\lambda_2 - E[q_1]E[q_2]\lambda_1'V\lambda_2 - 2E[q_1]\lambda_1'VA_2V\lambda_2 - E[q_1]E[q_2]\lambda_1'V\lambda_2 \\ &- 2E[q_2]\lambda_1'VA_1V\lambda_2 + E[q_1]E[q_2]\lambda_1'V\lambda_2 = 2\text{tr}(A_1VA_2V)\lambda_1'V\lambda_2 + 8\lambda_1'VA_1VA_2V\lambda_2 \\ &= \text{cov}(q_1, q_2)\text{cov}(s_1, s_2) + 8\lambda_1'VA_1VA_2V\lambda_2. \end{aligned}$$

Lema 4. Bajo las condiciones H1-H6, se verifica que

$$\text{cov}(s^{(i)}, s^{(j)}) = L^{(i)}VL^{(j)'} + [O(D^{-1})]_{DR \times DR}.$$

Demostración. Por el Lema 1 se sabe que $s^{(i)} = (L^{(i)} + F^{(i)})v$, donde $v \sim N(0, V)$ y

$$(a) L^{(i)} = \text{diag} (L_d^{(i)}), \text{ con } L_d^{(i)} = [O(1)]_{R \times R}, d = 1, \dots, D,$$

$$(b) F^{(i)} = [O(D^{-1})]_{DR \times DR},$$

y análogamente para $s^{(j)}$. Por otra parte, se tiene que

$$\text{cov}(s^{(i)}, s^{(j)}) = (L^{(i)} + F^{(i)})V(L^{(j)} + F^{(j)})' = L^{(i)}VL^{(j)'} + L^{(i)}VF^{(j)'} + F^{(i)}VL^{(j)'} + F^{(i)}VF^{(j)'}$$

Aplicando (a) y (b) se deduce que

$$(i) F^{(i)}VF^{(j)'} = [O(D^{-1})]_{DR \times DR},$$

$$(ii) L^{(i)}VF^{(j)'} = [O(D^{-1})]_{DR \times DR} \text{ y } F^{(i)}VL^{(j)'} = [O(D^{-1})]_{DR \times DR}.$$

por tanto $\text{cov}(s^{(i)}, s^{(j)}) = L^{(i)}VL^{(j)'} + [O(D^{-1})]_{DR \times DR}$.

Teorema 1. Se supone que el modelo (2.1) verifica H1-H6, y, además que

$$(H_7) (X'V^{-1}X)^{-1} = [O(D^{-1})]_{pr \times pr}, \text{ y}$$

$$(H_8) \hat{\sigma}_u^2 = k + y'Cy \text{ es un estimador de } \sigma_u^2 \text{ insesgado, consistente e invariante por traslaciones, donde } k = O(1) \text{ y } C = \text{diag} \{ [O(D^{-1})]_{r \times r}, \dots, [O(D^{-1})]_{r \times r} \} + [O(D^{-2})]_{Dr \times Dr},$$

Entonces

$$E[(\hat{\theta}_i - \theta_i)(\hat{\theta}_j - \theta_j)s^{(i)}s^{(j)'}] = \text{cov}(\hat{\theta}_i, \hat{\theta}_j)L^{(i)}VL^{(j)'} + [O(D^{-1})]_{DR \times DR}.$$

Demostración. Por el Lema 1, se tiene que las componentes del vector $s^{(i)}$ y $s^{(j)}$, definidos en (2.7), son funciones lineales de $v = Zu + e$. Es decir,

$$s_{dr_1}^{(i)} = (f_{dr_1}^{(i)} + l_{dr_1}^{(i)})'v, \quad r_1 = 1, \dots, R, \quad d = 1, \dots, D,$$

donde

$$s^{(i)} = \begin{pmatrix} s_{11}^{(i)} \\ \vdots \\ s_{DR}^{(i)} \end{pmatrix}, \quad F^{(i)} = \begin{pmatrix} f_{11}^{(i)'} \\ \vdots \\ f_{DR}^{(i)'} \end{pmatrix}, \quad L^{(i)} = \begin{pmatrix} l_{11}^{(i)'} \\ \vdots \\ l_{DR}^{(i)'} \end{pmatrix}$$

y análogamente para $s^{(j)}$. Por la hipótesis H8, se tiene que

$$\hat{\theta}_i = k + y'A_i y, \quad \text{con } E[\hat{\theta}_i] = \theta_i.$$

Como $\hat{\theta}_i$ es invariante por traslaciones, para $v = Zu + e = y - X\beta$ se verifica que

$$\hat{\theta}_i(y) = \hat{\theta}_i(y - X\beta) = \hat{\theta}_i(v).$$

Es decir,

$$\hat{\theta}_i(v) = k + v'A_i v, \quad \text{con } \theta_u = k + E[v'A_i v].$$

Restando, se obtiene

$$\hat{\theta}_i - \theta_i = v'A_i v - E[v'A_i v].$$

Sea $q_i = v'A_i v$, entonces

$$\hat{\theta}_i - \theta_i = q_i - E[q_i],$$

y análogamente para $\hat{\theta}_j$.

Aplicando el Lema 3 con $\lambda_1 = f_{dr_1}^{(i)} + l_{dr_1}^{(i)}$, $\lambda_2 = f_{dr_2}^{(j)} + l_{dr_2}^{(j)}$, $s_1 = s_{dr_1}^{(i)}$, $s_2 = s_{dr_2}^{(j)}$, $q_1 = v'A_i v$, $q_2 = v'A_j v$, y teniendo en cuenta que $v \sim N(0, V)$, se obtiene

$$E[s_{dr_1}^{(i)} s_{dr_2}^{(j)} (\hat{\theta}_i - \theta_i)(\hat{\theta}_j - \theta_j)] = \text{cov}(s_{dr_1}^{(i)}, s_{dr_2}^{(j)}) \text{cov}(\hat{\theta}_i, \hat{\theta}_j) + 8(f_{dr_1}^{(i)} + l_{dr_1}^{(i)})'VA_iVA_jV(f_{dr_2}^{(j)} + l_{dr_2}^{(j)}).$$

En notación matricial, se ha obtenido

$$E[(\hat{\theta}_i - \theta_i)(\hat{\theta}_j - \theta_j)s^{(i)}s^{(j)'}] = \text{cov}(s^{(i)}, s^{(j)})\text{cov}(\hat{\theta}_i, \hat{\theta}_j) + 8(F^{(i)} + L^{(i)})VA_iVA_jV(F^{(j)} + L^{(j)})'.$$

Aplicando el Lema 4 se tiene que

$$\text{cov}(s^{(i)}, s^{(j)}) = L^{(i)}VL^{(j)'} + [O(D^{-1})]_{DR \times DR}.$$

Aplicando el Lema 2 y el Lema A.3 de Prasad y Rao (1990), se demuestra que

$$(F^{(i)} + L^{(i)})VA_iVA_jV(F^{(j)} + L^{(j)})' = [O(D^{-2})]_{DR \times DR}.$$

Consecuentemente, se obtiene

$$E[(\hat{\theta}_i - \theta_i)(\hat{\theta}_j - \theta_j)s^{(i)}s^{(j)'}] = \text{cov}(\hat{\theta}_i, \hat{\theta}_j)L^{(i)}VL^{(j)'} + [o(D^{-1})]_{DR \times DR}.$$

Corolario 1. Bajo las condiciones H1-H8, se verifica que

$$E[(\hat{\mu}_E - \hat{\mu}_B)(\hat{\mu}_E - \hat{\mu}_B)'] \approx E[S(\hat{\theta} - \theta)(\hat{\theta} - \theta)'S'] = G_3(\theta) + [o(D^{-1})]_{DR \times DR},$$

donde

$$G_3(\theta) = \sum_{i=1}^m \sum_{j=1}^m \text{cov}(\hat{\theta}_i, \hat{\theta}_j)L^{(i)}VL^{(j)'}$$

A partir de lo anterior se tiene

$$MSE(\hat{\mu}_E) = G_1(\theta) + G_2(\theta) + G_3(\theta).$$

Además se verifica que

$$E[G_1(\hat{\theta})] \approx G_1(\theta) - G_3(\theta), \quad E[G_2(\hat{\theta})] \approx G_2(\theta), \quad E[G_3(\hat{\theta})] \approx G_3(\theta).$$

Por analogía, y, con el objeto de corregir el sesgo, el estimador que se propone para el error cuadrático medio de $\hat{\mu}_E$ es

$$mse(\hat{\mu}_E) = G_1(\hat{\theta}) + G_2(\hat{\theta}) + 2G_3(\hat{\theta}). \quad (2.8)$$

2.5.1. El caso univariante

Es oportuno indicar que en todo el desarrollo de la sección 2.5 se puede suponer que el modelo (2.1) consta de una única variable de interés ($R = 1$) y que sólo existe un parámetro desconocido en la matriz de varianzas V_u ($m = 1$). Por conveniencia, también se puede suponer que los errores son incorrelados y homocedásticos. Bajo estas condiciones el modelo (2.1) se convierte en el modelo Fay-Herriot descrito en el apéndice A y el error cuadrático medio $MSE(\hat{\mu}_d)$ del EBLUP univariante se puede obtener particularizando los resultados de la sección 2.5.

Para conseguir lo que se acaba de apuntar conviene presentar primero los cambios que se producen en los elementos que intervienen en las expresiones de G_1 , G_2 y G_3 en el caso $R = 1$ y $m = 1$. Se verifica que

$$V_u = \sigma_u^2 I_D \quad V_e = \sigma_e^2 I_D \quad V = (\sigma_u^2 \sigma_e^2) + I_D \quad Z = I_D,$$

y así se obtiene

$$\begin{aligned} T &= V_u - V_u Z' V^{-1} Z V_u = \frac{\sigma_u^2 \sigma_e^2}{(\sigma_u^2 + \sigma_e^2)} I_D, \\ G_1(\sigma_u^2) &= \frac{\sigma_u^2 \sigma_e^2}{(\sigma_u^2 + \sigma_e^2)} I_D, \\ G_2(\sigma_u^2) &= (X - Z T Z' V_e^{-1} X) Q (X' - X' V_e^{-1} Z T Z') = \frac{\sigma_e^4}{(\sigma_u^2 + \sigma_e^2)} X (X' X)^{-1} X'. \end{aligned}$$

Por otra parte, si se omiten los subíndices, ya que sólo hay un parámetro desconocido en la matriz V_u , se tiene que

$$L = (I_D - R) W V^{-1} = \frac{\sigma_e^2}{(\sigma_u^2 + \sigma_e^2)^2} I_D.$$

Así se llega a la expresión

$$G_3(\sigma_u^2) = \text{var}(\hat{\sigma}_u^2) L V L' = \frac{\sigma_e^4}{(\sigma_u^2 + \sigma_e^2)^3} \text{var}(\hat{\sigma}_u^2) I_D.$$

Además también se verifica lo siguiente:

$$E(G_1(\hat{\sigma}_u^2)) \approx G_1(\sigma_u^2) - G_3(\sigma_u^2), \quad E(G_2(\hat{\sigma}_u^2)) \approx G_2(\sigma_u^2), \quad E(G_3(\hat{\sigma}_u^2)) \approx G_3(\sigma_u^2).$$

Finalmente, para cada valor de d se tiene que

$$mse(\hat{\mu}_{Ed}) = G_{1d}(\hat{\sigma}_u^2) + G_{2d}(\hat{\sigma}_u^2) + 2G_{3d}(\hat{\sigma}_u^2),$$

donde

$$G_{1d}(\hat{\sigma}_u^2) = \frac{\hat{\sigma}_u^2 \sigma_e^2}{(\hat{\sigma}_u^2 + \sigma_e^2)}, \quad G_{2d}(\hat{\sigma}_u^2) = \frac{\sigma_e^4}{(\hat{\sigma}_u^2 + \sigma_e^2)} \frac{x_d^2}{\sum_{d=1}^D x_d^2}, \quad G_{3d}(\hat{\sigma}_u^2) = \frac{\sigma_e^4}{(\hat{\sigma}_u^2 + \sigma_e^2)^3} \text{var}(\hat{\sigma}_u^2).$$

La expresión que se acaba de obtener será objeto de estudio en los capítulos sucesivos, ya que servirá para comparar las estimaciones que se obtengan de los modelos multivariantes con las de los modelos univariantes que se derivan de los mismos.

2.6. Estimación bootstrap del MSE

Para estimar $MSE(\hat{\mu}_E)$ se puede utilizar un procedimiento del tipo bootstrap. A continuación se indican, de forma resumida, los pasos a seguir para obtener unas estimaciones alternativas para $MSE(\hat{\mu}_E)$, distintas de la que se ha propuesto en la sección 2.5. En los modelos estudiados en este trabajo se compararan los procedimientos bootstrap de este apartado con el estimador analítico propuesto en la sección 2.5.

Los pasos del algoritmo de bootstrap paramétrico para la estimación del error cuadrático medio son

- B1** Calcular las estimaciones $\hat{\theta}$ y $\hat{\beta}$ para θ y β respectivamente, siguiendo el método de la máxima verosimilitud residual.
- B2** Generar los vectores u_d^* , $d = 1, \dots, D$, utilizando la estimación $\hat{\theta}$ como vector verdadero de parámetros de varianza.
- B3** Generan vectores e_d^* , $d = 1, \dots, D$.
- B4** A partir de los vectores generados en los dos pasos anteriores, construir el modelo bootstrap

$$y^* = X\hat{\beta}_E + Zu^* + e^*.$$

Ahora, se propone una notación que ayuda a presentar con mayor claridad los pasos que siguen. El vector de medias del modelo bootstrap es

$$\mu^* = X\hat{\beta}_E + Zu^*$$

y su estimación BLUP es

$$\hat{\mu}_B^* = X\hat{\beta}_B^* + Z\hat{u}_B^*.$$

Si se sustituye el parámetro verdadero, $\hat{\theta}$, del modelo bootstrap por su estimación obtenida a partir la muestra bootstrap, $\hat{\theta}^*$, entonces se obtiene la estimación EBLUP del vector de medias en el modelo bootstrap; es decir,

$$\hat{\mu}_E^* = X\hat{\beta}_E^* + Z\hat{u}_E^*.$$

El error cuadrático medio de $\hat{\mu}_E^*$ en el modelo bootstrap es

$$MSE_*(\hat{\mu}_E^*) = E_* [(\hat{\mu}_E^* - \mu^*)(\hat{\mu}_E^* - \mu^*)']. \quad (2.9)$$

Finalmente, el estimador bootstrap del error cuadrático medio de $\hat{\mu}_E$ es $mse_*(\hat{\mu}_E) = MSE_*(\hat{\mu}_E^*)$. Puesto que la esperanza (2.9) no es directamente calculable, se aproxima por simulación Monte Carlo. El paso adicional del algoritmo de bootstrap paramétrico es

- B5** Generar a partir del modelo bootstrap B vectores $y^{*(b)}$, $b = 1, \dots, B$. Calcular las medias $\mu_d^{*(b)}$, los EBLUPs $\hat{\mu}_{Ed}^{*(b)}$ y los estimadores

$$mse^{*1}(\hat{\mu}_{Ed}) = \frac{1}{B} \sum_{b=1}^B (\hat{\mu}_{Ed}^{*(b)} - \mu_d^{*(b)})(\hat{\mu}_{Ed}^{*(b)} - \mu_d^{*(b)})'.$$

Como la expresión que no se determina de forma analítica en el estimador de Prasad y Rao es

$$E [(\hat{\mu}_E - \mu_B)(\hat{\mu}_E - \mu_B)']$$

es razonable proponer como segundo estimador bootstrap para el error cuadrático medio de $\hat{\mu}_E$

$$MSE^{*2}(\hat{\mu}_E) = G_1(\hat{\theta}) + G_2(\hat{\theta}) + E_* [(\hat{\mu}_E^* - \hat{\mu}_B^*)(\hat{\mu}_E^* - \hat{\mu}_B^*)'],$$

que también se aproxima por simulación Monte Carlo, dando lugar al estimador

$$mse^{*2}(\hat{\mu}_E) = G_1(\hat{\theta}) + G_2(\hat{\theta}) + \frac{1}{B} \sum_{b=1}^B (\hat{\mu}_E^{*(b)} - \hat{\mu}_B^{*(b)})(\hat{\mu}_E^{*(b)} - \hat{\mu}_B^{*(b)})'.$$

Dado que el valor esperado del término $G_1(\hat{\theta})$ es, de forma aproximada, $G_1(\theta) - G_3(\theta)$, entonces los estimadores bootstrap mse^{*1} y mse^{*2} pueden presentar sesgo. Con objeto de corregir este sesgo Pfefferman and Tiller (2005) propusieron el estimador corregido

$$MSE^{*3}(\hat{\mu}_E) = 2[G_1(\hat{\theta}) + G_2(\hat{\theta})] - E_*[G_1(\hat{\theta}^*) + G_2(\hat{\theta}^*)] + E_*[(\hat{\mu}_E^* - \hat{\mu}_B^*)(\hat{\mu}_E^* - \hat{\mu}_B^*)'],$$

que se aproxima por simulación Monte Carlo mediante la expresión siguiente:

$$mse^{*3}(\hat{\mu}_E) = 2(G_1(\hat{\theta}) + G_2(\hat{\theta})) - \frac{1}{B} \sum_{b=1}^B (G_1(\hat{\theta}^{*(b)}) + G_2(\hat{\theta}^{*(b)})) + \sum_{b=1}^B (\hat{\mu}_E^{*(b)} - \hat{\mu}_B^{*(b)})(\hat{\mu}_E^{*(b)} - \hat{\mu}_B^{*(b)})'.$$





3

Modelo diagonal

3.1. Definición del modelo

En el capítulo 2 se introduce un modelo multivariante lineal mixto que contempla un vector de errores e con componentes independientes. El modelo (2.1) no se especifica completamente debido a que se dejar libre la estructura de correlación del vector de efectos aleatorios. En este capítulo se contempla una posibilidad para tal estructura considerando que la matriz de covarianzas del vector u es diagonal. En primer lugar, se expone la representación completa del modelo. El *modelo diagonal* es

$$y = X\beta + Zu + e = X\beta + Z_1u_1 + \cdots + Z_Du_D + e, \quad (3.1)$$

donde e, u_1, \dots, u_D son independientes con distribuciones $e \sim N(0, V_e)$, $u \sim N(0, V_u)$, y $u_d \sim N(0, V_{ud})$, $d = 1, \dots, D$. Se supone que V_e es una matriz conocida y que

$$V_{ud} = \begin{pmatrix} \sigma_{u1}^2 & \cdots & 0 \\ \vdots & & \vdots \\ 0 & \cdots & \sigma_{uR}^2 \end{pmatrix} = \text{diag}(\sigma_{ur}^2)_{1 \leq r \leq R}.$$

En la notación del capítulo 2 se tiene que $m = R$ y $\theta = (\theta_1 = \sigma_{u1}^2, \dots, \theta_R = \sigma_{uR}^2)$.

Las derivadas de la matriz V_{ud} respecto de los parámetros de varianza aparecen en los vectores y matrices, $S(\theta)$ y $F(\theta)$, de la ecuación de actualización del algoritmo Fisher-scoring para el cálculo del estimador REML de θ . Las derivadas son

$$V_r = \frac{\partial V}{\partial \theta_r} = \frac{\partial V}{\partial \sigma_{ur}^2} = \delta_r \delta_r', \quad r = 1, \dots, R,$$

donde δ_r se definió en (2.2). La distribución asintótica del estimador REML de θ es $\hat{\theta} \sim N_R(\theta, F^{-1}(\theta))$, con lo cual

$$\hat{\sigma}_{ur_1}^2 - \hat{\sigma}_{ur_2}^2 \sim N(\sigma_{ur_1}^2 - \sigma_{ur_2}^2, \mathbf{v}_{r_1 r_1} + \mathbf{v}_{r_2 r_2} - 2\mathbf{v}_{r_1 r_2}), \quad (3.2)$$

donde v_{ij} es el término correspondiente a la fila i y columna j de la matriz $F^{-1}(\theta)$. La distribución asintótica (3.2) permite realizar el contraste $H_0 : \sigma_{ur_1}^2 = \sigma_{ur_2}^2$ para comprobar si hay diferencias significativas entre las varianzas de los efectos aleatorios u_{dr_1} y u_{dr_2} . Para un nivel de significación es α , se rechaza H_0 si

$$\frac{\hat{\sigma}_{u1}^2 - \hat{\sigma}_{u2}^2}{\sqrt{\hat{v}_{11} + \hat{v}_{22} - 2\hat{v}_{12}}} \notin (-z_{\alpha/2}, z_{\alpha/2}).$$

3.2. Experimentos de simulación

Para estudiar empíricamente el comportamiento de los algoritmos de ajuste y de los procedimientos de estimación del error cuadrático medio de los EBLUPs, en esta sección se presentan tres experimentos de simulación. En las simulaciones se compara el modelo diagonal (3.1) con el modelo con errores e_{dr} independientes. Este último modelo prescinde de toda estructura multivariante y equivale a aplicar por separado R modelos Fay-Herriot; es decir, uno por cada componente r , $r = 1, \dots, R$.

En las simulaciones, se ha programado un modelo diagonal (3.1) bivariante ($R = 2$) cuyas características se describen a continuación. La matriz de covarianzas del vector u_d es

$$V_{ud} = \begin{pmatrix} \sigma_{u1}^2 & 0 \\ 0 & \sigma_{u2}^2 \end{pmatrix}, \quad \sigma_{u1}^2 = 2, \quad \sigma_{u2}^2 = 4.$$

Las componentes del vector e_d verifican $\text{var}(e_{d1}) = 1$, $\text{var}(e_{d2}) = 2$ y $\text{corr}(e_{d1}, e_{d2}) = \rho_e$. La matriz de covarianzas del vector e_d es

$$V_{ed} = \begin{pmatrix} \sigma_{d11} & \sigma_{d12} \\ \sigma_{d21} & \sigma_{d22} \end{pmatrix}, \quad \sigma_{d11} = 1, \quad \sigma_{d22} = 2, \quad \sigma_{d12} = \sigma_{d21} = \rho_e \sqrt{\sigma_{d11} \sigma_{d22}}.$$

La matriz de covarianzas del vector y_d es

$$\begin{aligned} V_d &= V_{ud} + V_{ed} = \begin{pmatrix} \sigma_{u1}^2 & 0 \\ 0 & \sigma_{u2}^2 \end{pmatrix} + \begin{pmatrix} \sigma_{d11} & \sigma_{d12} \\ \sigma_{d21} & \sigma_{d22} \end{pmatrix} = \begin{pmatrix} 2 & 0 \\ 0 & 4 \end{pmatrix} + \begin{pmatrix} 1 & \rho_e \sqrt{2} \\ \rho_e \sqrt{2} & 2 \end{pmatrix} \\ &= \begin{pmatrix} 3 & \rho_e \sqrt{2} \\ \rho_e \sqrt{2} & 6 \end{pmatrix}, \quad d = 1, \dots, D. \end{aligned}$$

Finalmente, la matriz de covarianzas del vector y es

$$V = \text{diag}_{1 \leq d \leq D} \begin{pmatrix} 3 & \rho_e \sqrt{2} \\ \rho_e \sqrt{2} & 6 \end{pmatrix}.$$

Las derivadas parciales que se utilizan en el algoritmo de Fisher-scoring son

$$V_1 = \text{diag}_{1 \leq d \leq D} (V_{d1}) \quad \text{y} \quad V_2 = \text{diag}_{1 \leq d \leq D} (V_{d2}),$$

donde

$$V_{d1} = \frac{\partial V_{ud}}{\partial \sigma_{u1}^2} = \begin{pmatrix} 1 & 0 \\ 0 & 0 \end{pmatrix} \quad \text{y} \quad V_{d2} = \frac{\partial V_{ud}}{\partial \sigma_{u2}^2} = \begin{pmatrix} 0 & 0 \\ 0 & 1 \end{pmatrix}.$$

Los valores de los parámetros de regresión son $\beta_1 = 1$, $\beta_2 = 1$. Para estimar los parámetros σ_{u1}^2 y σ_{u2}^2 mediante el algoritmo Fisher-scoring se utilizan como valores iniciales (semillas) los verdaderos valores; es decir, $\sigma_{u1}^2 = 2$ y $\sigma_{u2}^2 = 4$.

Como se van a realizar comparaciones entre el modelo multivariante y los modelos univariantes marginales, a continuación se da la descripción de tales modelos así como los parámetros que intervienen en los mismos. Los modelos univariantes son

$$y_{dr} = x_{dr}\beta_r + u_{dr} + e_{dr}, \quad d = 1, \dots, D, \quad r = 1, 2,$$

donde $u_{dr} \sim N(0, \sigma_{ur}^2)$ y $e_{dr} \sim N(0, \sigma_{dr}^2)$ son independientes. Los parámetros de los modelos univariantes son los mismos que se usan en el modelo multivariante diagonal; es decir $\beta_1 = 1$, $\beta_2 = 1$, $\sigma_{u1}^2 = 2$, $\sigma_{u2}^2 = 4$, $\sigma_{d11} = 1$ y $\sigma_{d22} = 2$.

Para ambos modelos se utilizan las variables explicativas

$$x_{d1} = \mu_1 + \sigma_{x11}^{1/2}U_{d1}, \quad x_{d2} = \mu_2 + \sigma_{x22}^{1/2}(\rho_x U_{d1} + (1 - \rho_x^2)^{1/2}U_{d2}), \quad d = 1, \dots, D,$$

donde $\mu_1 = \mu_2 = 10$, $\sigma_{x11} = 1$, $\sigma_{x22} = 2$, $\rho_x = \frac{1}{2}$ y

$$U_{dr} = \frac{d-D}{D} + \frac{r}{3}, \quad r = 1, 2, \quad d = 1, \dots, D.$$

3.2.1. Experimento de simulación 1

El objetivo de este experimento es investigar empíricamente la pérdida de eficiencia en las estimaciones cuando no se tiene en cuenta la naturaleza multivariante de los datos. El objeto principal consiste en estudiar lo que ocurre cuando se asume incorrectamente los modelos marginales independientes, en lugar del modelo multivariante subyacente. Para ello, se simulan los datos o bien del modelo multivariante o bien del modelo producto de marginales, se estiman los parámetros y se calculan los EBLUP de ambos modelos. Hay dos conjuntos de parámetros estimados y EBLUP calculados. Según sea el caso, un conjunto se obtiene bajo el modelo correcto y otro bajo el modelo incorrecto. Cabe esperar que los mejores resultados se obtengan siempre cuando se usan los estimadores correspondientes al modelo correcto.

El experimento consiste en calcular por simulación Monte Carlo los sesgos y errores cuadráticos medios empíricos (BIAS y MSE) de los estimadores de los parámetros y de los EBLUP.

Para los valores $\rho_e = 0, 1/4, 1/2, 3/4$, los pasos del experimento de simulación son

1. Repetir $I = 10^4$ veces ($i = 1, \dots, I$)
 - 1.1. Generar una muestra (y_{dr}, x_{dr}) , $d = 1, \dots, D$, $r = 1, 2$.
 - 1.2. Calcular $\hat{\beta}_{Er}^{(i,0)}$, $\hat{\sigma}_{ur}^{2(i,0)}$, $\hat{\beta}_{Er}^{(i,1)}$, $\hat{\sigma}_{ur}^{2(i,1)}$, $r = 1, 2$, y $\hat{\mu}_d^{(i,0)}$, $\hat{\mu}_d^{(i,1)}$, donde los superíndices “0” y “1” se usan para denotar los estimadores y EBLUPs calculados asumiendo el modelo producto de marginales ($a = 0$) o el modelo multivariante diagonal ($a = 1$).

2. Salida:

$$MSE(\hat{\beta}_{Er}^{(a)}) = \frac{1}{I} \sum_{i=1}^I (\hat{\beta}_{Er}^{(i,a)} - \beta_r)^2, \quad MSE(\hat{\sigma}_{ur}^{2(a)}) = \frac{1}{I} \sum_{i=1}^I (\hat{\sigma}_{ur}^{2(i,a)} - \sigma_{ur}^2)^2, \quad r = 1, 2, a = 0, 1.$$

$$BIAS(\hat{\beta}_{Er}^{(a)}) = \frac{1}{I} \sum_{i=1}^I (\hat{\beta}_{Er}^{(i,a)} - \beta_r), \quad BIAS(\hat{\sigma}_{ur}^{2(a)}) = \frac{1}{I} \sum_{i=1}^I (\hat{\sigma}_{ur}^{2(i,a)} - \sigma_{ur}^2), \quad r = 1, 2, a = 0, 1.$$

$$MSE_{drr}^{(a)} = \frac{1}{I} \sum_{i=1}^I (\hat{\mu}_{Edr}^{(i,a)} - \mu_{dr}^{(i)})^2, \quad BIAS_{drr}^{(a)} = \frac{1}{I} \sum_{i=1}^I (\hat{\mu}_{Edr}^{(i,a)} - \mu_{dr}^{(i)}), \quad d = 1, D/2, D, r = 1, 2, a = 0, 1.$$

La tabla 3.1 presenta los errores cuadráticos medios y sesgos empíricos de los estimadores REML de los parámetros de los dos modelos considerados. La tabla está ordenada por columnas y filas. La primera columna da el valor de ρ_e , la segunda columna especifica el estimador del parámetro del modelo, las cuatro columnas siguientes muestran el error cuadrático medio y las cuatro últimas el sesgo. Cada uno de los dos grupos de cuatro columnas que se acaban de nombrar se corresponde con un valor distinto del número de áreas; es decir, $D = 50, 100, 200, 400$. Las filas se disponen en grupos de ocho, un grupo para cada valor de ρ_e . En el caso $\rho_e = 0$, los datos se simulan de los modelos univariantes independientes ($a = 0$) que también llamaremos modelo producto de univariantes. Por parte, en los casos $\rho_e > 0$ los datos se generan del modelo multivariante diagonal ($a = 1$). Dentro de cada grupo, las cuatro primeras se corresponden con las estimaciones para el modelo $a = 0$ y las cuatro últimas con las estimaciones para el modelo $a = 1$.

La tabla 3.1 muestra que los errores cuadráticos medios y sesgos de $\hat{\beta}_{E1}$ y $\hat{\beta}_{E2}$ son básicamente iguales para ambos modelos, $a = 0, 1$, independientemente del valor de ρ_e usado en la simulación de los datos. Para los estimadores de las varianzas, se observa que los errores cuadráticos medios son menores en el modelo multivariante ($a = 1$) conforme aumenta el valor de ρ_e .

Las tablas 3.2 y 3.3 presentan los errores cuadráticos medios y los sesgos, $MSE_{drr}^{(a)}$ y $BIAS_{drr}^{(a)}$, $a = 0, 1$, de los EBLUPS de las medias de las componentes $r = 1$ y $r = 2$ respectivamente. En ambas tablas se disponen las columnas de forma análoga a la tabla 3.1. La diferencia está en la segunda columna donde aparecen los valores de las áreas consideradas, $d = 1, d = D/2$ y $d = D$, y el valor medio a lo largo de ellas. El resto de columnas están estructuradas de la misma forma que en la tabla 3.1, pero el error cuadrático medio y el sesgo son de los estimadores $\hat{\mu}_{d1}$ en la tabla 3.2 y $\hat{\mu}_{d2}$ en la tabla 3.3.

En las tablas 3.2 y 3.3 se observa que conforme aumenta el valor de ρ_e los errores cuadráticos medios de los estimadores de $\hat{\mu}_{d1}$ y $\hat{\mu}_{d2}$ son menores en el modelo multivariante ($a = 1$). También conviene destacar el hecho de que la utilización del modelo multivariante cuando el modelo correcto es el producto de univariantes ($\rho_e = 0$) no aumenta el error cuadrático medio de los EBLUPS.

La figura 3.1 muestra las gráficas de los valores de $MSE_{drr}^{(a)}$, $a = 0, 1, r = 1, 2, d = 1, \dots, D, D = 100$. La figura está dividida en 4 partes y tiene una disposición en forma de tabla con dos filas y dos columnas. Las filas 1 y 2 presentan los valores de $MSE_{drr}^{(a)}$ para $r = 1$ y $r = 2$ respectivamente. Las columnas 1 y 2 presentan los valores de $MSE_{drr}^{(a)}$ cuando los datos se generan del modelo con $\rho_e = 0$ y $\rho_e = \frac{3}{4}$ respectivamente. Cada una de las 4 sub-figuras muestran los valores de $MSE_{drr}^{(a)}$ para $a = 0$ y $a = 1$.

ρ_e	D	MSE				BIAS			
		50	100	200	400	50	100	200	40
0	$\hat{\beta}_{E1}^{(0)}$	0.00060	0.00031	0.00015	0.00008	0.00000	0.00018	0.00014	-0.00005
	$\hat{\beta}_{E2}^{(0)}$	0.00112	0.00058	0.00029	0.00014	-0.00020	-0.00001	-0.00017	0.00003
	$\hat{\sigma}_{u1}^{2(0)}$	0.36575	0.18509	0.09067	0.04597	-0.00733	0.00084	0.00054	0.00039
	$\hat{\sigma}_{u2}^{2(0)}$	1.46395	0.72318	0.37740	0.18192	-0.01601	-0.00668	-0.00018	-0.00122
	$\hat{\beta}_{E1}^{(1)}$	0.00060	0.00031	0.00015	0.00008	0.00000	0.00018	0.00014	-0.00005
	$\hat{\beta}_{E2}^{(1)}$	0.00112	0.00058	0.00029	0.00014	-0.00020	-0.00001	-0.00017	-0.00003
	$\hat{\sigma}_{u1}^{2(1)}$	0.36575	0.18509	0.09067	0.04597	-0.00733	0.00084	0.00054	0.00039
	$\hat{\sigma}_{u2}^{2(1)}$	1.46395	0.72318	0.37740	0.18192	-0.01601	-0.00668	-0.00018	-0.00122
$\frac{1}{4}$	$\hat{\beta}_{E1}^{(0)}$	0.00062	0.00031	0.00016	0.00008	0.00001	-0.00035	-0.00015	0.00016
	$\hat{\beta}_{E2}^{(0)}$	0.00114	0.00058	0.00029	0.00014	0.00020	0.00027	-0.00038	0.00002
	$\hat{\sigma}_{u1}^{2(0)}$	0.36422	0.18128	0.09112	0.04504	-0.00013	-0.00343	-0.00046	0.00127
	$\hat{\sigma}_{u2}^{2(0)}$	1.48842	0.71515	0.37214	0.18704	-0.03230	0.00092	0.01069	-0.00447
	$\hat{\beta}_{E1}^{(1)}$	0.00062	0.00031	0.00016	0.00008	0.00001	-0.00035	-0.00015	0.00016
	$\hat{\beta}_{E2}^{(1)}$	0.00114	0.00058	0.00029	0.00014	0.00020	0.00027	-0.00038	0.00002
	$\hat{\sigma}_{u1}^{2(1)}$	0.35873	0.17784	0.08999	0.04454	-0.00086	-0.00306	-0.00090	0.00126
	$\hat{\sigma}_{u2}^{2(1)}$	1.46969	0.70686	0.36632	0.18448	-0.03346	0.00214	0.00969	-0.00452
$\frac{1}{2}$	$\hat{\beta}_{E1}^{(0)}$	0.00061	0.00031	0.00016	0.00008	0.00004	0.00032	0.00001	0.00003
	$\hat{\beta}_{E2}^{(0)}$	0.00110	0.00056	0.00029	0.00014	-0.00033	0.00003	0.00000	-0.00003
	$\hat{\sigma}_{u1}^{2(0)}$	0.36673	0.18287	0.09074	0.04465	-0.00481	-0.01045	-0.00293	0.00031
	$\hat{\sigma}_{u2}^{2(0)}$	1.45802	0.73550	0.36021	0.18211	0.00826	0.00847	-0.00575	0.00028
	$\hat{\beta}_{E1}^{(1)}$	0.00061	0.00031	0.00016	0.00008	0.00004	0.00032	0.00001	0.00003
	$\hat{\beta}_{E2}^{(1)}$	0.00110	0.00056	0.00029	0.00014	-0.00033	0.00003	0.00000	-0.00003
	$\hat{\sigma}_{u1}^{2(1)}$	0.34607	0.17345	0.08581	0.04203	-0.00456	-0.01118	0.00202	0.00122
	$\hat{\sigma}_{u2}^{2(1)}$	1.38517	0.70008	0.34176	0.17214	0.00940	0.00728	-0.00381	0.00214
$\frac{3}{4}$	$\hat{\beta}_{E1}^{(0)}$	0.00063	0.00031	0.00015	0.00008	-0.00006	0.00009	0.00017	-0.00001
	$\hat{\beta}_{E2}^{(0)}$	0.00113	0.00058	0.00029	0.00014	0.00030	-0.00011	0.00015	-0.00012
	$\hat{\sigma}_{u1}^{2(0)}$	0.37778	0.18409	0.09202	0.04403	-0.00408	0.00136	0.00294	0.00583
	$\hat{\sigma}_{u2}^{2(0)}$	1.48387	0.75015	0.36359	0.18147	0.01303	0.00668	-0.00616	0.00886
	$\hat{\beta}_{E1}^{(1)}$	0.00063	0.00031	0.00015	0.00008	-0.00006	0.00009	0.00017	-0.00001
	$\hat{\beta}_{E2}^{(1)}$	0.00113	0.00058	0.00029	0.00014	0.00030	-0.00011	0.00015	-0.00012
	$\hat{\sigma}_{u1}^{2(1)}$	0.33474	0.16081	0.08233	0.03885	-0.00195	0.00098	0.00108	0.00361
	$\hat{\sigma}_{u2}^{2(1)}$	1.30486	0.65723	0.32148	0.15925	0.01673	0.00596	-0.01017	0.00428

Tabla 3.1: MSE (izquierda) y BIAS (derecha) para $\rho_x = 1/2$.

ρ_e	a	d	MSE				BIAS			
			50	100	200	400	50	100	200	400
0	0	1	0.6739	0.6936	0.6772	0.6631	-0.0097	-0.0032	-0.0032	-0.0058
		$D/2$	0.6789	0.6798	0.6672	0.6737	0.0015	0.0094	-0.0051	0.0084
		D	0.6685	0.6785	0.6798	0.6810	0.0036	0.0071	0.0003	0.0031
		mean	0.6862	0.6789	0.6728	0.6693	-0.0008	0.0018	-0.0003	0.0003
	1	1	0.6739	0.6936	0.6772	0.6631	-0.0097	-0.0032	-0.0032	-0.0058
		$D/2$	0.6789	0.6798	0.6672	0.6737	0.0015	0.0094	-0.0051	0.0084
		D	0.6685	0.6785	0.6798	0.6810	0.0036	0.0071	0.0003	0.0031
		mean	0.6862	0.6789	0.6728	0.6693	-0.0008	0.0018	-0.0003	0.0003
$\frac{1}{4}$	0	1	0.6781	0.6642	0.6742	0.6629	0.0127	-0.0242	0.0128	0.0013
		$D/2$	0.6942	0.6735	0.6612	0.6545	0.0056	-0.0102	0.0147	-0.0061
		D	0.7048	0.6700	0.6727	0.6690	-0.0007	-0.0067	0.0054	-0.0103
		mean	0.6873	0.6762	0.6730	0.6692	0.0029	-0.0008	0.0008	-0.0005
	1	1	0.6707	0.6559	0.6640	0.6548	0.0106	-0.0255	0.0148	0.0018
		$D/2$	0.6848	0.6631	0.6499	0.6450	0.0053	-0.0094	0.0143	-0.0062
		D	0.6961	0.6586	0.6621	0.6584	-0.0023	-0.0049	0.0048	-0.0100
		mean	0.6777	0.6670	0.6636	0.6599	0.0029	-0.0008	0.0008	-0.0005
$\frac{1}{2}$	0	1	0.6719	0.6667	0.6623	0.6768	0.0040	-0.0073	0.0097	-0.0124
		$D/2$	0.6990	0.6741	0.6877	0.6703	-0.0002	-0.0058	0.0102	0.0021
		D	0.6905	0.6836	0.6799	0.6754	0.0018	-0.0180	0.0082	-0.0073
		mean	0.6840	0.6769	0.6712	0.6694	0.0021	-0.0002	0.0006	0.0007
	1	1	0.6292	0.6292	0.6285	0.6382	0.0046	-0.0069	0.0117	-0.0127
		$D/2$	0.6568	0.6393	0.6438	0.6329	-0.0009	-0.0041	0.0077	0.0043
		D	0.6522	0.6427	0.6441	0.6352	-0.0001	-0.0162	0.0057	-0.0060
		mean	0.6469	0.6396	0.6334	0.6314	0.0021	-0.0002	0.0006	0.0007
$\frac{3}{4}$	0	1	0.6994	0.6786	0.6519	0.6717	0.0045	-0.0203	-0.0035	-0.0068
		$D/2$	0.6826	0.6609	0.6705	0.6683	-0.0104	0.0029	-0.0024	0.0005
		D	0.6947	0.6673	0.6766	0.6653	-0.0061	-0.0088	0.0043	0.0064
		mean	0.6895	0.6775	0.6723	0.6701	0.0011	-0.0011	0.0002	-0.0001
	1	1	0.6080	0.5939	0.5700	0.5881	0.0019	-0.0152	-0.0039	-0.0075
		$D/2$	0.6066	0.5749	0.5895	0.5781	-0.0101	0.0085	-0.0046	0.0028
		D	0.6072	0.5803	0.5931	0.5745	-0.0055	-0.0063	-0.0006	0.0052
		mean	0.6025	0.5898	0.5840	0.5809	0.0011	-0.0011	0.0002	-0.0001

Tabla 3.2: $MSE_{11d}^{(a)}$ (izquierda) y $BIAS_{11d}^{(a)}$ (derecha) para $\rho_x = 1/2$, $a = 0, 1$.

ρ_e	a	d	MSE				BIAS			
			50	100	200	400	50	100	200	400
0	0	1	1.3817	1.3538	1.3734	1.3323	-0.0098	0.0275	0.0108	-0.0162
		$D/2$	1.3981	1.3708	1.3529	1.3434	-0.0034	0.0003	-0.0127	0.0035
		D	1.3673	1.3268	1.3652	1.3413	0.0158	-0.0186	0.0034	-0.0077
		mean	1.3786	1.3577	1.3435	1.3367	-0.0024	-0.0001	-0.0011	0.0007
	1	1	1.3817	1.3538	1.3734	1.3323	-0.0098	0.0275	0.0108	-0.0162
		$D/2$	1.3981	1.3708	1.3529	1.3434	-0.0034	0.0003	-0.0127	0.0035
		D	1.3673	1.3268	1.3652	1.3413	0.0158	-0.0186	0.0034	-0.0077
		mean	1.3786	1.3577	1.3435	1.3367	-0.0024	-0.0001	-0.0011	0.0007
$\frac{1}{4}$	0	1	1.3880	1.3655	1.3542	1.3647	-0.0175	-0.0238	0.0113	-0.0031
		$D/2$	1.3613	1.3201	1.3397	1.3304	-0.0013	-0.0074	0.0109	0.0026
		D	1.3472	1.3578	1.3371	1.3387	0.0018	-0.0134	0.0023	0.0220
		mean	1.3763	1.3507	1.3436	1.3372	-0.0012	-0.0014	-0.0010	-0.0001
	1	1	1.3683	1.3491	1.3439	1.3381	-0.0170	-0.0226	0.0110	-0.0022
		$D/2$	1.3417	1.3063	1.3173	1.3104	0.0005	-0.0072	0.0085	0.0013
		D	1.3307	1.3357	1.3153	1.3225	0.0021	-0.0103	0.0006	0.0197
		mean	1.3573	1.3321	1.3249	1.3188	-0.0012	-0.0014	-0.0010	-0.0001
$\frac{1}{2}$	0	1	1.3528	1.3207	1.3014	1.3353	-0.0206	-0.0018	-0.0034	-0.0118
		$D/2$	1.3404	1.3634	1.3506	1.3246	-0.0038	-0.0030	-0.0235	-0.0163
		D	1.3837	1.3554	1.3317	1.3201	0.0066	-0.0129	0.0031	-0.0036
		mean	1.3745	1.3538	1.3412	1.3385	-0.0024	0.0003	0.0003	-0.0007
	1	1	1.2724	1.2643	1.2251	1.2507	-0.0183	-0.0065	-0.0053	-0.0117
		$D/2$	1.2694	1.2795	1.2794	1.2450	-0.0078	-0.0049	-0.0208	-0.0193
		D	1.3128	1.2646	1.2598	1.2505	0.0054	-0.0149	0.0066	-0.0009
		mean	1.2994	1.2782	1.2658	1.2627	-0.0024	0.0003	0.0003	-0.0007
$\frac{3}{4}$	0	1	1.3540	1.3679	1.3550	1.3493	0.0176	-0.0076	-0.0004	0.0023
		$D/2$	1.3481	1.3382	1.3557	1.3422	0.0139	0.0232	0.0023	-0.0117
		D	1.3724	1.3755	1.3984	1.3083	0.0209	0.0090	-0.0096	-0.0070
		mean	1.3781	1.3527	1.3457	1.3387	0.0047	0.0011	0.0003	-0.0012
	1	1	1.1845	1.1743	1.1830	1.1702	0.0121	0.0023	-0.0006	0.0054
		$D/2$	1.1880	1.1596	1.1635	1.1552	0.0204	0.0223	0.0038	-0.0100
		D	1.1990	1.1880	1.2062	1.1384	0.0217	0.0072	-0.0099	-0.0162
		mean	1.2025	1.1768	1.1675	1.1601	0.0047	0.0011	0.0003	-0.0012

Tabla 3.3: $MSE_{22d}^{(a)}$ (izquierda) y $BIAS_{22d}^{(a)}$ (derecha) para $\rho_x = 1/2$, $a = 0, 1$.

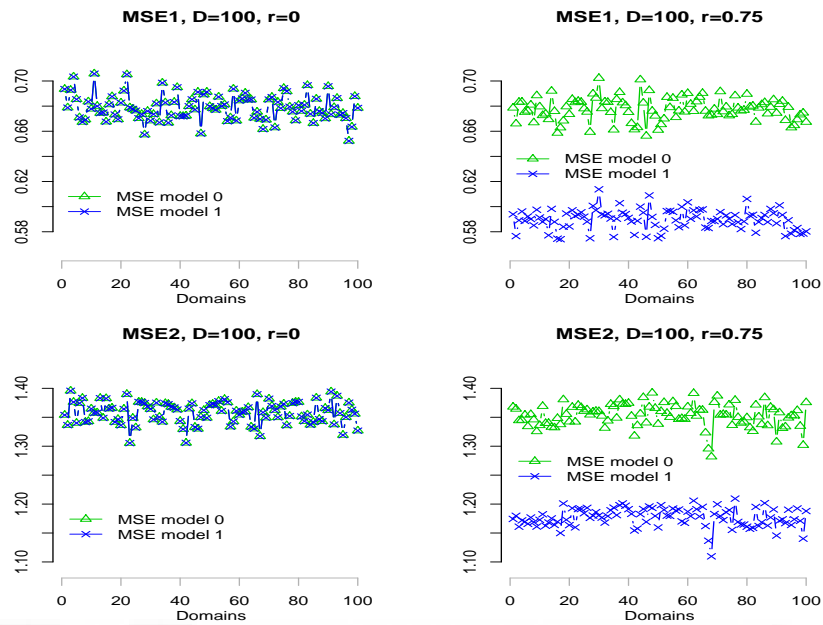


Figura 3.1: MSE_{drr} , para $a = 0, 1$, $r = 1, 2$, $\rho_e = 0, 3/4$, $\rho_x = 1/2$, $D = 100$.

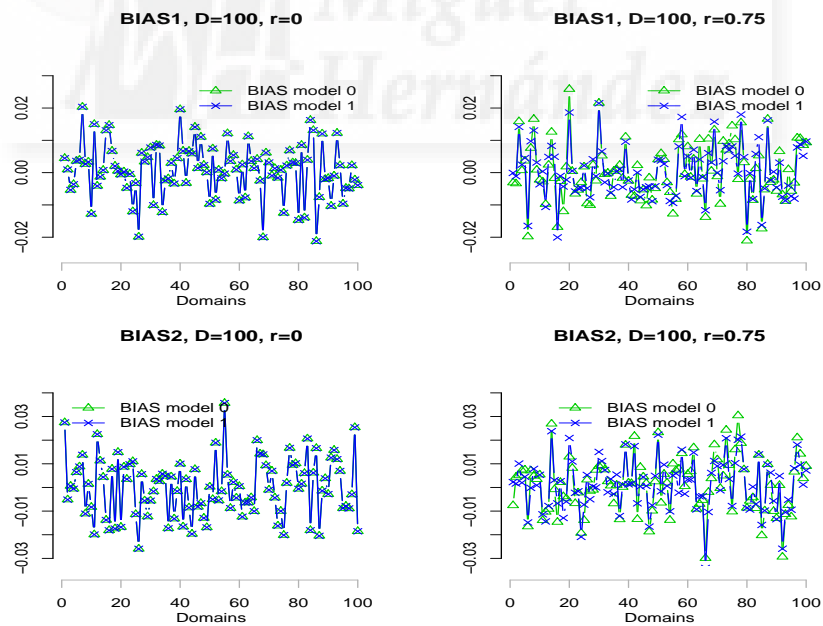


Figura 3.2: $BIAS_{drr}$, para $a = 0, 1$, $r = 1, 2$, $\rho_e = 0, 3/4$, $\rho_x = 1/2$, $D = 100$.

En la figura 3.1 se observa que la diferencia de los errores cuadráticos medios entre el modelo univariante ($a = 0$) y el modelo multivariante ($a = 1$) cuando ρ_e cambia de 0 a 0.75 es bastante pronunciada. También hay que destacar que si los datos se generan del modelo $a = 0$ con $\rho_e = 0$, entonces no se aprecia un aumento del error cuadrático medio al utilizar el modelo multivariante $a = 1$.

La figura 3.2 muestra las gráficas los valores de $BIAS_{drr}^{(a)}$, $a = 0, 1$, $r = 1, 2$, $d = 1, \dots, D$, $D = 100$. La figura está estructurada de la misma forma que la figura 3.1. En la figura 3.2 se observa una leve diferencia en los sesgos del modelo univariante ($a = 0$) y del modelo multivariante ($a = 1$) cuando ρ_e cambia de 0 a 0.75. También hay que destacar que si los datos se generan del modelo $a = 0$ con $\rho_e = 0$, entonces no se aprecia un aumento de sesgo al utilizar el modelo multivariante $a = 1$.

3.2.2. Experimento de simulación 2

El objetivo de este experimento es investigar empíricamente la pérdida de eficiencia en las estimaciones cuando no se tiene en cuenta la naturaleza multivariante de los datos. Para ello se simulan los datos o bien del modelo multivariante ($a = 1$) o bien del modelo producto de marginales ($a = 0$), se estiman los parámetros y se calculan los EBLUP de ambos modelos. El experimento consiste en calcular por simulación Monte Carlo los sesgos y errores cuadráticos medios empíricos (B y E) de los estimadores analíticos (2.8) en la estimación de los errores cuadráticos medios del EBLUP de μ_{dr} .

En esta simulación se consideran los valores $\rho_e = 0, 1/4, 1/2, 3/4$. El resto de parámetros son los mismos que en la simulación 1; es decir, $\beta_{E1} = 1$, $\beta_{E2} = 1$, $\sigma_{u1}^2 = 2$ y $\sigma_{u1}^2 = 4$, $\sigma_{d11} = 1$, $\sigma_{d22} = 2$. Las variables auxiliares x_{dr} también se generan de la misma forma que en la simulación 1. Los pasos del experimento de simulación son

1. Repetir $I = 500$ veces ($i = 1, \dots, 500$)
 - 1.1. Generar una muestra $(y_{dr}^{(i)}, x_{dr}^{(i)})$, $d = 1, \dots, D$, $r = 1, 2$.
 - 1.2. Calcular $\hat{\sigma}_{u1}^{2(i,a)}$, $\hat{\sigma}_{u2}^{2(i)}$, $\hat{\beta}_1^{(i,a)}$ y $\hat{\beta}_2^{(i,a)}$, $a = 0, 1$.
 - 1.3. Para $d = 1, \dots, D$, $a = 0, 1$, $r = 1, 2$, calcular

$$mse_{drr}^{(i,a)} = g_{1drr}^{(i,a)}(\hat{\sigma}_{u1}^{2(i,a)}, \hat{\sigma}_{u2}^{2(i,a)}) + g_{2drr}^{(i,a)}(\hat{\sigma}_{u1}^{2(i,a)}, \hat{\sigma}_{u2}^{2(i,a)}) + 2g_{3drr}^{(i,a)}(\hat{\sigma}_{u1}^{2(i,a)}, \hat{\sigma}_{u2}^{2(i,a)}).$$

2. Leer los valores $MSE_{drr}^{(a)}$ obtenidos en la simulación 1.

3. Salida:

$$B_{drr}^{(a)} = \frac{1}{I} \sum_{i=1}^I (mse_{drr}^{(i,a)} - MSE_{drr}^{(a)}), \quad E_{drr}^{(a)} = \frac{1}{I} \sum_{i=1}^I (mse_{drr}^{(i,a)} - MSE_{drr}^{(a)})^2, \quad r = 1, 2, a = 0, 1.$$

Las tablas 3.4 y 3.5 presentan los errores cuadráticos medios y los sesgos, $E_{drr}^{(a)}$ y $B_{drr}^{(a)}$, $a = 0, 1$, de los estimadores del error cuadrático medio de los EBLUP de las medias de las componentes $r = 1$ y $r = 2$ respectivamente. La tabla está ordenada por columnas y filas. La primera columna da el valor de ρ_e , la

segunda columna especifica el modelo univariante ($a = 0$) o multivariante ($a = 1$) bajo el cual se calcula el EBLUP, la tercera columna señala el área $d = 1$, $d = D/2$, $d = D$ y el valor medio de todas las áreas, las cuatro columnas siguientes muestran el error cuadrático medio $E_{drr}^{(a)}$ y las cuatro últimas el sesgo $B_{drr}^{(a)}$. Cada uno de los dos grupos de cuatro columnas que acabamos de nombrar se corresponde con un valor distinto del número de áreas; es decir, $D = 50, 100, 200, 400$. Las filas se disponen en grupos de ocho, un grupo para cada valor de ρ_e . En el caso $\rho_e = 0$, los datos se simulan de los modelos univariantes independientes ($a = 0$) que también llamaremos modelo producto de univariantes. En cambio, en los casos $\rho_e > 0$ los datos se generan del modelo multivariante diagonal ($a = 1$). Dentro de cada grupo, las cuatro primeras se corresponden con las estimaciones para $a = 0$ y las cuatro últimas con las estimaciones para $a = 1$.

En las tablas 3.4 y 3.5 se observa que conforme aumenta el valor de ρ_e los errores cuadráticos medios de los estimadores de mse_{d1} y mse_{d2} son menores en el modelo multivariante diagonal. También conviene destacar el hecho de que la utilización del modelo multivariante cuando el modelo correcto es el producto de univariantes ($\rho_e = 0$) no aumenta el error cuadrático medio de los EBLUPs.

La figura 3.3 muestra las gráficas los valores de $E_{drr}^{(a)}$, $a = 0, 1$, $r = 1, 2$, $d = 1, \dots, D$, $D = 100$. La figura está dividida en cuatro partes y tiene una disposición en forma de tabla con dos filas y dos columnas. La primera y la segunda fila presentan los valores de $E_{drr}^{(a)}$ para $r = 1$ y $r = 2$ respectivamente. La primera y la segunda columna presentan los valores de $E_{drr}^{(a)}$ cuando los datos se generan del modelo con $\rho_e = 0$ y $\rho_e = \frac{3}{4}$ respectivamente. Cada una de las cuatro sub-figuras muestran los valores de $E_{drr}^{(a)}$ para $a = 0$ y $a = 1$.

En la figura 3.3 se observa que la diferencia de los errores cuadráticos medios entre el modelo univariante ($a = 0$) y el modelo multivariante ($a = 1$) cuando ρ_e cambia de 0 a $\frac{3}{4}$ es bastante pronunciada. También hay que destacar que si los datos se generan del modelo $a = 0$ con $\rho_e = 0$, entonces no se aprecia un aumento del error cuadrático medio de $mse_{dr}^{(a)}$ al utilizar el modelo multivariante $a = 1$.

La figura 3.4 muestra las gráficas los valores de $B_{drr}^{(a)}$, $a = 0, 1$, $r = 1, 2$, $d = 1, \dots, D$, $D = 100$. Esta figura está estructurada de la misma forma que la figura 3.3. En la figura 3.4 se observa una leve diferencia en los sesgos del modelo univariante ($a = 0$) y del modelo multivariante ($a = 1$) cuando ρ_e cambia de 0 a 0,75. También hay que destacar que si los datos se generan del modelo $a = 0$ con $\rho_e = 0$, entonces no se aprecia un aumento de sesgo al utilizar el modelo multivariante $a = 1$.

ρ_e	a	d	E				B			
			50	100	200	400	50	100	200	400
0	0	1	0.0048	0.0025	0.0012	0.0006	0.0121	-0.0157	-0.0090	0.0049
		$D/2$	0.0047	0.0022	0.0011	0.0006	0.0078	-0.0016	0.0012	-0.0055
		D	0.0050	0.0022	0.0013	0.0008	0.0189	0.0000	-0.0112	-0.0127
		mean	0.0047	0.0023	0.0012	0.0007	0.0005	-0.0007	-0.0044	-0.0011
	1	1	0.0048	0.0025	0.0012	0.0006	0.0121	-0.0157	-0.0090	0.0049
		$D/2$	0.0047	0.0022	0.0011	0.0006	0.0078	-0.0016	0.0012	-0.0055
		D	0.0050	0.0022	0.0013	0.0008	0.0189	0.0000	-0.0112	-0.0127
		mean	0.0047	0.0023	0.0012	0.0007	0.0005	-0.0007	-0.0044	-0.0011
$\frac{1}{4}$	0	1	0.0041	0.0021	0.0012	0.0006	0.0142	0.0150	-0.0024	0.0071
		$D/2$	0.0039	0.0020	0.0013	0.0008	-0.0012	0.0060	0.0109	0.0155
		D	0.0040	0.0020	0.0012	0.0005	-0.0111	0.0099	-0.0005	0.0011
		mean	0.0041	0.0020	0.0013	0.0006	0.0057	0.0033	-0.0009	0.0009
	1	1	0.0038	0.0020	0.0011	0.0005	0.0117	0.0136	-0.0016	0.0058
		$D/2$	0.0036	0.0019	0.0013	0.0007	-0.0018	0.0067	0.0126	0.0158
		D	0.0038	0.0020	0.0011	0.0005	-0.0123	0.0116	0.0006	0.0025
		mean	0.0038	0.0019	0.0012	0.0006	0.0054	0.0028	-0.0011	0.0009
$\frac{1}{2}$	0	1	0.0045	0.0024	0.0010	0.0007	0.0154	0.0122	0.0097	-0.0075
		$D/2$	0.0044	0.0022	0.0012	0.0006	-0.0111	0.0052	-0.0155	-0.0009
		D	0.0043	0.0022	0.0010	0.0006	-0.0018	-0.0039	-0.0075	-0.0059
		mean	0.0044	0.0023	0.0010	0.0007	0.0039	0.0024	0.0010	0.0000
	1	1	0.0036	0.0018	0.0007	0.0005	0.0193	0.0120	0.0057	-0.0070
		$D/2$	0.0033	0.0016	0.0008	0.0004	-0.0076	0.0023	-0.0094	-0.0016
		D	0.0032	0.0016	0.0008	0.0005	-0.0022	-0.0008	-0.0095	-0.0038
		mean	0.0033	0.0017	0.0008	0.0005	0.0023	0.0019	0.0010	-0.0002
$\frac{3}{4}$	0	1	0.0048	0.0026	0.0014	0.0006	-0.0143	0.0003	0.0206	-0.0025
		$D/2$	0.0046	0.0029	0.0010	0.0006	0.0032	0.0183	0.0021	0.0009
		D	0.0046	0.0027	0.0010	0.0006	-0.0081	0.0123	-0.0039	0.0040
		mean	0.0046	0.0026	0.0011	0.0007	-0.0037	0.0017	0.0003	-0.0009
	1	1	0.0027	0.0014	0.0007	0.0004	-0.0070	-0.0037	0.0123	-0.0075
		$D/2$	0.0026	0.0017	0.0006	0.0003	-0.0048	0.0157	-0.0070	0.0027
		D	0.0026	0.0015	0.0007	0.0004	-0.0044	0.0108	-0.0104	0.0064
		mean	0.0027	0.0015	0.0007	0.0004	-0.0006	0.0009	-0.0015	-0.0002

Tabla 3.4: $E_{11d}^{(a)}$ (izquierda) y $B_{11d}^{(a)}$ (derecha) para $\rho_x = 1/2$, $a = 0, 1$.

ρ_e	a	d	E				B			
			50	100	200	400	50	100	200	400
0	0	1	0.0182	0.0080	0.0052	0.0024	-0.0101	0.0023	-0.0307	0.0053
		$D/2$	0.0187	0.0082	0.0043	0.0024	-0.0254	-0.0141	-0.0098	-0.0057
		D	0.0179	0.0090	0.0047	0.0024	0.0077	0.0310	-0.0216	-0.0033
		mean	0.0184	0.0084	0.0046	0.0028	-0.0057	-0.0009	-0.0004	0.0010
	1	1	0.0182	0.0080	0.0052	0.0024	-0.0101	0.0023	-0.0307	0.0053
		$D/2$	0.0187	0.0082	0.0043	0.0024	-0.0254	-0.0141	-0.0098	-0.0057
		D	0.0179	0.0090	0.0047	0.0024	0.0077	0.0310	-0.0216	-0.0033
		mean	0.0184	0.0084	0.0046	0.0028	-0.0057	-0.0009	-0.0004	0.0010
$\frac{1}{4}$	0	1	0.0155	0.0095	0.0048	0.0030	-0.0049	-0.0124	-0.0118	-0.0260
		$D/2$	0.0160	0.0104	0.0047	0.0024	0.0229	0.0335	0.0030	0.0084
		D	0.0169	0.0093	0.0047	0.0023	0.0392	-0.0030	0.0061	0.0003
		mean	0.0158	0.0097	0.0051	0.0027	0.0081	0.0030	-0.0009	0.0016
	1	1	0.0152	0.0089	0.0048	0.0025	-0.0054	-0.0151	-0.0205	-0.0180
		$D/2$	0.0156	0.0095	0.0044	0.0023	0.0223	0.0284	0.0064	0.0099
		D	0.0163	0.0086	0.0044	0.0022	0.0355	0.0000	0.0089	-0.0020
		mean	0.0155	0.0091	0.0048	0.0026	0.0069	0.0026	-0.0012	0.0015
$\frac{1}{2}$	0	1	0.0182	0.0097	0.0060	0.0025	0.0140	0.0301	0.0441	0.0040
		$D/2$	0.0187	0.0090	0.0040	0.0027	0.0275	-0.0120	-0.0049	0.0148
		D	0.0180	0.0088	0.0042	0.0028	-0.0136	-0.0030	0.0145	0.0196
		mean	0.0183	0.0092	0.0044	0.0028	-0.0064	-0.0024	0.0045	0.0009
	1	1	0.0134	0.0066	0.0051	0.0020	0.0191	0.0127	0.0443	0.0121
		$D/2$	0.0136	0.0064	0.0032	0.0021	0.0235	-0.0019	-0.0097	0.0180
		D	0.0132	0.0066	0.0032	0.0020	-0.0176	0.0141	0.0105	0.0127
		mean	0.0134	0.0067	0.0034	0.0021	-0.0063	-0.0005	0.0040	0.0003
$\frac{3}{4}$	0	1	0.0179	0.0099	0.0047	0.0023	0.0352	-0.0178	-0.0187	-0.0104
		$D/2$	0.0184	0.0097	0.0047	0.0022	0.0422	0.0125	-0.0191	-0.0031
		D	0.0169	0.0101	0.0081	0.0031	0.0200	-0.0238	-0.0613	0.0310
		mean	0.0171	0.0100	0.0048	0.0025	0.0124	-0.0019	-0.0092	0.0004
	1	1	0.0115	0.0056	0.0029	0.0012	0.0256	0.0024	-0.0224	-0.0086
		$D/2$	0.0114	0.0059	0.0024	0.0012	0.0236	0.0179	-0.0025	0.0066
		D	0.0110	0.0056	0.0044	0.0017	0.0150	-0.0093	-0.0446	0.0237
		mean	0.0111	0.0058	0.0027	0.0014	0.0093	0.0008	-0.0065	0.0017

Tabla 3.5: $E_{22d}^{(a)}$ (izquierda) y $B_{22d}^{(a)}$ (derecha) para $\rho_x = 1/2$, $a = 0, 1$.

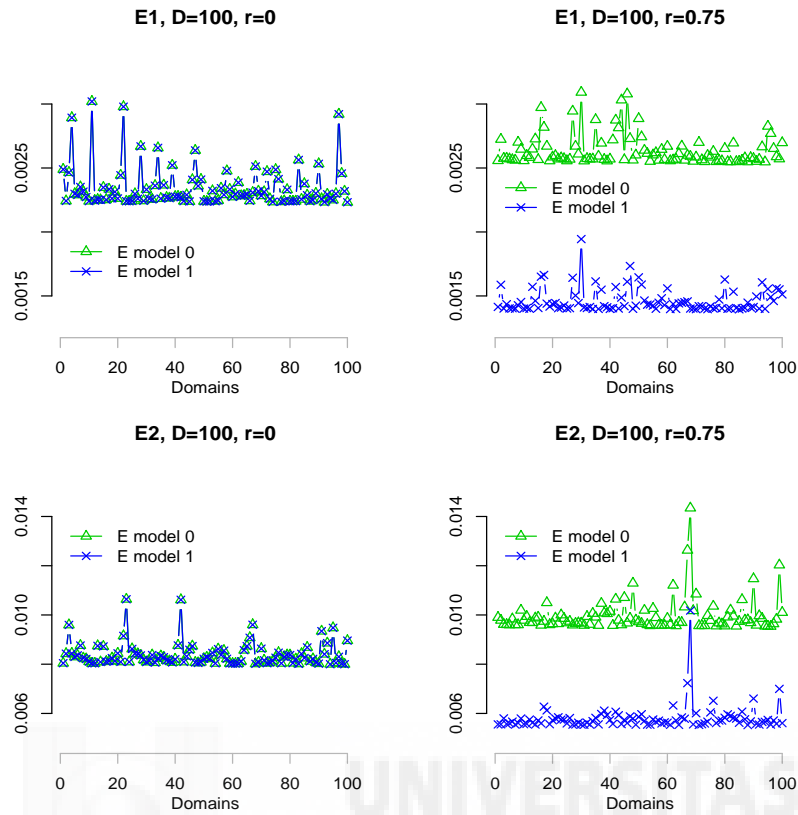


Figura 3.3: E_{drr} , para $a = 0, 1$, $r = 1, 2$, $\rho_e = 0, 3/4$, $\rho_x = 1/2$, $D = 100$.

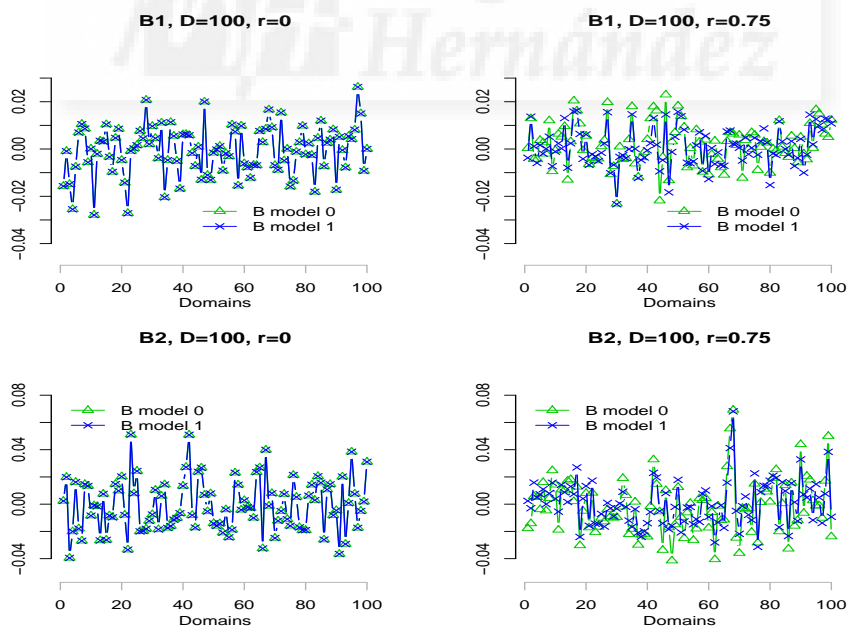


Figura 3.4: B_{drr} , para $a = 0, 1$, $r = 1, 2$, $\rho_e = 0, 3/4$, $\rho_x = 1/2$, $D = 100$.

3.2.3. Experimento de simulación 3

El objetivo de este experimento es comprobar el funcionamiento del bootstrap paramétrico en la estimación de los errores cuadráticos medios del EBLUP de las medias poblacionales en un modelo multivariante diagonal. En esta simulación se consideran los valores $\rho_e = 0, 1/4, 1/2, 3/4$. El resto de parámetros son los mismos que en la simulación 1; es decir, $\beta_1 = 1, \beta_2 = 1, \sigma_{u1}^2 = 2$ y $\sigma_{u2}^2 = 4, \sigma_{d11} = 1, \sigma_{d22} = 2$. Las variables auxiliares x_{dr} también se generan de la misma forma que en la simulación 1. Los pasos del experimento de simulación son

1. Repetir $I = 500$ veces ($i = 1, \dots, 500$)

1.1. Generar una muestra $(y_{dr}^{(i)}, x_{dr}^{(i)})$, $d = 1, \dots, D, r = 1, 2$.

1.2. Calcular $\mu_d^{(i)} = X_d^{(i)} \beta + I_2 u_d^{(i)}$.

1.3. Calcular $\hat{\sigma}_{u1}^{2(i)}, \hat{\sigma}_{u2}^{2(i)}, \hat{\beta}_{E1}^{(i)}$ y $\hat{\beta}_{E2}^{(i)}$.

1.4. Para $d = 1, \dots, D$, calcular $\hat{u}_{Ed}^{(i)}$, usando $\hat{\sigma}_{ur}^{2(i)}, \hat{\beta}_{Er}^{(i)}, r = 1, 2$. Calcular

$$\begin{aligned} \hat{\mu}_{Ed}^{(i)} &= X_d^{(i)} \hat{\beta}_E^{(i)} + I_2 \hat{u}_{Ed}^{(i)} \\ mse_d^{(i)} &= G_{1d}^{(i)}(\hat{\sigma}_{u1}^{2(i)}, \hat{\sigma}_{u2}^{2(i)}) + G_{2d}^{(i)}(\hat{\sigma}_{u1}^{2(i)}, \hat{\sigma}_{u2}^{2(i)}) + 2G_{3d}^{(i)}(\hat{\sigma}_{u1}^{2(i)}, \hat{\sigma}_{u2}^{2(i)}) \end{aligned}$$

1.5. Repetir $B = 200$ veces ($b = 1, \dots, B$)

1.5.1. Generar $u_d^{*(ib)}, e_{dr}^{*(ib)}$, $d = 1, \dots, D, r = 1, 2$ (cf. B-C en Sección 1), pero usando $\hat{\sigma}_{ur}^{2(i)}$ en lugar de σ_{ur}^2 .

1.5.2. Generar una muestra bootstrap $(y_{dr}^{*(ib)}, x_{dr}^{(i)})$, $d = 1, \dots, D, r = 1, 2$, del modelo

$$y_{dr}^{*(ib)} = x_{dr}^{(i)} \hat{\beta}_{Er}^{(i)} + u_{dr}^{*(ib)} + e_{dr}^{*(ib)}.$$

1.5.3. Calcular $\mu_d^{*(ib)} = X_d^{(i)} \hat{\beta}_E^{(i)} + u_d^{*(ib)}$.

1.5.4. Calcular $\hat{\sigma}_{ur}^{2*(ib)}$ a partir de $\hat{\sigma}_{ur}^{2(i)}$, reemplazando convenientemente los elementos de la muestra bootstrap.

1.5.5. Calcular $\hat{\beta}_{Br}^{*(ib)}$ y $\hat{\beta}_{Er}^{*(ib)}$, $r = 1, 2$; es decir, las versiones bootstrap de $\hat{\beta}_{Br}$ y $\hat{\beta}_{Er}$ respectivamente. Se calculan usando $\hat{V}_d^{(i)}$ e $y_d^{*(ib)}$ para el cálculo de $\hat{\beta}_{Br}^{*(ib)}$ y $\hat{V}_d^{*(ib)}$ e $y_d^{*(ib)}$ para el cálculo de $\hat{\beta}_{Er}^{*(ib)}$.

1.5.6. Para $d = 1, \dots, D$ y $r = 1, 2$, calcular $\hat{u}_{Bd}^{*(ib)}$ y $\hat{u}_{Ed}^{*(ib)}$, a partir de $\hat{\sigma}_{ur}^{2(i)}$ y $\hat{\beta}_{Br}^{*(ib)}, \hat{\sigma}_{ur}^{2*(ib)}$ y $\hat{\beta}_{Er}^{*(ib)}$, respectivamente.

1.5.7. Para $d = 1, \dots, D$, calcular

$$\hat{\mu}_{Bd}^{*(ib)} = X_d^{(i)} \hat{\beta}_B^{*(ib)} + I_2 \hat{u}_{Bd}^{*(ib)} \quad \text{y} \quad \hat{\mu}_{Ed}^{*(ib)} = X_d^{(i)} \hat{\beta}_E^{*(ib)} + I_2 \hat{u}_{Ed}^{*(ib)}.$$

1.5.8. Para $d = 1, \dots, D$, calcular

$$\delta_{Ed}^{*(ib)} = (\hat{\mu}_{Ed}^{*(ib)} - \mu_{Ed}^{*(ib)}), \quad \delta_{Bd}^{*(ib)} = (\hat{\mu}_{Bd}^{*(ib)} - \mu_{Bd}^{*(ib)}), \quad \delta_{EBd}^{*(ib)} = (\hat{\mu}_{Ed}^{*(ib)} - \hat{\mu}_{Bd}^{*(ib)}).$$

1.6 Para $d = 1, \dots, D$, calcular

$$\begin{aligned} mse_d^{*1(i)} &= \frac{1}{B} \sum_{b=1}^B \delta_{Ed}^{*(ib)} \delta_{Ed}^{*(ib)t} \\ mse_d^{*2(i)} &= G_{1d}^{(i)}(\hat{\sigma}_u^{2(i)}) + G_{2d}(i)(\hat{\sigma}_u^{2(i)}) + \frac{1}{B} \sum_{b=1}^B \delta_{EBd}^{*(ib)} \delta_{EBd}^{*(ib)t} \\ mse_d^{*3(i)} &= 2[G_{1d}^{(i)}(\hat{\sigma}_u^{2(i)}) + G_{2d}(i)(\hat{\sigma}_u^{2(i)})] - \frac{1}{B} \sum_{b=1}^B [G_1(\hat{\sigma}_u^{2*(ib)}) + G_2(\hat{\sigma}_u^{2*(ib)})] \\ &\quad + \frac{1}{B} \sum_{b=1}^B \delta_{EBd}^{*(ib)} \delta_{EBd}^{*(ib)t}. \end{aligned}$$

2. Salida:

$$mse_d = \frac{1}{I} \sum_{i=1}^I mse_d^{(i)}, \quad mse_d^{*\ell} = \frac{1}{I} \sum_{i=1}^I mse_d^{*\ell(i)}, \quad \ell = 1, 2, 3.$$

3. Leer los MSE_{drr} obtenidos en la simulación 1 para el caso $\rho_e = \frac{1}{2}$ y hacer

$$\begin{aligned} B_{drr}^0 &= \frac{1}{I} \sum_{i=1}^I (mse_{drr}^{(i)} - MSE_{drr}), \quad B_{drr}^{*\ell} = \frac{1}{I} \sum_{i=1}^I (mse_{drr}^{*\ell(i)} - MSE_{drr}), \quad \ell = 1, 2, 3, r = 1, 2, \\ E_{drr}^0 &= \frac{1}{I} \sum_{i=1}^I (mse_{drr}^{(i)} - MSE_{drr})^2, \quad E_{drr}^{*\ell} = \frac{1}{I} \sum_{i=1}^I (mse_{drr}^{*\ell(i)} - MSE_{drr})^2, \quad \ell = 1, 2, 3, r = 1, 2, \\ B_{rr}^0 &= \frac{1}{D} \sum_{d=1}^D B_{drr}^0, \quad B_{rr}^{*\ell} = \frac{1}{D} \sum_{d=1}^D B_{drr}^{*\ell}, \quad E_{rr}^0 = \frac{1}{D} \sum_{d=1}^D E_{drr}^0, \quad E_{rr}^{*\ell} = \frac{1}{D} \sum_{d=1}^D E_{drr}^{*\ell}, \quad \ell = 1, 2, 3, r = 1, 2. \end{aligned}$$

D	E_{11}^0	E_{11}^{*1}	E_{11}^{*2}	E_{11}^{*3}	E_{22}^0	E_{22}^{*1}	E_{22}^{*2}	E_{22}^{*3}
50	0.00345	0.00783	0.00367	0.00346	0.01328	0.03119	0.01477	0.01337
100	0.00172	0.00581	0.00185	0.00172	0.00755	0.02392	0.00805	0.00768
200	0.00083	0.00484	0.00085	0.00083	0.00390	0.02002	0.00394	0.00393
400	0.00051	0.00447	0.00051	0.00051	0.00202	0.01801	0.00206	0.00201

Tabla 3.6: $E_{rr}^0, E_{rr}^{*\ell}, \ell = 1, 2, 3, r = 1, 2.$

D	B_{11}^0	B_{11}^{*1}	B_{11}^{*2}	B_{11}^{*3}	B_{22}^0	B_{22}^{*1}	B_{22}^{*2}	B_{22}^{*3}
50	0.00812	-0.00495	-0.00520	0.00853	0.00503	-0.02131	-0.02184	0.00588
100	-0.00233	-0.00825	-0.00895	-0.00205	-0.00130	-0.01528	-0.01463	-0.00068
200	0.00148	-0.00193	-0.00182	0.00157	0.00469	-0.00166	-0.00186	0.00482
400	-0.00013	-0.00175	-0.00177	-0.00004	-0.00147	-0.00465	-0.00478	-0.00108

Tabla 3.7: $B_{rr}^0, B_{rr}^{*\ell}, \ell = 1, 2, 3, r = 1, 2.$

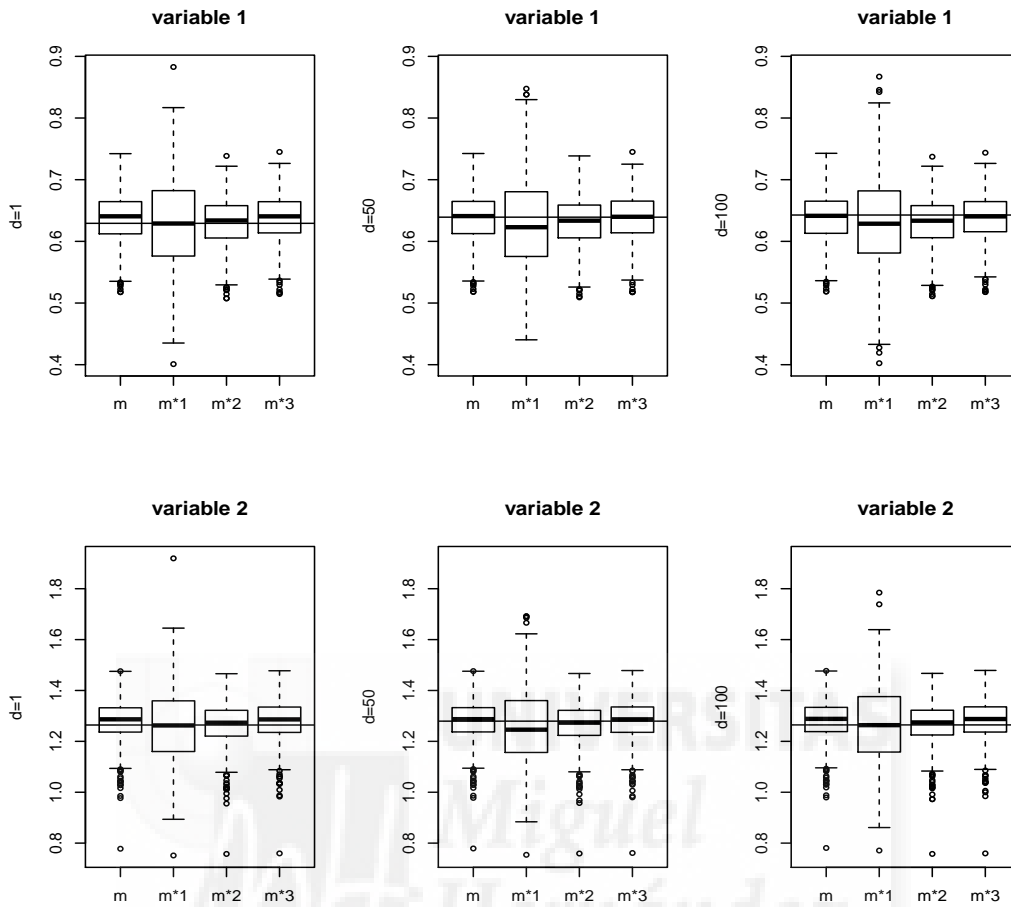


Figura 3.5: Diagrama de cajas de mse_{drr} 's, $msed_{drr}^{*1}$'s, $msed_{drr}^{*2}$'s y $msed_{drr}^{*3}$ para $D = 100$.

La primera columna de la tabla 3.6 da el número de áreas consideradas en la simulación; es decir, $D = 50$, $D = 100$, $D = 200$ y $D = 400$. Las cuatro columnas siguientes muestran el error cuadrático medio de los estimadores $msed$, $msed^{*1}$, $msed^{*2}$ y $msed^{*3}$ para $\hat{\mu}_{d1}$, donde el valor teórico considerado es el obtenido en la simulación 1. Las cuatro columnas siguientes presentan lo mismo, pero esta vez para el estimador $\hat{\mu}_{d2}$. La disposición de las columnas en la tabla 3.7 es la misma pero esta vez para el sesgo.

En la tabla 3.6 se observa que el estimador $msed^{*1}$ produce los mayores errores y los estimadores $msed$ y $msed^{*3}$ los menores. Asimismo se observa que el error disminuye de forma considerable en los cuatro estimadores al aumentar el número de áreas consideradas. En la tabla 3.7 se observa que los sesgos para los estimadores $msed^{*1}$ y $msed^{*2}$ son negativos, lo cual no resulta nada sorprendente, ya que les afecta que el valor esperado del término $G_1(\hat{\theta})$ es de forma aproximada $G_1(\theta) - G_3(\theta)$.

La figura 3.5 contiene los diagramas de cajas de los cuatro estimadores considerados. Se observa que las distribuciones de $msed$, $msed^{*2}$ y $msed^{*3}$ son mejores que la de $msed^{*1}$. También se nota la presencia de sesgo negativo para los estimadores $msed^{*1}$ y $msed^{*2}$.

4

Modelo AR(1)

4.1. Definición del modelo

En el trabajo de Rao-Yu (1994), y posteriormente en el de Esteban et al. (2012), se estudia un modelo lineal mixto donde los efectos aleatorios se distribuyen según un proceso AR(1). En ambos trabajos se realiza un enfoque temporal, es decir, en cada una de las áreas pequeñas consideradas se estudia una variable objetivo en varios instantes de tiempo. En este capítulo se estudia un modelo lineal mixto, donde los efectos aleatorios asociados a las áreas se distribuyen según un proceso AR(1). El modelo de este capítulo, además del enfoque temporal, admite el enfoque multivariante; es decir, en cada área se consideran R variables que tienen asociados efectos aleatorios con correlación de tipo AR(1). El enfoque temporal sigue siendo válido si por cada área existe una única variable medida en R instantes de tiempo. En este último caso los efectos aleatorios tienen correlación temporal AR(1).

El modelo lineal multivariante mixto, que se denominará *modelo AR(1)*, es

$$y = X\beta + Zu + e = X\beta + Z_1u_1 + \dots + Z_Du_D + e, \quad (4.1)$$

donde e, u_1, \dots, u_D son independientes con distribuciones $e \sim N(0, V_e)$, $u \sim N(0, V_u)$, y $u_d \sim N(0, V_{ud})$, $d = 1, \dots, D$. Se supone que V_e es una matriz conocida y que

$$V_{ud} = \sigma_u^2 \Omega_d(\rho), \quad \Omega_d(\rho) = \frac{1}{1 - \rho^2} \begin{pmatrix} 1 & \rho & \dots & \rho^{R-1} \\ \rho & 1 & \dots & \rho^{R-2} \\ \vdots & \vdots & & \vdots \\ \rho^{R-1} & \rho^{R-2} & \dots & 1 \end{pmatrix}.$$

En la notación del capítulo 2 se tiene que el número de componentes de la varianza es $m = 2$ y el vector de componentes es $\theta = (\theta_1, \theta_2)$, donde $\theta_1 = \sigma_u^2$ y $\theta_2 = \rho$.

Las derivadas de la matriz V_{ud} respecto de los parámetros de varianza aparecen en los vectores y matrices, $S(\theta)$ y $F(\theta)$, de la ecuación de actualización del algoritmo Fisher-scoring para el cálculo del estimador REML de θ . Las derivadas son

$$V_1 = \frac{\partial V}{\partial \theta_1} = \frac{\partial V}{\partial \sigma_u^2} = \text{diag}(\Omega_d(\rho)), \quad V_2 = \frac{\partial V}{\partial \theta_2} = \frac{\partial V}{\partial \rho} = \sigma_u^2 \text{diag}(\Omega'_d(\rho)),$$

donde

$$\Omega'_d(\rho) = \frac{1}{1-\rho^2} \begin{pmatrix} 0 & 1 & \dots & (R-1)\rho^{R-2} \\ 1 & 0 & \dots & (R-2)\rho^{R-3} \\ \vdots & \vdots & \ddots & \vdots \\ (R-1)\rho^{R-2} & (R-2)\rho^{R-3} & \dots & 0 \end{pmatrix} - \frac{2\rho}{1-\rho^2} \Omega_d(\rho).$$

La distribución asintótica de los estimadores REML de θ es $\hat{\theta} \sim N_2(\theta, F^{-1}(\theta))$. Por tanto, la distribución asintótica de $\hat{\rho}$ es

$$\hat{\rho} \sim N(\rho, v_{22}) \quad (4.2)$$

donde v_{22} es el elemento correspondiente de la matriz $F^{-1}(\theta)$. La distribución (4.2) se puede usar para comprobar la no nulidad del parámetro de correlación mediante el contraste de la hipótesis $H_0 : \rho = 0$. Si el nivel de significación se fija en α , entonces se tiene que se rechaza H_0 si

$$\frac{\hat{\rho}}{\sqrt{\hat{v}_{22}}} \notin (-z_{\alpha/2}, z_{\alpha/2}).$$

4.2. Experimentos de simulación

Para estudiar empíricamente el comportamiento de los algoritmos de ajuste y de los procedimientos de estimación del error cuadrático medio de los EBLUPs, en esta sección se presentan tres experimentos de simulación. En las simulaciones se compara el modelo (4.1) con el modelo con errores e_{dr} independientes. Este último modelo prescinde de toda estructura multivariante y equivale a aplicar por separado R modelos Fay-Herriot; es decir, uno por cada componente r , $r = 1, \dots, R$.

En las simulaciones, se ha programado un modelo AR(1) bivalente ($R = 2$) cuyas características se describen en lo que sigue.

La matriz de covarianzas del vector u_d es

$$V_{ud} = \sigma_u^2 \Omega_d(\rho), \quad \Omega_d(\rho) = \frac{1}{1-\rho^2} \begin{pmatrix} 1 & \rho \\ \rho & 1 \end{pmatrix}, \quad \sigma_u^2 = 2, \quad \rho = \frac{1}{2}.$$

Las componentes del vector e_d verifican $\text{var}(e_{d1}) = 1$, $\text{var}(e_{d2}) = 2$ y $\text{corr}(e_{d1}, e_{d2}) = \rho_e$. Por tanto, la matriz de covarianzas del vector e_d es

$$V_{ed} = \begin{pmatrix} \sigma_{d11} & \sigma_{d12} \\ \sigma_{d21} & \sigma_{d22} \end{pmatrix}, \quad \sigma_{d11} = 1, \quad \sigma_{d22} = 2, \quad \sigma_{d12} = \sigma_{d21} = \rho_e \sqrt{\sigma_{d11} \sigma_{d22}}.$$

y la matriz de covarianzas del vector y_d es

$$\begin{aligned} V_d &= V_{ud} + V_{ed} = \frac{\sigma_u^2}{1-\rho^2} \begin{pmatrix} 1 & \rho \\ \rho & 1 \end{pmatrix} + \begin{pmatrix} \sigma_{d11} & \sigma_{d12} \\ \sigma_{d21} & \sigma_{d22} \end{pmatrix} = \begin{pmatrix} \frac{8}{3} & \frac{4}{3} \\ \frac{4}{3} & \frac{8}{3} \end{pmatrix} + \begin{pmatrix} 1 & \rho_e\sqrt{2} \\ \rho_e\sqrt{2} & 2 \end{pmatrix} \\ &= \begin{pmatrix} 3 & \rho_e\sqrt{2} \\ \rho_e\sqrt{2} & 6 \end{pmatrix}, \quad d = 1, \dots, D. \end{aligned}$$

Finalmente, la matriz de covarianzas del vector y es

$$V = \text{diag}_{1 \leq d \leq D} \begin{pmatrix} \frac{11}{3} & \frac{4}{3} + \rho_e\sqrt{2} \\ \frac{4}{3} + \rho_e\sqrt{2} & \frac{14}{3} \end{pmatrix}.$$

Las derivadas parciales que se utilizan en el algoritmo de Fisher-scoring son

$$V_1 = \text{diag}_{1 \leq d \leq D} (V_{d1}) \quad \text{y} \quad V_2 = \text{diag}_{1 \leq d \leq D} (V_{d2}),$$

donde

$$\begin{aligned} V_{d1} &= \frac{\partial V_{ud}}{\partial \sigma_u^2} = \frac{1}{1-\rho^2} \begin{pmatrix} 1 & \rho \\ \rho & 1 \end{pmatrix}, \\ V_{d2} &= \frac{\partial V_{ud}}{\partial \rho} = \sigma_u^2 \frac{2\rho}{(1-\rho^2)^2} \begin{pmatrix} 1 & \rho \\ \rho & 1 \end{pmatrix} + \sigma_u^2 \frac{1}{1-\rho^2} \begin{pmatrix} 0 & 1 \\ 1 & 0 \end{pmatrix}. \end{aligned}$$

Puesto que $\sigma_u^2 = 2$ y $\rho = 1/2$, se tiene que

$$V_{d1} = \frac{\partial V_{ud}}{\partial \sigma_u^2} = \begin{pmatrix} \frac{4}{3} & \frac{2}{3} \\ \frac{2}{3} & \frac{4}{3} \end{pmatrix} \quad \text{y} \quad V_{d2} = \frac{\partial V_{ud}}{\partial \rho} = \begin{pmatrix} -\frac{3}{2} & \frac{3}{2} \\ \frac{3}{2} & -\frac{3}{2} \end{pmatrix}.$$

Los valores de los parámetros de regresión son $\beta_1 = 1$, $\beta_2 = 1$. Para estimar los parámetros σ_u^2 y ρ mediante el algoritmo Fisher-scoring se utilizan como valores iniciales (semillas) los verdaderos valores; es decir, $\sigma_u^2 = 2$ y $\rho = \frac{1}{2}$.

Como se van a realizar comparaciones entre el modelo multivariante y los modelos univariantes marginales, a continuación se da la descripción de tales modelos así como los parámetros que intervienen en los mismos. Los modelos univariantes son

$$y_{dr} = x_{dr}\beta_r + u_{dr} + e_{dr}, \quad d = 1, \dots, D, \quad r = 1, 2,$$

donde $u_{dr} \sim N(0, \frac{\sigma_u^2}{1-\rho^2})$ y $e_{dr} \sim N(0, \sigma_{drr}^2)$ son independientes. Los parámetros de los modelos univariantes son los mismos que se usan en el modelo multivariante AR(1); es decir, $\beta_1 = 1$, $\beta_2 = 1$, $\sigma_u^2 = 2$, $\rho = \frac{1}{2}$, $\sigma_{d11} = 1$ y $\sigma_{d22} = 2$.

Para ambos modelos se utilizan las variables explicativas

$$x_{d1} = \mu_1 + \sigma_{x11}^{1/2} U_{d1}, \quad x_{d2} = \mu_2 + \sigma_{x22}^{1/2} (\rho_x U_{d1} + (1-\rho_x^2)^{1/2} U_{d2}), \quad d = 1, \dots, D,$$

donde $\mu_1 = \mu_2 = 10$, $\sigma_{x11} = 1$, $\sigma_{x22} = 2$, $\rho_x = 0,5$ y $U_{dr} = \frac{d-D}{D} + \frac{r}{3}$, $r = 1, 2$, $d = 1, \dots, D$.

4.2.1. Experimento de simulación 1

El objetivo de este experimento es investigar empíricamente la pérdida de eficiencia en las estimaciones cuando no se tiene en cuenta la naturaleza multivariante de los datos. Nos interesa estudiar lo que ocurre cuando se asume incorrectamente los modelos marginales independientes, en lugar del modelo multivariante subyacente. Para ello, se simulan los datos o bien del modelo multivariante o bien del modelo con $\rho_e = \rho = 0$, se estiman los parámetros y se calculan los EBLUP de ambos modelos. Hay dos conjuntos de parámetros estimados y EBLUP calculados. Según sea el caso, un conjunto se obtiene bajo el modelo correcto y otro bajo el modelo incorrecto. Cabe esperar que los mejores resultados se obtengan siempre cuando se usan los estimadores correspondientes al modelo correcto. El experimento consiste en calcular por simulación Monte Carlo los sesgos y errores cuadráticos medios empíricos (BIAS y MSE) de los estimadores de los parámetros y de los EBLUP.

Los datos se simulan del modelo multivariante (4.1). Se consideran los casos: (1) $\rho_e = 0, \rho = 0$, (2) $\rho_e = 1/2, \rho = 0$, (3) $\rho_e = 0, \rho = 1/2$ y (4) $\rho_e = 1/2, \rho = 1/2$. En el caso 1, los datos se simulan del modelo producto de modelos marginales, pero restringido a $\sigma_{u1}^2 = \sigma_{u2}^2$. Los pasos de la simulación son

1. Repetir $I = 10^4$ veces ($i = 1, \dots, I$)
 - 1.1. Generar una muestra $(y_{dr}, x_{dr}), d = 1, \dots, D, r = 1, 2$.
 - 1.2. Calcular $\{\hat{\beta}_{E1}^{(i,0)}, \hat{\beta}_{E2}^{(i,0)}, \hat{\sigma}_{u1}^{2(i,0)}, \hat{\sigma}_{u2}^{2(i,0)}\}, \{\hat{\beta}_{E1}^{(i,1)}, \hat{\beta}_{E2}^{(i,1)}, \hat{\sigma}_u^{2(i,1)}, \hat{\rho}^{(i,1)}\}$ y $\{\hat{\mu}_d^{(i,0)}, \hat{\mu}_d^{(i,1)}\}$, donde los superíndices “0” y “1” se usan para denotar los estimadores y EBLUPs calculados asumiendo el modelo producto de marginales ($a = 0$) o el modelo multivariante AR(1) ($a = 1$). El modelo $a = 0$ no asume la igualdad $\hat{\sigma}_{u1}^{2(0)} = \hat{\sigma}_{u2}^{2(0)}$.
2. Salida: Para todo $\tau \in \{\beta_1^{(0)}, \beta_2^{(0)}, \sigma_{u1}^{2(0)}, \sigma_{u2}^{2(0)}, \beta_1^{(1)}, \beta_2^{(1)}, \sigma_u^{2(1)}, \hat{\rho}^{(1)}\}$, calcular

$$MSE(\hat{\tau}) = \frac{1}{I} \sum_{i=1}^I (\hat{\tau}^{(i)} - \tau)^2, \quad BIAS(\hat{\tau}) = \frac{1}{I} \sum_{i=1}^I (\hat{\tau}^{(i)} - \tau),$$

$$MSE_{rrd}^{(a)} = \frac{1}{I} \sum_{i=1}^I (\hat{\mu}_{rd}^{(i,a)} - \mu_{rd}^{(i,a)})^2, \quad BIAS_{rrd}^{(a)} = \frac{1}{I} \sum_{i=1}^I (\hat{\mu}_{rd}^{(i,a)} - \mu_{rd}^{(i,a)}), \quad d = 1, D/2, D, r = 1, 2, a = 0, 1.$$

La tabla 4.1 presenta los errores cuadráticos medios y los sesgos empíricos de los estimadores REML de los parámetros de los dos modelos considerados. La tabla está ordenada por columnas y filas. La primera columna señala el caso que se está considerando (modelo que genera los datos), la segunda columna especifica el estimador del parámetro del modelo, las cuatro columnas siguientes muestran el error cuadrático medio y las cuatro últimas el sesgo. Cada uno de los dos grupos de cuatro columnas que se acaban de nombrar se corresponde con un valor distinto del número de áreas; es decir, $D = 50, 100, 200, 400$. Las filas se disponen en grupos de ocho, un grupo para cada caso. En el primer caso, los datos se simulan del modelo con $\rho = \rho_e = 0$. En los casos 2, 3 y 4 los datos se generan del modelo multivariante con $\rho > 0$ o $\rho_e > 0$. Dentro de cada caso, las cuatro primeras filas se corresponden con las estimaciones basadas en el modelo producto de univariantes ($a = 0$) y las cuatro últimas con las estimaciones basadas en el modelo multivariante ($a = 1$).

La tabla 4.1 muestra que los errores cuadráticos medios y sesgos de $\hat{\beta}_{E1}$ y $\hat{\beta}_{E2}$ son básicamente iguales para las estimaciones que se hacen asumiendo los modelos $a = 0$ o $a = 1$, independientemente del caso que se esté considerando en la simulación de los datos. Para los estimadores de las varianzas y del parámetro de correlación, ρ , se observa que sus errores cuadráticos medios son siempre menores cuando se asume el modelo multivariante ($a = 1$). Esto ocurre incluso en el primer caso, donde los datos se generan de un modelo producto de marginales. Este resultado no contradice la intuición, pues el modelo correcto es el modelo (4.1) con parámetros $\rho = \rho_e = 0$ y $\sigma_{u1}^2 = \sigma_{u2}^2 = \sigma_u^2$, mientras que el modelo $a = 0$ es el producto de dos modelos marginales independientes con $\sigma_{u1}^2 \neq \sigma_{u2}^2$. De hecho, el modelo $a = 0$ no es un caso particular del modelo (4.1). Así pues, en el caso 1 se puede afirmar que ambos modelos son incorrectos por estar sobreparametrizados. El modelo producto de univariantes ($a = 0$) ignora la restricción $\sigma_{u1}^2 = \sigma_{u2}^2$ y el multivariante ($a = 1$) hace caso omiso de la restricción $\rho = \rho_e = 0$. Por todo ello, se observa que el error cuadrático medio en el modelo $a = 1$ es menor, pues la estimación del parámetro ρ realiza cierta corrección sobre el parámetro σ_u^2 que no se tienen en cuenta en el modelo $a = 0$. La diferencia apuntada también se aprecia, con mayores motivos, en el resto de casos. Sobre los sesgos se puede decir, a la vista de las tablas, que tienden a ser mayores en el modelo $a = 1$ en todos los casos considerados.

Las tablas 4.2 y 4.3 presentan los errores cuadráticos medios y los sesgos, $MSE_{drr}^{(a)}$ y $BIAS_{drr}^{(a)}$, $a = 0, 1$, de los EBLUPS de las medias de las componentes $r = 1$ y $r = 2$ respectivamente. En ambas tablas, las columnas se disponen de forma análoga a la tabla 4.1. La diferencia está en la segunda columna donde aparecen los valores de las áreas consideradas, $d = 1$, $d = D/2$ y $d = D$, y el valor medio a lo largo de ellas. El resto de columnas está estructurado de la misma forma que en la tabla 4.1, pero el error cuadrático medio y el sesgo son de los estimadores $\hat{\mu}_{d1}$ en la tabla 4.2 y $\hat{\mu}_{d2}$ en la tabla 4.3. En las tablas 4.2 y 4.3 se observa en el primer caso ($\rho_e = 0, \rho = 0$) que los errores cuadráticos medios de los estimadores de $\hat{\mu}_{d1}$ y $\hat{\mu}_{d2}$ apenas se diferencian en ambos modelos; en los casos 2 y 3 la diferencia que se aprecia es considerable y en el caso 4 se aprecia diferencia, pero en menor grado comparado con los dos casos anteriores.

La figura 4.1 muestra las gráficas de los valores de $MSE_{drr}^{(a)}$, $a = 0, 1$, $r = 1, 2$, $d = 1, \dots, D$, $D = 100$. La figura está dividida en cuatro partes y tiene una disposición en forma de tabla con dos filas y dos columnas. La primera y segunda fila presentan los valores de $MSE_{drr}^{(a)}$ para $r = 1$ y $r = 2$ respectivamente. La primera y segunda columna presentan los valores de $MSE_{drr}^{(a)}$ cuando los datos se generan del modelo teniendo en cuenta los casos primero ($\rho_e = 0, \rho = 0$) y tercero ($\rho_e = 0, \rho = 1/2$) respectivamente. Cada una de las cuatro sub-figuras muestran los valores de $MSE_{drr}^{(a)}$ para $a = 0$ y $a = 1$. En la figura 4.1 se observa que la diferencia de los errores cuadráticos medios entre el modelo producto de univariantes ($a = 0$) y el modelo multivariante ($a = 1$) cuando ρ_e cambia de 0 a $\frac{1}{2}$ es bastante pronunciada. También hay que destacar que si los datos se generan del modelo $a = 0$ con $\rho_e = 0$, entonces no se aprecia un aumento del error cuadrático medio al utilizar el modelo sobre-parametrizado $a = 1$.

La figura 4.2 muestra las gráficas de los valores de $BIAS_{drr}^{(a)}$, $a = 0, 1$, $r = 1, 2$, $d = 1, \dots, D$, $D = 100$. La figura está estructurada de la misma forma que la figura 4.1. En la figura 4.2 se observa una leve diferencia en los sesgos de los modelos $a = 0$ y $a = 1$ cuando ρ_e cambia de 0 a $\frac{1}{2}$. También hay que destacar que si los datos se generan del modelo $a = 0$ con $\rho_e = 0$, entonces no se aprecia un aumento de sesgo al utilizar el modelo sobre-parametrizado $a = 1$.

(ρ_e, ρ)	D	MSE				BIAS			
		50	100	200	400	50	100	200	400
(0, 0)	$\hat{\beta}_{E1}^{(0)}$	0.00061	0.00031	0.00016	0.00008	0.00003	0.00002	0.00006	-0.00004
	$\hat{\beta}_{E2}^{(0)}$	0.00076	0.00038	0.00019	0.00010	-0.00020	0.00015	-0.00022	-0.00006
	$\hat{\sigma}_{u1}^{2(0)}$	0.36230	0.18075	0.08987	0.04586	0.00236	-0.00394	0.00177	0.00105
	$\hat{\sigma}_{u2}^{2(0)}$	0.65777	0.32322	0.16432	0.08074	0.01287	-0.00554	0.00591	0.00015
	$\hat{\beta}_{E1}^{(1)}$	0.00061	0.00031	0.00016	0.00008	0.00003	0.00002	0.00006	-0.00004
	$\hat{\beta}_{E2}^{(1)}$	0.00076	0.00038	0.00019	0.00010	-0.00021	0.00015	-0.00022	-0.00006
	$\hat{\sigma}_u^{2(1)}$	0.26669	0.12858	0.06092	0.02964	-0.11491	-0.06443	-0.02698	-0.01421
	$\hat{\rho}^{(1)}$	0.06366	0.03087	0.01534	0.00751	-0.00047	0.00029	0.00219	-0.00078
$(\frac{1}{2}, 0)$	$\hat{\beta}_{E1}^{(0)}$	0.00063	0.00032	0.00015	0.00008	0.00013	0.00016	0.00006	-0.00010
	$\hat{\beta}_{E2}^{(0)}$	0.00077	0.00038	0.00019	0.00009	0.00037	0.00001	0.00021	-0.00006
	$\hat{\sigma}_{u1}^{2(0)}$	0.36308	0.18262	0.08983	0.04499	-0.00300	0.00359	-0.00412	-0.00610
	$\hat{\sigma}_{u2}^{2(0)}$	0.65263	0.32245	0.15882	0.08105	0.00531	0.00011	-0.00150	-0.00041
	$\hat{\beta}_{E1}^{(1)}$	0.00063	0.00032	0.00015	0.00008	0.00013	0.00016	0.00006	-0.00010
	$\hat{\beta}_{E2}^{(1)}$	0.00076	0.00038	0.00019	0.00009	0.00037	0.00001	0.00021	-0.00006
	$\hat{\sigma}_u^{2(1)}$	0.28601	0.13318	0.06263	0.03103	-0.12294	-0.05988	-0.03527	-0.01939
	$\hat{\rho}^{(1)}$	0.06796	0.03249	0.01640	0.00777	-0.02134	-0.01221	-0.00724	-0.00553
$(0, \frac{1}{2})$	$\hat{\beta}_{E1}^{(0)}$	0.00077	0.00038	0.00019	0.00009	-0.00013	0.00011	-0.00042	-0.00008
	$\hat{\beta}_{E2}^{(0)}$	0.00089	0.00044	0.00022	0.00011	0.00001	-0.00008	-0.00011	-0.00003
	$\hat{\sigma}_{u1}^{2(0)}$	0.54270	0.27436	0.13581	0.06766	-0.02254	-0.00196	-0.00011	-0.00063
	$\hat{\sigma}_{u2}^{2(0)}$	0.87959	0.42655	0.22152	0.10953	0.00066	-0.00998	0.00461	-0.00355
	$\hat{\beta}_{E1}^{(1)}$	0.00077	0.00038	0.00019	0.00009	-0.00012	0.00011	-0.00042	-0.00008
	$\hat{\beta}_{E2}^{(1)}$	0.00089	0.00044	0.00022	0.00011	0.00004	-0.00008	-0.00011	-0.00003
	$\hat{\sigma}_u^{2(1)}$	0.40765	0.20745	0.10464	0.05038	-0.09783	-0.05323	-0.02293	-0.01242
	$\hat{\rho}^{(1)}$	0.03771	0.01858	0.00921	0.00430	-0.00692	-0.00162	-0.00108	-0.00062
$(\frac{1}{2}, \frac{1}{2})$	$\hat{\beta}_{E1}^{(0)}$	0.00075	0.00037	0.00019	0.00009	0.00047	-0.00026	-0.00005	0.00002
	$\hat{\beta}_{E2}^{(0)}$	0.00089	0.00045	0.00023	0.00011	0.00030	-0.00014	0.00009	-0.00011
	$\hat{\sigma}_{u1}^{2(0)}$	0.55375	0.27216	0.13499	0.06627	0.00436	-0.00471	0.00167	0.00052
	$\hat{\sigma}_{u2}^{2(0)}$	0.89180	0.44946	0.22126	0.11081	0.00659	0.00007	0.00217	0.00059
	$\hat{\beta}_{E1}^{(1)}$	0.00075	0.00037	0.00019	0.00009	0.00047	-0.00026	-0.00005	0.00002
	$\hat{\beta}_{E2}^{(1)}$	0.00089	0.00045	0.00023	0.00011	0.00030	-0.00014	0.00009	-0.00011
	$\hat{\sigma}_u^{2(1)}$	0.25160	0.12228	0.06173	0.03027	-0.07514	-0.03698	-0.01765	-0.00847
	$\hat{\rho}^{(1)}$	0.03068	0.01481	0.00717	0.00347	-0.01605	-0.01014	-0.00420	-0.00224

Tabla 4.1: MSE (izquierda) y BIAS (derecha) para $\rho_x = 1/2$.

(ρ_e, ρ)	a	d	MSE				BIAS			
			50	100	200	400	50	100	200	400
(0,0)	0	1	0.6845	0.6747	0.6710	0.6683	0.0031	0.0053	-0.0101	0.0004
		$D/2$	0.6858	0.6931	0.6664	0.6744	-0.0028	-0.0033	0.0184	-0.0070
		D	0.6966	0.6677	0.6713	0.6784	-0.0044	0.0051	0.0081	0.0026
		mean	0.6879	0.6768	0.6720	0.6696	0.0007	-0.0011	-0.0012	-0.0005
	1	1	0.6853	0.6753	0.6706	0.6690	0.0046	0.0062	-0.0095	0.0010
		$D/2$	0.6886	0.6957	0.6674	0.6750	-0.0030	-0.0024	0.0183	-0.0066
		D	0.6993	0.6687	0.6713	0.6785	-0.0061	0.0050	0.0085	0.0019
		mean	0.6896	0.6778	0.6727	0.6698	0.0007	-0.0011	-0.0012	-0.0005
$(\frac{1}{2}, 0)$	0	1	0.6872	0.6986	0.6769	0.6677	0.0070	-0.0070	0.0027	0.0031
		$D/2$	0.6980	0.6851	0.6754	0.6686	0.0074	0.0045	0.0089	-0.0101
		D	0.6819	0.6704	0.6540	0.6689	-0.0078	0.0126	0.0168	-0.0054
		mean	0.6898	0.6770	0.6719	0.6695	0.0004	0.0017	-0.0002	-0.0002
	1	1	0.6301	0.6349	0.6154	0.6179	0.0027	-0.0077	0.0009	0.0016
		$D/2$	0.6420	0.6234	0.6176	0.6143	0.0084	0.0040	0.0081	-0.0121
		D	0.6336	0.6110	0.5991	0.6154	-0.0071	0.0123	0.0164	-0.0009
		mean	0.6340	0.6209	0.6148	0.6125	0.0004	0.0017	-0.0002	-0.0002
$(0, \frac{1}{2})$	0	1	0.7608	0.7278	0.7416	0.7310	-0.0036	0.0044	-0.0145	-0.0076
		$D/2$	0.7579	0.7216	0.7223	0.7254	-0.0023	0.0002	0.0093	0.0014
		D	0.7592	0.7280	0.7440	0.7388	-0.0013	-0.0117	0.0021	-0.0019
		mean	0.7500	0.7336	0.7320	0.7295	-0.0020	-0.0027	-0.0006	-0.0005
	1	1	0.7243	0.6999	0.7064	0.7070	-0.0019	0.0020	-0.0148	-0.0059
		$D/2$	0.7313	0.6971	0.6879	0.7001	-0.0037	-0.0017	0.0107	0.0015
		D	0.7315	0.6993	0.7151	0.7035	-0.0017	-0.0070	0.0045	-0.0017
		mean	0.7173	0.7042	0.7014	0.6984	-0.0020	-0.0027	-0.0006	-0.0005
$(\frac{1}{2}, \frac{1}{2})$	0	1	0.7417	0.7565	0.7286	0.7273	0.0026	-0.0033	-0.0006	0.0058
		$D/2$	0.7382	0.7283	0.7570	0.7217	-0.0018	0.0033	-0.0032	-0.0002
		D	0.7445	0.7484	0.7344	0.7264	-0.0026	-0.0076	0.0259	-0.0041
		mean	0.7457	0.7366	0.7311	0.7285	0.0020	-0.0010	0.0013	-0.0003
	1	1	0.7370	0.7490	0.7207	0.7229	0.0023	-0.0021	-0.0003	0.0081
		$D/2$	0.7325	0.7202	0.7507	0.7150	-0.0036	0.0024	-0.0033	-0.0004
		D	0.7348	0.7429	0.7282	0.7221	-0.0027	-0.0070	0.0277	-0.0043
		mean	0.7393	0.7309	0.7248	0.7223	0.0020	-0.0010	0.0013	-0.0003

Tabla 4.2: $MSE_{11d}^{(a)}$ (izquierda) y $BIAS_{11d}^{(a)}$ (derecha) para $\rho_x = 1/2$, $a = 0, 1$.

(ρ_e, ρ)	a	d	MSE				BIAS			
			50	100	200	400	50	100	200	400
(0, 0)	0	1	1.0501	1.0425	1.0229	1.0130	-0.0040	0.0115	0.0158	0.0186
		$D/2$	1.0567	1.0192	1.0087	1.0052	-0.0050	-0.0009	-0.0134	-0.0143
		D	1.0676	1.0235	1.0013	1.0131	-0.0252	0.0078	-0.0024	0.0007
		mean	1.0604	1.0309	1.0147	1.0077	-0.0009	0.0016	-0.0011	0.0001
	1	1	1.0484	1.0406	1.0229	1.0120	-0.0041	0.0119	0.0151	0.0175
		$D/2$	1.0492	1.0181	1.0077	1.0071	-0.0016	-0.0018	-0.0129	-0.0145
		D	1.0571	1.0220	0.9993	1.0114	-0.0222	0.0067	-0.0034	0.0012
		mean	1.0547	1.0277	1.0130	1.0070	-0.0010	0.0016	-0.0012	0.0001
$(\frac{1}{2}, 0)$	0	1	1.0691	1.0427	1.0060	0.9943	0.0062	0.0128	0.0143	0.0018
		$D/2$	1.0661	1.0191	1.0079	1.0023	0.0032	-0.0114	-0.0037	-0.0072
		D	1.0587	1.0033	1.0261	1.0507	0.0104	0.0098	-0.0152	-0.0052
		mean	1.0619	1.0307	1.0157	1.0083	0.0029	-0.0007	0.0011	-0.0005
	1	1	1.0179	1.0063	0.9617	0.9548	0.0031	0.0096	0.0147	0.0018
		$D/2$	1.0321	0.9676	0.9729	0.9585	0.0001	-0.0105	-0.0043	-0.0102
		D	0.9999	0.9537	0.9767	0.9982	0.0107	0.0094	-0.0142	-0.0048
		mean	1.0127	0.9855	0.9721	0.9646	0.0029	-0.0007	0.0011	-0.0005
$(0, \frac{1}{2})$	0	1	1.1793	1.1647	1.1611	1.1576	0.0129	0.0081	-0.0081	0.0230
		$D/2$	1.2122	1.1542	1.1390	1.1272	-0.0133	0.0050	-0.0054	0.0017
		D	1.1934	1.1506	1.1787	1.1418	0.0033	-0.0056	-0.0015	-0.0127
		mean	1.2005	1.1668	1.1575	1.1501	0.0005	-0.0019	0.0006	-0.0009
	1	1	1.0824	1.0742	1.0586	1.0573	0.0071	0.0071	-0.0094	0.0235
		$D/2$	1.1081	1.0486	1.0460	1.0373	-0.0095	0.0108	0.0002	-0.0027
		D	1.0889	1.0581	1.0785	1.0497	0.0065	-0.0033	0.0013	-0.0099
		mean	1.0960	1.0687	1.0580	1.0507	0.0008	-0.0019	0.0006	-0.0009
$(\frac{1}{2}, \frac{1}{2})$	0	1	1.1896	1.1667	1.1708	1.1428	0.0010	-0.0031	0.0117	0.0016
		$D/2$	1.1806	1.1730	1.1923	1.1227	-0.0212	0.0034	-0.0051	-0.0193
		D	1.2222	1.1919	1.1554	1.1403	0.0163	0.0008	0.0196	0.0046
		mean	1.1962	1.1723	1.1559	1.1491	0.0017	-0.0013	0.0012	-0.0009
	1	1	1.1687	1.1580	1.1610	1.1336	-0.0035	-0.0035	0.0109	0.0038
		$D/2$	1.1504	1.1604	1.1837	1.1156	-0.0185	0.0054	-0.0049	-0.0190
		D	1.1977	1.1767	1.1455	1.1304	0.0155	0.0008	0.0201	0.0030
		mean	1.1787	1.1592	1.1448	1.1386	0.0016	-0.0013	0.0012	-0.0009

Tabla 4.3: $MSE_{22d}^{(a)}$ (izquierda) y $BIAS_{22d}^{(a)}$ (derecha) para $\rho_x = 1/2$, $a = 0, 1$.

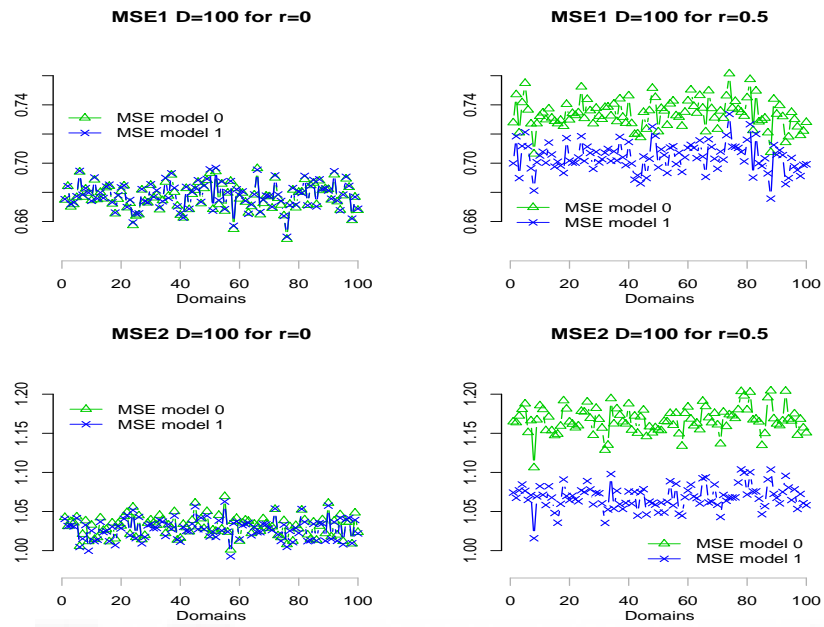


Figura 4.1: MSE_{drr} , para $a = 0, 1$, $r = 1, 2$, $(\rho_e, \rho) = (0, 0), (0, \frac{1}{2})$, $\rho_x = \frac{1}{2}$, $D = 100$.

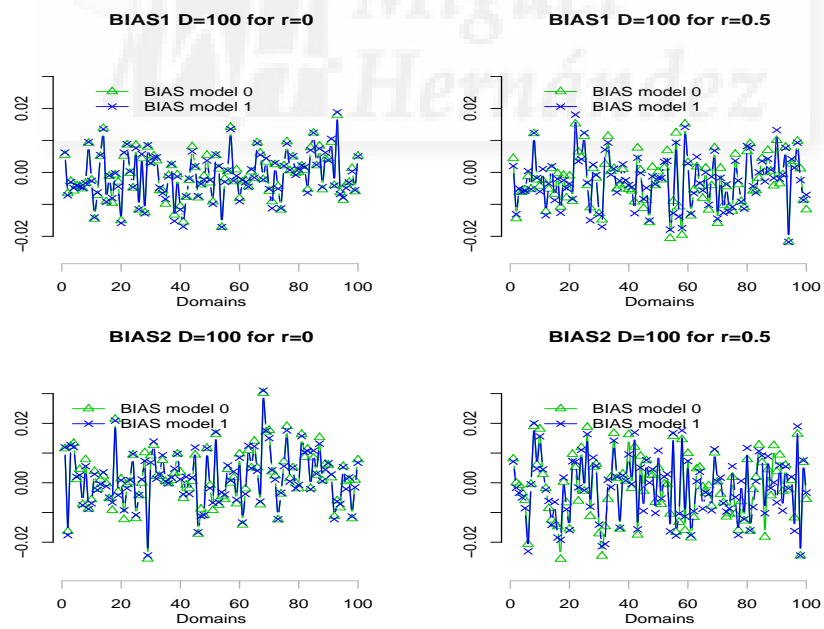


Figura 4.2: $BIAS_{drr}$, para $a = 0, 1$, $r = 1, 2$, $(\rho_e, \rho) = (0, 0), (0, \frac{1}{2})$, $\rho_x = \frac{1}{2}$, $D = 100$.

4.2.2. Experimento de simulación 2

El objetivo de este experimento es investigar empíricamente la pérdida de eficiencia en las estimaciones cuando no se tiene en cuenta la naturaleza multivariante de los datos. Para ello se simulan los datos o bien del modelo multivariante o bien del modelo producto de marginales restringido a $\sigma_{u1}^2 = \sigma_{u2}^2$, se estiman los parámetros y se calculan los EBLUP de ambos modelos. El experimento consiste en calcular por simulación Monte Carlo los sesgos y errores cuadráticos medios empíricos (B y E) de los estimadores analíticos (2.8) en la estimación de los errores cuadráticos medios del EBLUP de μ_{dr} .

En esta simulación se consideran los casos: (1) $\rho_e = 0, \rho = 0$, (2) $\rho_e = 1/2, \rho = 0$, (3) $\rho_e = 0, \rho = 1/2$ y (4) $\rho_e = 1/2, \rho = 1/2$. En el caso 1 los datos se simulan del modelo producto de modelos marginales, pero restringido a $\sigma_{u1}^2 = \sigma_{u2}^2$. El resto de parámetros son los mismos que en la simulación 1; es decir, $\beta_1 = 1, \beta_2 = 1, \sigma_{u1}^2 = 2$ y $\sigma_{u1}^2 = 4, \sigma_{d11} = 1, \sigma_{d22} = 2$. Las variables auxiliares x_{dr} también se generan de la misma forma que en la simulación 1.

Los pasos del experimento de simulación son

1. Repetir $I = 500$ veces ($i = 1, \dots, 500$)

1.1. Generar una muestra $(y_{dr}^{(i)}, x_{dr}^{(i)})$, $d = 1, \dots, D$, $r = 1, 2$.

1.2. Calcular $\{\hat{\beta}_{E1}^{(i,0)}, \hat{\beta}_{E2}^{(i,0)}, \hat{\sigma}_{u1}^{2(i,0)}, \hat{\sigma}_{u2}^{2(i,0)}\}$, $\{\hat{\beta}_{E1}^{(i,1)}, \hat{\beta}_{E2}^{(i,1)}, \hat{\sigma}_u^{2(i,1)}, \hat{\rho}^{(i,1)}\}$.

1.3. Para $d = 1, \dots, D$, $a = 1, 2$, $r = 1, 2$, calcular

$$mse_{drr}^{(i,a)} = g_{1drr}^{(i,a)}(\hat{\theta}^{(i,a)}) + g_{2drr}^{(i,a)}(\hat{\theta}^{(i,a)}) + 2g_{3drr}^{(i,a)}(\hat{\theta}^{(i,a)}),$$

donde $\hat{\theta}^{(i,0)} = (\hat{\sigma}_{u1}^{2(i,0)}, \hat{\sigma}_{u2}^{2(i,0)})$ y $\hat{\theta}^{(i,1)} = (\hat{\sigma}_u^{2(i,1)}, \hat{\rho}^{(i,1)})$.

2. Leer los valores $MSE_{drr}^{(a)}$ obtenidos en la simulación 1.

3. Salida:

$$B_{drr}^{(a)} = \frac{1}{I} \sum_{i=1}^I (mse_{drr}^{(i,a)} - MSE_{drr}^{(a)}), \quad E_{drr}^{(a)} = \frac{1}{I} \sum_{i=1}^I (mse_{drr}^{(i,a)} - MSE_{drr}^{(a)})^2, \quad r = 1, 2, a = 0, 1.$$

Las tablas 4.4 y 4.5 presentan los errores cuadráticos medios y los sesgos, $E_{drr}^{(a)}$ y $B_{drr}^{(a)}$, $a = 0, 1$, de los estimadores del error cuadrático medio de los EBLUP de las medias de las componentes $r = 1$ y $r = 2$ respectivamente. Las tablas están ordenadas por columnas y filas. La primera columna apunta cada uno de los modelos que generan los datos, la segunda columna especifica el modelo producto de univariantes ($a = 0$) o multivariante ($a = 1$) bajo el cual se calcula el EBLUP, la tercera columna señala el área $d = 1$, $d = D/2$, $d = D$ y el valor medio de todas las áreas, las cuatro columnas siguientes muestran el error cuadrático medio $E_{drr}^{(a)}$ y las cuatro últimas el sesgo $B_{drr}^{(a)}$. Cada uno de los dos grupos de cuatro columnas que se acaban de nombrar se corresponde con un valor distinto del número de áreas; es decir, $D = 50, 100, 200, 400$. Las filas se disponen en grupos de ocho, un grupo para cada uno de los casos que se están considerando. En el primer caso, los datos se simulan del modelo con $\rho = \rho_e = 0$, mientras que en los casos 2, 3 y 4 los datos se generan

del modelo multivariante con $\rho \neq 0$ o $\rho_e \neq 0$. Dentro de cada grupo, las cuatro primeras filas se corresponden con las estimaciones para el modelo $a = 0$ y las cuatro últimas con las estimaciones para el modelo $a = 1$.

En las tablas 4.4 y 4.5 se observa, en el primer caso ($\rho_e = 0, \rho = 0$), unos errores cuadráticos medios menores para el modelo $a = 1$. Eso se debe a que el modelo multivariante $a = 1$ es menos incorrecto que el modelo producto de univariantes $a = 0$. En los casos tercero y cuarto se observan unos errores cuadráticos medios menores para el modelo multivariante. En el caso segundo de la tabla 4.4 se aprecian unos errores cuadráticos medios menores para el modelo $a = 0$. En ambas tablas no se aprecian diferencias significativas en los sesgos.

En la figura 4.3 se vuelve a observar lo mismo que se ha apuntado en el párrafo anterior. Cabe añadir, para el caso $D = 100$, que se observa con claridad en el gráfico 4.4 que los sesgos para el modelo $a = 0$ aumentan cuando se cambia del caso primero al caso tercero.

La figura 4.3 muestra las gráficas los valores de $E_{drr}^{(a)}$, $a = 0, 1$, $r = 1, 2$, $d = 1, \dots, D$, $D = 100$. La figura está dividida en cuatro partes y tiene una disposición en forma de tabla con dos filas y dos columnas. La primera y segunda fila presentan los valores de $E_{drr}^{(a)}$ para $r = 1$ y $r = 2$ respectivamente. Las primera y segunda columna presentan los valores de $E_{drr}^{(a)}$ cuando los datos se generan del modelo definido por los casos 1 y 3 respectivamente. Cada una de las cuatro sub-figuras muestran los valores de $E_{drr}^{(a)}$ para $a = 0$ y $a = 1$.

En la figura 4.3 se observa que la diferencia de los errores cuadráticos medios entre los modelos $a = 0$ y $a = 1$, cuando ρ_e cambia del caso 1 al caso 3, es bastante pronunciada. También hay que destacar que si los datos se generan del modelo $a = 0$ con $\rho_e = 0$, entonces se aprecia un aumento del error cuadrático medio de $mse_{dr}^{(a)}$ al utilizar el modelo $a = 1$, siendo la diferencia mayor para la segunda variable del modelo.

La figura 4.4 grafica los valores de $B_{drr}^{(a)}$, $a = 0, 1$, $r = 1, 2$, $d = 1, \dots, D$, $D = 100$. Esta figura está estructurada de la misma forma que la figura 4.1. En la figura 4.4 se observa una leve diferencia en los sesgos de los modelos $a = 0$ y $a = 1$ cuando se cambia del caso 1 al caso 3. También hay que destacar que si los datos se generan del modelo definido por el caso 1, entonces no se aprecia un aumento de sesgo al utilizar el modelo sobre-parametrizado $a = 1$.

(ρ_e, ρ)	a	d	E				B			
			50	100	200	400	50	100	200	400
(0, 0)	0	1	0.0047	0.0022	0.0011	0.0005	0.0049	-0.0002	-0.0001	0.0004
		$D/2$	0.0047	0.0026	0.0011	0.0005	0.0042	-0.0183	0.0047	-0.0056
		D	0.0047	0.0023	0.0011	0.0006	-0.0059	0.0075	0.0000	-0.0096
		mean	0.0048	0.0023	0.0012	0.0006	0.0021	-0.0019	-0.0010	-0.0008
	1	1	0.0048	0.0029	0.0013	0.0007	0.0003	0.0050	-0.0008	-0.0001
		$D/2$	0.0028	0.0018	0.0007	0.0004	0.0024	-0.0208	0.0033	-0.0067
		D	0.0029	0.0014	0.0007	0.0004	-0.0076	0.0065	-0.0004	-0.0102
		mean	0.0029	0.0014	0.0008	0.0004	0.0014	-0.0029	-0.0020	-0.0015
$(\frac{1}{2}, 0)$	0	1	0.0039	0.0021	0.0011	0.0006	0.0005	-0.0163	-0.0038	0.0005
		$D/2$	0.0040	0.0019	0.0011	0.0006	-0.0096	-0.0025	-0.0020	-0.0004
		D	0.0039	0.0020	0.0014	0.0006	0.0072	0.0125	0.0195	-0.0006
		mean	0.0040	0.0020	0.0012	0.0006	-0.0015	0.0056	0.0015	-0.0013
	1	1	0.0038	0.0046	0.0031	0.0015	0.0008	0.0101	-0.0113	-0.0010
		$D/2$	0.0045	0.0030	0.0015	0.0007	-0.0011	0.0005	-0.0030	-0.0036
		D	0.0045	0.0031	0.0018	0.0007	0.0083	0.0134	0.0157	-0.0047
		mean	0.0047	0.0031	0.0016	0.0008	0.0070	0.0031	-0.0002	-0.0018
$(0, \frac{1}{2})$	0	1	0.0031	0.0015	0.0009	0.0004	-0.0160	0.0063	-0.0114	-0.0010
		$D/2$	0.0029	0.0017	0.0008	0.0004	-0.0126	0.0128	0.0079	0.0047
		D	0.0030	0.0015	0.0009	0.0005	-0.0133	0.0066	-0.0136	-0.0086
		mean	0.0030	0.0016	0.0009	0.0005	-0.0046	0.0008	-0.0018	0.0006
	1	1	0.0025	0.0012	0.0007	0.0004	-0.0085	0.0078	-0.0062	-0.0092
		$D/2$	0.0027	0.0013	0.0008	0.0003	-0.0149	0.0109	0.0124	-0.0022
		D	0.0026	0.0012	0.0009	0.0003	-0.0144	0.0091	-0.0147	-0.0056
		mean	0.0026	0.0012	0.0007	0.0004	-0.0009	0.0038	-0.0011	-0.0005
$(\frac{1}{2}, \frac{1}{2})$	0	1	0.0034	0.0018	0.0008	0.0004	-0.0021	-0.0204	0.0024	0.0014
		$D/2$	0.0034	0.0015	0.0014	0.0004	0.0019	0.0082	-0.0260	0.0071
		D	0.0034	0.0015	0.0008	0.0004	-0.0038	-0.0117	-0.0032	0.0025
		mean	0.0035	0.0015	0.0009	0.0005	-0.0056	-0.0002	0.0000	0.0002
	1	1	0.0027	0.0017	0.0007	0.0003	0.0013	-0.0180	0.0043	-0.0005
		$D/2$	0.0027	0.0014	0.0013	0.0004	0.0063	0.0110	-0.0255	0.0074
		D	0.0027	0.0015	0.0007	0.0003	0.0047	-0.0113	-0.0029	0.0004
		mean	0.0028	0.0014	0.0008	0.0004	-0.0005	0.0003	0.0004	0.0001

Tabla 4.4: $E_{11d}^{(a)}$ (izquierda) y $B_{11d}^{(a)}$ (derecha) para $\rho_x = 1/2$, $a = 0, 1$.

		E				B				
(ρ_e, ρ)	a	d	50	100	200	400	50	100	200	400
(0,0)	0	1	0.0373	0.0194	0.0106	0.0050	0.0123	-0.0249	-0.0153	-0.0108
		$D/2$	0.0370	0.0187	0.0104	0.0049	0.0074	-0.0007	-0.0007	-0.0028
		D	0.0367	0.0187	0.0104	0.0050	-0.0002	-0.0034	0.0075	-0.0103
		mean	0.0372	0.0190	0.0106	0.0051	0.0040	-0.0123	-0.0066	-0.0053
	1	1	0.0151	0.0074	0.0035	0.0017	0.0087	-0.0208	-0.0132	-0.0080
		$D/2$	0.0151	0.0070	0.0034	0.0017	0.0096	0.0025	0.0024	-0.0029
		D	0.0149	0.0069	0.0035	0.0017	0.0050	0.0002	0.0116	-0.0068
		mean	0.0152	0.0072	0.0036	0.0019	0.0044	-0.0070	-0.0028	-0.0028
$(\frac{1}{2}, 0)$	0	1	0.0399	0.0205	0.0108	0.0053	0.0057	-0.0121	0.0006	0.0084
		$D/2$	0.0398	0.0204	0.0108	0.0052	0.0104	0.0123	-0.0008	0.0006
		D	0.0399	0.0211	0.0111	0.0075	0.0210	0.0297	-0.0182	-0.0473
		mean	0.0401	0.0205	0.0111	0.0055	0.0149	0.0008	-0.0086	-0.0053
	1	1	0.0191	0.0125	0.0061	0.0029	0.0086	-0.0150	0.0080	0.0060
		$D/2$	0.0189	0.0128	0.0060	0.0028	-0.0038	0.0245	-0.0028	0.0025
		D	0.0198	0.0138	0.0060	0.0042	0.0317	0.0401	-0.0058	-0.0369
		mean	0.0193	0.0125	0.0062	0.0030	0.0159	0.0068	-0.0019	-0.0036
$(0, \frac{1}{2})$	0	1	0.0269	0.0145	0.0069	0.0037	0.0194	0.0070	-0.0062	-0.0140
		$D/2$	0.0266	0.0148	0.0072	0.0037	-0.0123	0.0182	0.0162	0.0166
		D	0.0263	0.0150	0.0074	0.0035	0.0094	0.0232	-0.0228	0.0023
		mean	0.0266	0.0148	0.0072	0.0038	-0.0002	0.0057	-0.0022	-0.0062
	1	1	0.0176	0.0077	0.0043	0.0021	0.0133	0.0024	-0.0027	-0.0079
		$D/2$	0.0175	0.0085	0.0044	0.0022	-0.0109	0.0288	0.0103	0.0123
		D	0.0174	0.0081	0.0047	0.0021	0.0115	0.0208	-0.0215	0.0003
		mean	0.0175	0.0080	0.0045	0.0023	0.0016	0.0089	-0.0017	-0.0011
$(\frac{1}{2}, \frac{1}{2})$	0	1	0.0281	0.0141	0.0078	0.0040	0.0121	0.0078	-0.0160	0.0044
		$D/2$	0.0284	0.0141	0.0090	0.0046	0.0226	0.0021	-0.0371	0.0247
		D	0.0279	0.0142	0.0076	0.0041	-0.0162	-0.0154	0.0005	0.0074
		mean	0.0282	0.0144	0.0078	0.0043	0.0072	0.0029	-0.0006	-0.0017
	1	1	0.0128	0.0061	0.0033	0.0015	0.0073	0.0011	-0.0177	0.0037
		$D/2$	0.0134	0.0061	0.0046	0.0019	0.0270	-0.0006	-0.0400	0.0218
		D	0.0129	0.0063	0.0030	0.0015	-0.0173	-0.0154	-0.0012	0.0074
		mean	0.0130	0.0064	0.0033	0.0017	-0.0010	0.0008	-0.0011	-0.0011

Tabla 4.5: $E_{22d}^{(a)}$ (izquierda) y $B_{22d}^{(a)}$ (derecha) para $\rho_x = 1/2$, $a = 0, 1$.

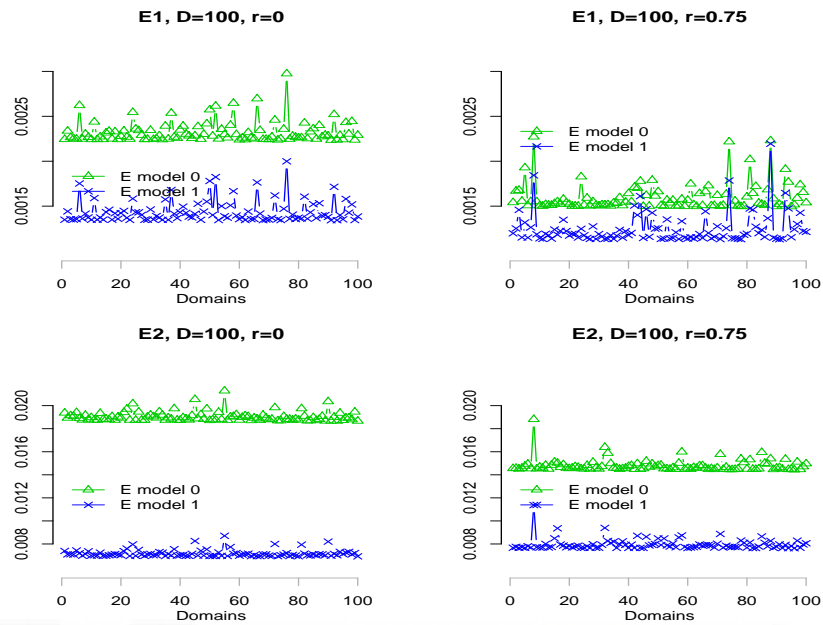


Figura 4.3: E_{drr} , para $a = 0, 1$, $r = 1, 2$, $(\rho_e, \rho) = (0, 0), (0, \frac{1}{2})$, $\rho_x = \frac{1}{2}$, $D = 100$.

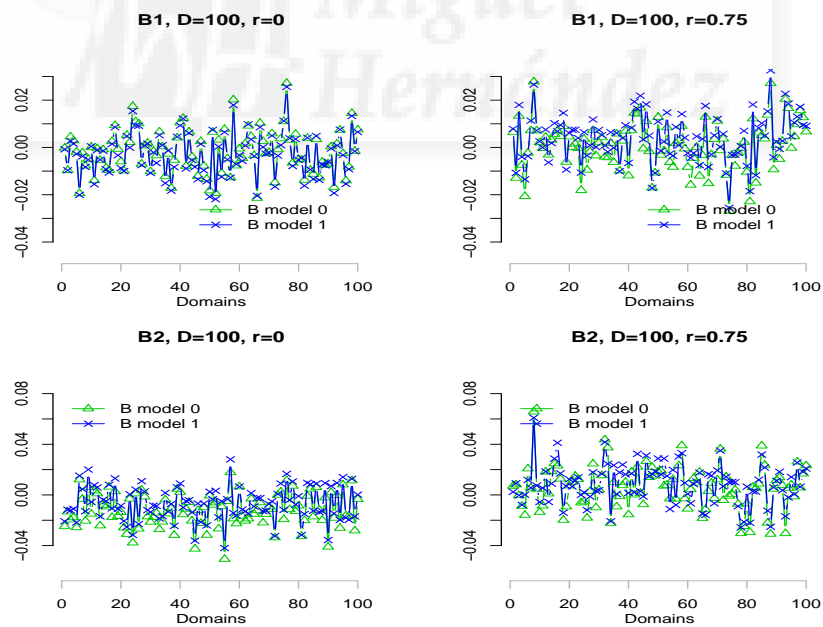


Figura 4.4: B_{drr} , para $a = 0, 1$, $r = 1, 2$, $(\rho_e, \rho) = (0, 0), (0, \frac{1}{2})$, $\rho_x = \frac{1}{2}$, $D = 100$.

4.2.3. Experimento de simulación 3

El objetivo de este experimento es comprobar el funcionamiento del bootstrap paramétrico en la estimación de los errores cuadráticos medios del EBLUP de las medias poblacionales en un modelo multivariante AR(1). En esta simulación se consideran las correlaciones $\rho_e = \rho = 1/2$ del caso 4. Los valores de los restantes parámetros son los mismos que en la simulación 1; es decir, $\beta_1 = 1$, $\beta_2 = 1$, $\sigma_u^2 = 2$, $\sigma_{d11} = 1$ y $\sigma_{d22} = 2$. En este apartado se usa la notación $\theta = (\theta_1, \theta_2)$, donde $\theta_1 = \sigma_u^2$ y $\theta_2 = \rho$. Las variables auxiliares x_{dr} se generan de la misma forma que en la simulación 1.

Los pasos del experimento de simulación son

1. Repetir $I = 500$ veces ($i = 1, \dots, 500$)

1.1. Generar una muestra $(y_{dr}^{(i)}, x_{dr}^{(i)})$, $d = 1, \dots, D$, $r = 1, 2$ (cf. A-D en Sección 1).

1.2. Calcular $\mu_d^{(i)} = X_d^{(i)} \beta + I_2 u_d^{(i)}$.

1.3. Calcular $\hat{\sigma}_u^{2(i)}$, $\hat{\rho}^{(i)}$, $\hat{\beta}_{E1}^{(i)}$ y $\hat{\beta}_{E2}^{(i)}$.

1.4. Para $d = 1, \dots, D$, calcular $\hat{u}_{Ed}^{(i)}$, usando $\hat{\sigma}_u^{2(i)}$, $\hat{\rho}^{(i)}$, $\hat{\beta}_{Er}^{(i)}$. Calcular

$$\hat{\mu}_d^{(i)} = X_d^{(i)} \hat{\beta}_E^{(i)} + I_2 \hat{u}_d^{(i)}, \text{mse}_d^{(i)} = G_{1d}^{(i)}(\hat{\sigma}_u^{2(i)}, \hat{\rho}^{(i)}) + G_{2d}^{(i)}(\hat{\sigma}_u^{2(i)}, \hat{\rho}^{(i)}) + 2G_{3d}^{(i)}(\hat{\sigma}_u^{2(i)}, \hat{\rho}^{(i)}).$$

1.5. Repetir $B = 200$ veces ($b = 1, \dots, B$)

1.5.1. Generar $u_d^{*(ib)}$, $e_{dr}^{*(ib)}$, $d = 1, \dots, D$, $r = 1, 2$ (cf. B-C en Sección 1), pero usando $\hat{\sigma}_u^{2(i)}$ y $\hat{\rho}^{(i)}$ en lugar de σ_u^2 y ρ .

1.5.2. Generar una muestra bootstrap $(y_{dr}^{*(ib)}, x_{dr}^{*(ib)})$, $d = 1, \dots, D$, $r = 1, 2$, del modelo

$$y_{dr}^{*(ib)} = x_{dr}^{(i)} \hat{\beta}_{Rr}^{(i)} + u_{dr}^{*(ib)} + e_{dr}^{*(ib)}.$$

1.5.3. Calcular $\mu_d^{*(ib)} = X_d^{(i)} \hat{\beta}_E^{(i)} + u_d^{*(ib)}$.

1.5.4. Calcular $\hat{\theta}^{*(ib)}$ a partir de $\hat{\theta}$, reemplazando convenientemente los elementos de la muestra bootstrap.

1.5.5. Calcular $\hat{\beta}_{Br}^{*(ib)}$ y $\hat{\beta}_{Er}^{*(ib)}$, las versiones bootstrap $\hat{\beta}_{Br}$ y $\hat{\beta}_{Er}$, $r = 1, 2$, respectivamente. Calculados usando $\hat{V}_d^{(i)}$ e $y_d^{*(ib)}$ para el cálculo de $\hat{\beta}_{Br}^{*(ib)}$, y $\hat{V}_d^{*(ib)}$ e $y_d^{*(ib)}$ para el cálculo de $\hat{\beta}_{Er}^{*(ib)}$.

1.5.6. Para $d = 1, \dots, D$ y $r = 1, 2$; calcular $\hat{u}_{Bd}^{*(ib)}$ y $\hat{u}_d^{*(ib)}$, a partir de $\hat{\theta}^{(i)}$ y $\hat{\beta}_{Br}^{*(ib)}$, $\hat{\theta}^{*(ib)}$ y $\hat{\beta}_{Er}^{*(ib)}$, $r=1,2$ respectivamente.

1.5.7. Para $d = 1, \dots, D$, calcular

$$\hat{\mu}_{Bd}^{*(ib)} = X_d^{(i)} \hat{\beta}_B^{*(ib)} + I_2 \hat{u}_{Bd}^{*(ib)} \quad \text{y} \quad \hat{\mu}_{Ed}^{*(ib)} = X_d^{(i)} \hat{\beta}_E^{*(ib)} + I_2 \hat{u}_d^{*(ib)}.$$

1.5.8. Para $d = 1, \dots, D$, calcular

$$\delta_{Ed}^{*(ib)} = (\hat{\mu}_d^{*(ib)} - \mu_d^{*(ib)}), \quad \delta_{Bd}^{*(ib)} = (\hat{\mu}_{Bd}^{*(ib)} - \mu_d^{*(ib)}), \quad \delta_{EBd}^{*(ib)} = (\hat{\mu}_d^{*(ib)} - \hat{\mu}_{Bd}^{*(ib)}).$$

1.6 Para $d = 1, \dots, D$, calcular

$$\begin{aligned} mse_d^{*1(i)} &= \frac{1}{B} \sum_{b=1}^B \delta_{Ed}^{*(ib)} \delta_{Ed}^{*(ib)t} \\ mse_d^{*2(i)} &= G_{1d}^{(i)}(\hat{\theta}^{(i)}) + G_{2d}^{(i)}(\hat{\theta}^{(i)}) + \frac{1}{B} \sum_{b=1}^B \delta_{EBd}^{*(ib)} \delta_{EBd}^{*(ib)t} \\ mse_d^{*3(i)} &= 2[G_{1d}^{(i)}(\hat{\theta}^{(i)}) + G_{2d}^{(i)}(\hat{\theta}^{(i)})] - \frac{1}{B} \sum_{b=1}^B [G_1(\hat{\theta}^{*(ib)}) + G_2(\hat{\theta}^{*(ib)})] + \frac{1}{B} \sum_{b=1}^B \delta_{EBd}^{*(ib)} \delta_{EBd}^{*(ib)t}. \end{aligned}$$

2. Salida:

$$mse_d = \frac{1}{I} \sum_{i=1}^I mse_d^{(i)}, \quad mse_d^{*\ell} = \frac{1}{I} \sum_{i=1}^I mse_d^{*\ell(i)}, \quad \ell = 1, 2, 3.$$

3. Leer los MSE_{drr} obtenidos en la simulación 1 para el caso $\rho = \rho_e = \frac{1}{2}$ y hacer

$$\begin{aligned} B_{drr}^0 &= \frac{1}{I} \sum_{i=1}^I (mse_{dr}^{(i)} - MSE_{drr}), \quad B_{drr}^{*\ell} = \frac{1}{I} \sum_{i=1}^I (mse_{dr}^{*\ell(i)} - MSE_{drr}), \quad \ell = 1, 2, 3, \quad r = 1, 2, \\ E_{drr}^0 &= \frac{1}{I} \sum_{i=1}^I (mse_{dr}^{(i)} - MSE_{drr})^2, \quad E_{drr}^{*\ell} = \frac{1}{I} \sum_{i=1}^I (mse_{dr}^{*\ell(i)} - MSE_{drr})^2, \quad \ell = 1, 2, 3, \quad r = 1, 2, \\ B_{rr}^0 &= \frac{1}{D} \sum_{d=1}^D B_{drr}^0, \quad B_{rr}^{*\ell} = \frac{1}{D} \sum_{d=1}^D B_{drr}^{*\ell}, \quad E_{rr}^0 = \frac{1}{D} \sum_{d=1}^D E_{drr}^0, \quad E_{rr}^{*\ell} = \frac{1}{D} \sum_{d=1}^D E_{drr}^{*\ell}, \quad \ell = 1, 2, 3, \quad r = 1, 2. \end{aligned}$$

D	E_{11}^0	E_{11}^{*1}	E_{11}^{*2}	E_{11}^{*3}	E_{22}^0	E_{22}^{*1}	E_{22}^{*2}	E_{22}^{*3}
50	0.00292	0.00863	0.00347	0.00297	0.01306	0.02824	0.01502	0.01347
100	0.00147	0.00674	0.00160	0.00146	0.00621	0.01969	0.00674	0.00618
200	0.00077	0.00600	0.00080	0.00077	0.00314	0.01635	0.00325	0.00315
400	0.00044	0.00567	0.00045	0.00044	0.00182	0.01476	0.00185	0.00183

Tabla 4.6: $E_{rr}^0, E_{rr}^{*\ell}, \ell = 1, 2, 3$.

D	B_{11}^0	B_{11}^{*1}	B_{11}^{*2}	B_{11}^{*3}	B_{22}^0	B_{22}^{*1}	B_{22}^{*2}	B_{22}^{*3}
50	-0.00240	-0.01591	-0.01542	-0.00164	0.00073	-0.03057	-0.03096	0.00139
100	-0.00184	-0.00790	-0.00808	-0.00164	-0.00371	-0.01828	-0.01905	-0.00332
200	-0.00041	-0.00365	-0.00348	-0.00021	-0.00029	-0.00877	-0.00788	0.00009
400	0.00075	-0.00074	-0.00077	0.00070	-0.00040	-0.00414	-0.00416	-0.00050

Tabla 4.7: $B_{rr}^0, B_{rr}^{*\ell}, \ell = 1, 2, 3$.

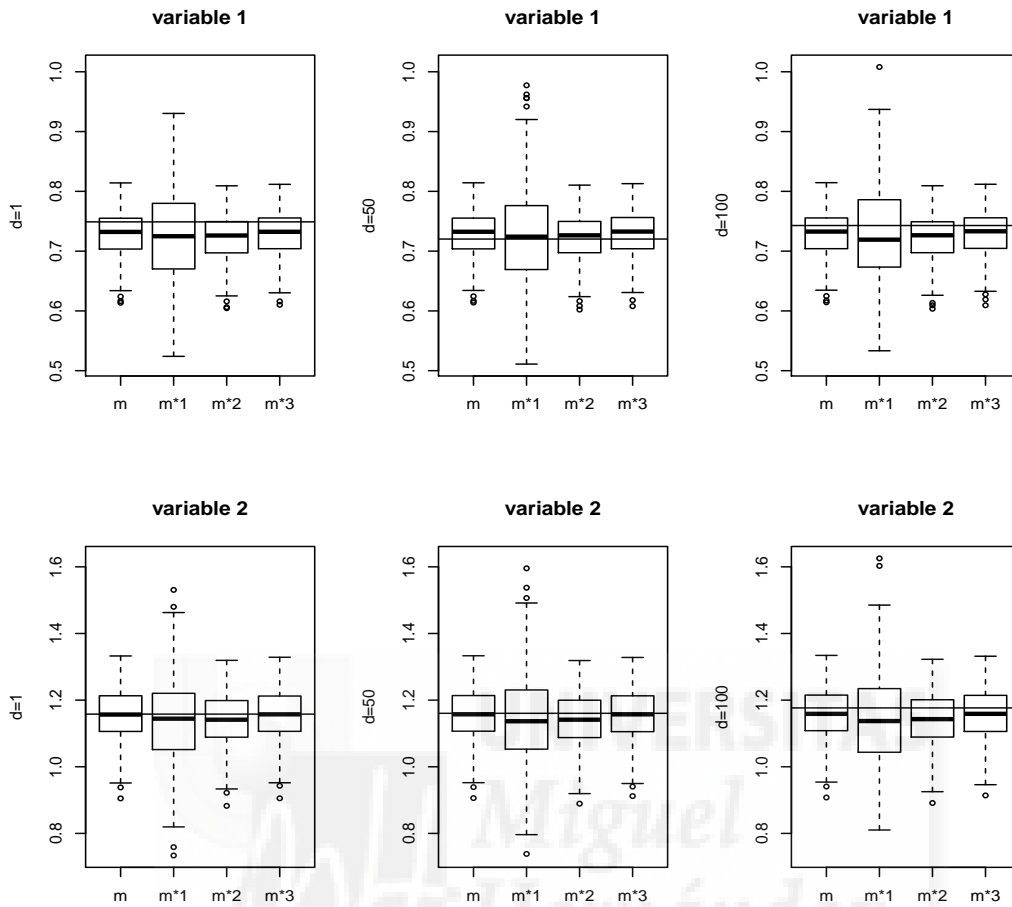


Figura 4.5: Diagrama de cajas de $msed_{drr}$ y $msed_{drr}^{\ell}$, $\ell = 1, 2, 3$, para $\rho_e = \rho = \rho_x = \frac{1}{2}$, $D = 100$.

En la tabla 4.6 se apunta en la primera columna el número de áreas consideradas en la simulación; es decir, $D = 50$, $D = 100$, $D = 200$ y $D = 400$. En las cuatro columnas siguientes se muestra el error cuadrático medio de los estimadores $msed$, $msed^{*1}$, $msed^{*2}$ y $msed^{*3}$ del error cuadrático medio de $\hat{\mu}_{d1}$, donde el valor teórico considerado es el obtenido en la simulación 1. En las cuatro columnas siguientes se muestra lo mismo, pero esta vez para el estimador $\hat{\mu}_{d2}$. La disposición de las columnas en la tabla 4.7 es la misma pero esta vez para el sesgo.

En la tabla 4.6 se observa que el estimador $msed^{*1}$ produce los mayores errores y el estimador $msed^{*3}$ los menores. Asimismo se observa que el error disminuye de forma considerable en los cuatro estimadores al aumentar el número de áreas consideradas. En la tabla 4.7 se observa que los sesgos para los estimadores $msed^{*1}$ y $msed^{*2}$ son negativos, lo cual era de esperar debido a que les afecta que el valor esperado del término $G_1(\hat{\theta})$, que es de forma aproximada $G_1(\theta) - G_3(\theta)$. También se aprecia una presencia mayor del sesgo negativo en el estimador $msed^{*3}$ posiblemente debida a la naturaleza de la matriz de varianzas de los efectos aleatorios del modelo, ya que dicha matriz afecta considerablemente a la estimación del término G_3 y consecuentemente a sus correcciones.

La figura 4.5 contiene los diagramas de cajas de los cuatro estimadores considerados. Se observa que la distribución del estimador $msed^{*3}$ es mejor que la de los tres restantes. También se nota de forma evidente la presencia de sesgo negativo para los estimadores $msed^{*1}$ y $msed^{*2}$ y se aprecia una ligera tendencia negativa para el sesgo del estimador $msed^{*3}$ sobre todo en la primera variable del modelo.

4.2.4. Experimento de simulación 4

En el presente apartado se realizan una simulación adicional relacionada con la simulación 1, en ella se comparan el modelo multivariante ($a = 1$) para el caso ($\rho_e = 0, \rho = 0$) con el modelo producto de marginales sin la restricción $\sigma_{u1}^2 = \sigma_{u2}^2$ ($a = 0$); es decir, como en el *modelo diagonal* que se estudió en el capítulo segundo. El objetivo de este experimento es comparar la eficiencia de las estimaciones cuando se utiliza el modelo AR(1) cuando $\rho = 0$ y $\rho_e = 0$ con la de las estimaciones del modelo producto de marginales que se deduce del modelo con matriz de varianzas de los efectos diagonal.

Los datos se simulan teniendo en cuenta el modelo producto de marginales del modelo diagonal se estiman los parámetros y se calculan los EBLUP de ambos modelos. Hay dos conjuntos de parámetros estimados y EBLUP calculados. Según sea el caso, un conjunto se obtiene bajo el modelo correcto y otro bajo el modelo incorrecto. Cabe esperar que los mejores resultados se obtengan siempre cuando se usan los estimadores correspondientes al modelo correcto. El experimento consiste en calcular por simulación Monte Carlo los sesgos y errores cuadráticos medios empíricos (BIAS y MSE) de los estimadores de los parámetros y de los EBLUP.

Se considera $\sigma_{u1}^2 = 2$, $\sigma_{u2}^2 = 4$, $\rho_e = 0$ y $\rho = 0$. Los valores de los restantes parámetros son los mismos que en la simulación 1; es decir, $\beta_1 = 1$, $\beta_2 = 1$, $\sigma_u^2 = 2$, $\sigma_{d11} = 1$ y $\sigma_{d22} = 2$. Las variables auxiliares x_{dr} se generan de la misma forma que en la simulación 1.

Los pasos del experimento de simulación son

1. Repetir $I = 10^4$ veces ($i = 1, \dots, I$)
 - 1.1. Generar una muestra (y_{dr}, x_{dr}) , $d = 1, \dots, D$, $r = 1, 2$, del modelo $a = 0$.
 - 1.2. Calcular $\{\hat{\beta}_{E1}^{(i,0)}, \hat{\beta}_{E2}^{(i,0)}, \hat{\sigma}_{u1}^{2(i,0)}, \hat{\sigma}_{u2}^{2(i,0)}\}$, $\{\hat{\beta}_{E1}^{(i,1)}, \hat{\beta}_{E2}^{(i,1)}, \hat{\sigma}_u^{2(i,1)}, \hat{\rho}^{(i,1)}\}$ y $\{\hat{\mu}_d^{(i,0)}, \hat{\mu}_d^{(i,1)}\}$, donde los superíndices "0" y "1" se usan para denotar los estimadores y EBLUPs calculados asumiendo el modelo producto de marginales ($a = 0$) o el modelo multivariante AR(1) ($a = 1$).
2. Salida: Para todo $\tau \in \{\beta_1^{(0)}, \beta_2^{(0)}, \sigma_{u1}^{2(0)}, \sigma_{u2}^{2(0)}, \beta_1^{(1)}, \beta_2^{(1)}, \sigma_u^{2(1)}, \hat{\rho}^{(1)}\}$, calcular

$$MSE(\hat{\tau}) = \frac{1}{I} \sum_{i=1}^I (\hat{\tau}^{(i)} - \tau)^2, \quad BIAS(\hat{\tau}) = \frac{1}{I} \sum_{i=1}^I (\hat{\tau}^{(i)} - \tau),$$

$$MSE_{rrd}^{(a)} = \frac{1}{I} \sum_{i=1}^I (\hat{\mu}_{rd}^{(i,a)} - \mu_{rd}^{(i,a)})^2, \quad BIAS_{rrd}^{(a)} = \frac{1}{I} \sum_{i=1}^I (\hat{\mu}_{rd}^{(i,a)} - \mu_{rd}^{(i,a)}), \quad d = 1, D/2, D, r = 1, 2, a = 0, 1.$$

a	D	MSE				BIAS			
		50	100	200	400	50	100	200	400
0	$\hat{\beta}_{E1}^{(0)}$	0,0006	0,0003	0,0002	0,0000	0,0000	0,0002	0,0000	0,0001
	$\hat{\beta}_{E2}^{(0)}$	0,0011	0,0006	0,0003	0,0001	-0,0002	0,0002	0,0000	0,0000
	$\hat{\sigma}_{u1}^{2(0)}$	0,3737	0,1799	0,0919	0,0446	0,0036	-0,0035	0,0002	0,0024
	$\hat{\sigma}_{u2}^{2(0)}$	1,4976	0,7129	0,3635	0,1805	-0,0085	0,0106	-0,0032	-0,0080
1	$\hat{\beta}_{E1}^{(1)}$	0,0006	0,0003	0,0002	0,0000	0,0000	0,0002	0,0000	0,0001
	$\hat{\beta}_{E2}^{(1)}$	0,0011	0,0006	0,0003	0,0001	-0,0002	0,0002	0,0000	0,0000
	$\hat{\sigma}_u^{2(1)}$	0,8559	0,7050	0,6471	0,6157	0,6543	0,7119	0,7389	0,7532
	$\hat{\rho}^{(1)}$	0,0432	0,0218	0,0106	0,0051	0,0011	0,0006	0,0015	0,0003

Tabla 4.8: MSE (izquierda) y BIAS (derecha) para $\rho_e = \rho = 0$, $\rho_x = 1/2$.

a	d	MSE				BIAS			
		50	100	200	400	50	100	200	400
0	1	0,6790	0,6741	0,6755	0,6749	0,0020	0,0088	-0,0062	-0,0077
	$D/2$	0,6826	0,6771	0,6668	0,6780	-0,0064	0,0125	-0,0095	-0,0060
	D	0,7007	0,6762	0,6840	0,6800	-0,0120	-0,0022	-0,0127	-0,0019
	mean	0,6870	0,6767	0,6711	0,6693	-0,0008	0,0015	0,0008	0,0004
1	1	0,6947	0,6882	0,6890	0,6898	0,0030	0,0068	-0,0025	-0,0088
	$D/2$	0,6933	0,6918	0,6799	0,6889	-0,0057	0,0114	-0,0107	-0,0058
	D	0,7103	0,6909	0,6952	0,6906	-0,0110	-0,0042	-0,0134	-0,0026
	mean	0,6982	0,6898	0,6845	0,6828	-0,0008	0,0015	0,0008	0,0004

Tabla 4.9: $MSE_{11d}^{(a)}$ (izquierda) y $BIAS_{11d}^{(a)}$ (derecha) para $\rho_e = \rho = 0$, $\rho_x = 1/2$.

a	d	MSE				BIAS			
		50	100	200	400	50	100	200	400
0	1	1,3609	1,3790	1,3304	1,3388	0,0187	-0,0119	-0,0184	-0,0044
	$D/2$	1,3783	1,3421	1,3276	1,3254	0,0115	0,0098	-0,0111	0,0151
	D	1,3887	1,3638	1,3117	1,3012	0,0011	-0,0020	0,0175	-0,0090
	mean	1,3740	1,3519	1,3435	1,3388	-0,0007	0,0002	0,0000	0,0004
1	1	1,4061	1,4286	1,3806	1,3840	0,0148	-0,0144	-0,0204	-0,0027
	$D/2$	1,4253	1,3772	1,3663	1,3769	0,0094	0,0072	-0,0130	0,0134
	D	1,4308	1,4081	1,3559	1,3367	0,0000	-0,0041	0,0138	-0,0056
	mean	1,4176	1,3979	1,3884	1,3835	-0,0007	0,0002	0,0000	0,0004

Tabla 4.10: $MSE_{22d}^{(a)}$ (izquierda) y $BIAS_{22d}^{(a)}$ (derecha) para $\rho_e = \rho = 0$, $\rho_x = 1/2$.

La información que se presenta en las tablas 4.8, 4.9 y 4.10 se dispone de la misma forma que la que se presenta en las tablas 4.1, 4.2 y 4.3, en la simulación que se está contemplando solo se atiende el caso primero, es decir, $(\rho_e = 0, \rho = 0)$. De la misma forma que en el resto de simulaciones, se puede ver en la tabla 4.8 que los errores cuadráticos medios y sesgos de $\hat{\beta}_{E1}$ y $\hat{\beta}_{E2}$ son básicamente iguales para las estimaciones que se hacen asumiendo los modelos $a = 0$ o $a = 1$. Para los estimadores de las varianzas y del parámetro de correlación, ρ , se observa que sus errores cuadráticos medios son siempre menores cuando se asume el modelo producto de marginales ($a = 0$); respecto a los sesgos no se aprecia una diferencia notable. En las tablas 4.9 y 4.10 se observa que los errores cuadráticos medios de los estimadores de $\hat{\mu}_{d1}$ y $\hat{\mu}_{d2}$ son notablemente menores para el modelo producto de marginales. Lo anterior era lo esperado, ya que el modelo correcto es el producto de marginales sin la restricción $\sigma_{u1}^2 = \sigma_{u2}^2$ que se impone en el modelo AR(1) como consecuencia de la matriz de varianzas de los efectos aleatorios 4.2 De la misma forma que se ha hecho para las tablas, para la figura 4.6, se puede afirmar, con el objeto de no repetir aclaraciones, que en ella se dispone los resultados obtenidos de la misma forma que en las figuras 4.1, pero solo para el caso $(\rho_e = 0, \rho = 0)$.

En la figura 4.6 se muestra de forma gráfica lo que ya se ha visto en las tablas 4.9 y 4.10, es decir, que los errores cuadráticos medios que produce el modelo producto de univariantes ($a = 0$) son menores que los que produce el modelo multivariante ($a = 1$). Dados los resultados que ofrecen la simulación que se ha planteado, se puede concluir que el ajuste del modelo AR(1), cuando se parte de unos datos simulados de un modelo univariante, produce estimaciones con mayores errores que el del modelo producto de marginales; en los sesgos no se aprecia una diferencia destacable entre los dos modelos que se han considerado.

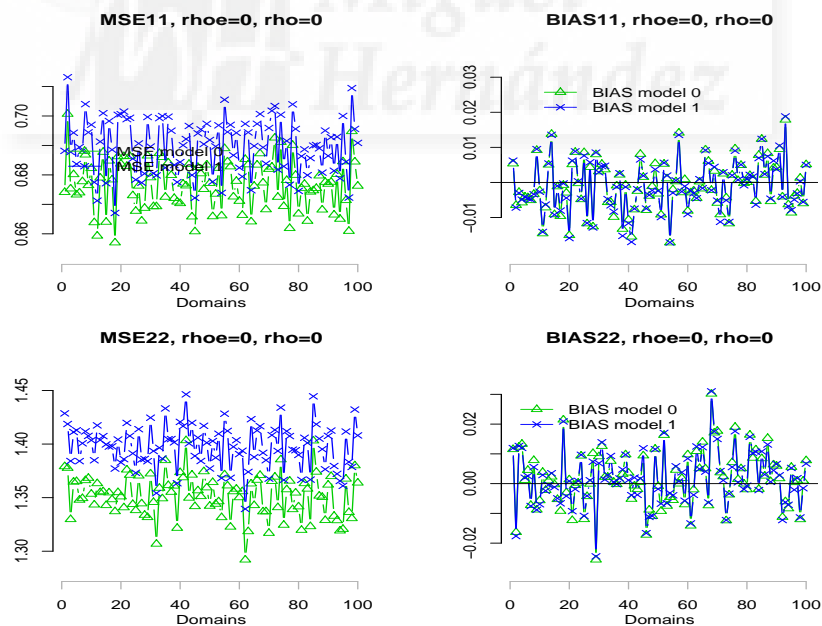


Figura 4.6: E_{drr} , para $a = 0, 1, r = 1, 2, (\rho_e, \rho) = (0, 0), \rho_x = \frac{1}{2}, D = 100$.

5

Modelo AR(1) heterocedástico

5.1. Definición del modelo

Se ha avanzado en el estudio de casos particulares del modelo de área multivariante presentado en el capítulo 2. Para ello se ha estudiado un modelo de área en el capítulo 3 (*modelo diagonal*) en el cual la matriz de varianzas de los efectos es diagonal; es decir, no presenta correlación, y la matriz depende de dos parámetros ($m = 2$) en el caso bivariante. Por otra parte se ha estudiado en el capítulo 4 un modelo de área (*modelo AR(1)*) en el que la matriz de varianzas de los efectos presenta correlación y depende de dos parámetros ($m = 2$).

En este capítulo se estudia un modelo de área en el que la matriz de varianzas de los efectos presenta correlación y depende de $m = R + 1$ parámetros, donde R es el número de variables. En lo que sigue se formaliza detalladamente el modelo.

El *modelo AR(1) heterocedástico* es

$$y = X\beta + Zu + e = X\beta + Z_1u_1 + \cdots + Z_Du_D + e, \quad (5.1)$$

donde e, u_1, \dots, u_D son independientes con distribuciones $e \sim N(0, V_e)$, $u \sim N(0, V_u)$ y $u_d \sim N(0, V_{ud})$, $d = 1, \dots, D$. Se supone que V_e es una matriz conocida y que, para $d = 1, \dots, D$, el vector de efectos aleatorios u_d se distribuye según un proceso estocástico tal que

$$u_{dr} = \rho u_{dr-1} + a_{dr}, \quad u_{d0} \sim N(0, 1), \quad a_{dr} \sim N(0, \sigma_r^2), \quad r = 1, \dots, R,$$

donde las variables aleatorias a_{dr} , $r = 1, \dots, R$, son independientes e independientes de u_{d0} .

El número de parámetros que intervienen en la matriz de varianzas del vector u_d es $m = R + 1$ y se representan por

$$\theta = (\theta_1 = \sigma_1^2, \dots, \theta_R = \sigma_R^2, \theta_{R+1} = \rho).$$

Las componentes de la matriz de varianzas de u_d son

$$\begin{aligned} \text{Var}(u_{dr}) &= E(u_{dr}^2) = E((\rho u_{dr-1} + a_{dr})u_{dr}) = \rho E(u_{dr-1}u_{dr}) + E(a_{dr}u_{dr}) \\ &= \rho E(u_{dr-1}u_{dr}) + E(a_{dr}^2) = \rho \text{Cov}(u_{dr}, u_{dr-1}) + \sigma_r^2, \\ \text{Cov}(u_{dr}, u_{dr-1}) &= E(u_{dr}u_{dr-1}) = E((\rho u_{dr-1} + a_{dr})u_{dr-1}) = \rho E(u_{dr-1}u_{dr-1}) + E(a_{dr}u_{dr-1}) \\ &= \rho E(u_{dr-1}u_{dr-1}) = \rho E((\rho u_{dr-2} + a_{dr-1})u_{dr-1}) = \rho^2 E(u_{dr-2}u_{dr-1}) \\ &+ \rho E(a_{dr-1}u_{dr-1}) = \rho^2 E(u_{dr-2}u_{dr-1}). \end{aligned}$$

Si se repite lo anterior de forma sucesiva se obtiene la siguiente expresión para la covarianza

$$\text{Cov}(u_{dr}, u_{dr-1}) = \rho^l E(u_{dr-1}^2) = \rho^l \text{Var}(u_{dr-1}).$$

Combinando las expresiones de $\text{Var}(u_{dr})$ y $\text{Cov}(u_{dr}, u_{dr-1})$, se obtiene

$$\begin{aligned} \text{Var}(u_{dr}) &= \rho \text{Cov}(u_{dr}, u_{dr-1}) + \sigma_r^2 = \rho^2 \text{Var}(u_{dr-1}) + \sigma_r^2 = \rho^2 (\rho \text{Cov}(u_{dr-1}, u_{dr-2}) + \sigma_{r-1}^2) + \sigma_r^2 \\ &= \rho^3 \text{Cov}(u_{dr-1}, u_{dr-2}) + \rho^2 \sigma_{r-1}^2 + \sigma_r^2 = \rho^4 \text{Var}(u_{dr-2}) + \rho^2 \sigma_{r-1}^2 + \sigma_r^2. \end{aligned}$$

Repitiendo lo anterior de forma sucesiva se obtiene

$$\text{Var}(u_{dr}) = \sum_{k=0}^r \rho^{2k} \sigma_{r-k}^2.$$

Por otra parte, se tiene que

$$\text{Cov}(u_{dr}, u_{dr-1}) = \rho^l \sum_{k=0}^{r-1} \rho^{2k} \sigma_{r-l-k}^2.$$

Para $i \leq j$, los elementos de la matriz de varianzas del vector u_d , v_{udij} , son

$$v_{udij} = \begin{cases} \sum_{k=0}^i \rho^{2k} \sigma_{i-k}^2 & \text{si } i = j, \\ \sum_{k=0}^{j-i} \rho^{2k+j-i} \sigma_{j-i-k}^2 & \text{si } i < j, \end{cases}$$

y para $i > j$ se verifica que $v_{udij} = v_{udji}$.

Las derivadas de la matriz V_{ud} respecto de los parámetros que intervienen en la misma son

$$V_{udr} = \frac{\partial V_{ud}}{\partial \theta_1} = \frac{\partial V_{ud}}{\partial \sigma_r^2}, \quad r = 1, \dots, R, \quad V_{udR+1} = \frac{\partial V_{ud}}{\partial \theta_{R+1}} = \frac{\partial V_{ud}}{\partial \rho}.$$

Para $i \leq j$, los elementos de las matrices anteriores son

$$\frac{\partial v_{udij}}{\partial \theta_r} = \frac{\partial v_{udij}}{\partial \sigma_r^2} = \begin{cases} \rho^{2(i-r)} & \text{si } i = j \geq r, \\ \rho^{3(j-i)-2r} & \text{si } i < j \text{ y } j-i \geq r, \\ 0 & \text{en caso contrario,} \end{cases} \quad r = 1, \dots, R,$$

$$\frac{\partial v_{udij}}{\partial \theta_{R+1}} = \begin{cases} \sum_{k=1}^i 2k \rho^{2k-1} \sigma_{i-k}^2 & \text{si } i = j, \\ \sum_{k=0}^r (2k+j-i) \rho^{2k+j-i-1} \sigma_{r-k}^2 & \text{si } i < j, \end{cases}$$

y para $i > j$ son

$$\frac{\partial v_{udij}}{\partial \theta_r} = \frac{\partial v_{udji}}{\partial \theta_r}, \quad r = 1, \dots, R, \quad \frac{\partial v_{udij}}{\partial \theta_{R+1}} = \frac{\partial v_{udji}}{\partial \theta_{R+1}}.$$

La distribución asintótica del estimador REML de θ es

$$\hat{\theta} \sim N_R(\theta, F^{-1}(\theta)).$$

En el caso bivalente, $R = 2$, se deduce que

$$\hat{\sigma}_1^2 - \hat{\sigma}_2^2 \sim N(\sigma_1^2 - \sigma_2^2, v_{11} + v_{22} - 2v_{12}), \quad (5.2)$$

donde v_{ij} el elemento correspondiente de la matriz $F^{-1}(\theta)$ definida en la sección 2.3. La distribución (5.2) se puede usar para comprobar la igualdad de varianzas del modelo mediante el contraste de la hipótesis $H_0 : \sigma_1^2 = \sigma_2^2$. Si el nivel de significación se fija en α , entonces se rechaza H_0 si

$$\frac{\hat{\sigma}_1^2 - \hat{\sigma}_2^2}{\sqrt{\hat{v}_{11} + \hat{v}_{22} - 2\hat{v}_{12}}} \notin (-z_{\alpha/2}, z_{\alpha/2}).$$

Para comprobar la presencia de correlación en los efectos aleatorios del modelo se puede realizar el contraste $H_0 : \rho = 0$. Para ello, se utiliza la distribución asintótica

$$\hat{\rho} \sim N(\rho, v_{33}), \quad (5.3)$$

donde v_{33} es el elemento correspondiente de la matriz $F^{-1}(\theta)$. La distribución (5.3) se puede usar para comprobar la no nulidad del parámetro de correlación mediante el contraste de la hipótesis $H_0 : \rho = 0$. Si el nivel de significación se fija en α , entonces se tiene que se rechaza H_0 si se verifica

$$\frac{\hat{\rho}}{\sqrt{\hat{v}_{33}}} \notin (-z_{\alpha/2}, z_{\alpha/2}).$$

En el caso general interesa contrastar si las varianzas σ_r^2 son iguales; es decir, $H_0 : \sigma_1^2 = \dots = \sigma_R^2$. Para ello, se utiliza el estadístico T de la razón de verosimilitudes. En primer lugar, se descompone el vector de parámetros en la forma $\theta = (\sigma, \rho)$, donde $\sigma = (\sigma_1^2, \dots, \sigma_R^2)$ y se consideran los conjuntos Θ , Θ_0 y Θ_1 , donde

$$\Theta = \{\theta \in \mathbb{R}^{R+1} : \sigma > 0 \text{ y } \rho \in [-1, 1]\}, \quad \Theta_0 = \{\theta \in \Theta : \sigma_1^2 = \dots = \sigma_R^2\}, \quad \Theta_1 = \{\theta \in \Theta : \theta \notin \Theta_0\}.$$

El estadístico T es

$$T = -2 \ln \frac{\sup_{\theta \in \Theta_0} f(y_1, \dots, y_{DR})}{\sup_{\theta \in \Theta} f(y_1, \dots, y_{DR})} \sim \chi_{\dim \Theta - \dim \Theta_0}^2.$$

El estadístico AIC se puede usar para seleccionar el modelo con un conjunto de variables explicativas más apropiado. El AIC-REML es

$$AIC = -2\ell_{reml}(\hat{\theta}) + 2 \dim(\Theta).$$

Para terminar el apartado, conviene notar que el modelo AR(1) heterocedástico no es una generalización del modelo AR(1) expuesto en el capítulo 4. En cada área d el modelo AR(1) supone que u_{dr} , $r = 1, 2$, es parte de un proceso estocástico AR(1) estacionario que comienza en $r = -\infty$. Sin embargo, en cada área d el modelo AR(1) heterocedástico supone que u_{dr} , $r = 1, 2$, comienza en $r = 0$, donde $u_{d0} \sim N(0, 1)$.

5.2. Experimentos de simulación

Para estudiar empíricamente el comportamiento de los algoritmos de ajuste y de los procedimientos de estimación del error cuadrático medio de los EBLUPs, en esta sección se presentan tres experimentos de simulación. En las simulaciones se compara el modelo AR(1) heterocedástico (5.1) con el modelo con errores e_{dr} independientes. Este último modelo prescinde de toda estructura multivariante y equivale a aplicar por separado R modelos Fay-Herriot; es decir, uno por cada componente r , $r = 1, \dots, R$.

En las simulaciones, se ha programado un modelo AR(1) heterocedástico (5.1) bivalente ($R = 2$) cuyas características se describen a continuación.

La matriz de covarianzas del vector u_d es

$$V_{ud} = \begin{pmatrix} \sigma_1^2 + \rho^2 & \rho\sigma_1^2 + \rho^3 \\ \rho\sigma_1^2 + \rho^3 & \sigma_2^2 + \rho^2\sigma_1^2 + \rho^4 \end{pmatrix}, \quad \sigma_1^2 = 2, \quad \sigma_2^2 = 4, \quad \rho = 1/2.$$

Las componentes del vector e_d verifican $\text{var}(e_{d1}) = 1$, $\text{var}(e_{d2}) = 2$ y $\text{corr}(e_{d1}, e_{d2}) = \rho_e$. Por tanto, la matriz de covarianzas del vector e_d es

$$V_{ed} = \begin{pmatrix} \sigma_{d11} & \sigma_{d12} \\ \sigma_{d21} & \sigma_{d22} \end{pmatrix}, \quad \sigma_{d11} = 1, \quad \sigma_{d22} = 2, \quad \sigma_{d12} = \sigma_{d21} = \rho_e \sqrt{\sigma_{d11}\sigma_{d22}}.$$

La matriz de covarianzas del vector y_d es

$$\begin{aligned} V_d &= V_{ud} + V_{ed} = \begin{pmatrix} \sigma_1^2 + \rho^2 & \rho\sigma_1^2 + \rho^3 \\ \rho\sigma_1^2 + \rho^3 & \sigma_2^2 + \rho^2\sigma_1^2 + \rho^4 \end{pmatrix} + \begin{pmatrix} \sigma_{d11} & \sigma_{d12} \\ \sigma_{d21} & \sigma_{d22} \end{pmatrix} \\ &= \begin{pmatrix} \frac{9}{4} & \frac{9}{8} \\ \frac{9}{8} & \frac{73}{16} \end{pmatrix} + \begin{pmatrix} 1 & \rho_e \sqrt{2} \\ \rho_e \sqrt{2} & 2 \end{pmatrix} = \begin{pmatrix} \frac{13}{4} & \frac{9}{8} + \rho_e \sqrt{2} \\ \frac{9}{8} + \rho_e \sqrt{2} & \frac{105}{16} \end{pmatrix}, \quad d = 1, \dots, D. \end{aligned}$$

La matriz de covarianzas del vector y es

$$V = \text{diag}_{1 \leq d \leq D} \begin{pmatrix} \frac{13}{4} & \frac{9}{8} + \rho_e \sqrt{2} \\ \frac{9}{8} + \rho_e \sqrt{2} & \frac{105}{16} \end{pmatrix}.$$

Las derivadas parciales que se utilizan en el algoritmo de Fisher-scoring son

$$V_1 = \text{diag}_{1 \leq d \leq D} V_{d1}, \quad V_2 = \text{diag}_{1 \leq d \leq D} V_{d2}, \quad V_3 = \text{diag}_{1 \leq d \leq D} V_{d3},$$

donde

$$\begin{aligned} V_{d1} &= \frac{\partial V_{ud}}{\partial \theta_1} = \frac{\partial V_{ud}}{\partial \sigma_1^2} = \begin{pmatrix} 1 & \rho \\ \rho & \rho^2 \end{pmatrix}, \\ V_{d2} &= \frac{\partial V_{ud}}{\partial \theta_2} = \frac{\partial V_{ud}}{\partial \sigma_2^2} = \begin{pmatrix} 0 & 0 \\ 0 & 1 \end{pmatrix}, \\ V_{d3} &= \frac{\partial V_{ud}}{\partial \theta_3} = \frac{\partial V_{ud}}{\partial \rho} = \begin{pmatrix} 2\rho & \sigma_1^2 + 3\rho^2 \\ \sigma_1^2 + 3\rho^2 & 2\rho\sigma_1^2 + 4\rho^3 \end{pmatrix}. \end{aligned}$$

Teniendo en cuenta los valores especificados de los parámetros, se obtiene

$$V_{d1} = \frac{\partial V_{ud}}{\partial \sigma_1^2} = \begin{pmatrix} 1 & \frac{1}{2} \\ \frac{1}{2} & \frac{1}{4} \end{pmatrix}, \quad V_{d2} = \frac{\partial V_{ud}}{\partial \sigma_2^2} = \begin{pmatrix} 0 & 0 \\ 0 & 1 \end{pmatrix}, \quad V_{d3} = \frac{\partial V_{ud}}{\partial \rho} = \begin{pmatrix} 1 & \frac{11}{4} \\ \frac{11}{4} & \frac{5}{2} \end{pmatrix}.$$

Los valores de los parámetros de regresión son $\beta_1 = 1$, $\beta_2 = 1$. Para estimar los parámetros σ_1^2 , σ_2^2 y ρ mediante el algoritmo Fisher-scoring, se utilizan como valores iniciales (semillas) los verdaderos valores; es decir, $\sigma_1^2 = 2$ y $\sigma_2^2 = 4$ y $\rho = \frac{1}{2}$.

Como se van a realizar comparaciones entre el modelo multivariante y los modelos univariantes marginales, a continuación se da la descripción de tales modelos así como los parámetros que intervienen en los mismos. Los modelos univariantes son

$$y_{dr} = x_{dr}\beta_r + u_{dr} + e_{dr}, \quad d = 1, \dots, D, \quad r = 1, 2,$$

donde $u_{d1} \sim N(0, \sigma_1^2 + \rho^2)$, $u_{d2} \sim N(0, \sigma_2^2 + \rho^2\sigma_1^2 + \rho^4)$ y $e_{dr} \sim N(0, \sigma_{dr}^2)$, $r = 1, 2$, son independientes.

Los parámetros de los modelos univariantes son los mismos que se usan en el modelo multivariante diagonal; es decir $\beta_1 = 1$, $\beta_2 = 1$, $\sigma_1^2 = 2$, $\sigma_2^2 = 4$, $\rho = 0,5$, $\sigma_{d11} = 1$ y $\sigma_{d22} = 2$.

Para ambos modelos se utilizan las variables explicativas

$$x_{d1} = \mu_1 + \sigma_{x11}^{1/2}U_{d1}, \quad x_{d2} = \mu_2 + \sigma_{x22}^{1/2}(\rho_x U_{d1} + (1 - \rho_x^2)^{1/2}U_{d2}), \quad d = 1, \dots, D,$$

donde $\mu_1 = \mu_2 = 10$, $\sigma_{x11} = 1$, $\sigma_{x22} = 2$, $\rho_x = 0,5$ y

$$U_{dr} = \frac{d-D}{D} + \frac{r}{3}, \quad r = 1, 2, \quad d = 1, \dots, D.$$

5.2.1. Experimento de simulación 1

El objetivo de este experimento es investigar empíricamente la pérdida de eficiencia en las estimaciones cuando no se tiene en cuenta la naturaleza multivariante de los datos. Nos interesa estudiar lo que ocurre cuando se asume incorrectamente los modelos marginales independientes, en lugar del modelo multivariante subyacente. Para ello se simulan los datos o bien del modelo multivariante o bien del modelo con $\rho_e = \rho = 0$, se estiman los parámetros y se calculan los EBLUP de ambos modelos. Hay dos conjuntos de parámetros estimados y EBLUP calculados. Según sea el caso, un conjunto se obtiene bajo el modelo correcto y otro bajo el modelo incorrecto. Cabe esperar que los mejores resultados se obtengan siempre cuando se usan los estimadores correspondientes al modelo correcto. El experimento consiste en calcular por simulación Monte Carlo los sesgos y errores cuadráticos medios empíricos (BIAS y MSE) de los estimadores de los parámetros y de los EBLUP.

Los datos se simulan del modelo multivariante (5.1). Se consideran los casos: (1) $\rho_e = 0, \rho = 0$, (2) $\rho_e = 1/2, \rho = 0$, (3) $\rho_e = 0, \rho = 1/2$ y (4) $\rho_e = 1/2, \rho = 1/2$. Los pasos de la simulación son

1. Repetir $I = 10^4$ veces ($i = 1, \dots, I$)

1.1. Generar una muestra (y_{dr}, x_{dr}) , $d = 1, \dots, D$, $r = 1, 2$.

1.2. Calcular $\{\hat{\beta}_{E1}^{(i,0)}, \hat{\beta}_{E2}^{(i,0)}, \hat{\sigma}_{u1}^{2(i,0)}, \hat{\sigma}_{u2}^{2(i,0)}\}$, $\{\hat{\beta}_{E1}^{(i,1)}, \hat{\beta}_{E2}^{(i,1)}, \hat{\sigma}_{u1}^{2(i,1)}, \hat{\sigma}_{u2}^{2(i,1)}, \hat{\rho}^{(i,1)}\}$ y $\{\hat{\mu}_d^{(i,0)}, \hat{\mu}_d^{(i,1)}\}$, donde los superíndices “0” y “1” se usan para denotar los estimadores y EBLUPs calculados asumiendo el modelo producto de marginales ($a = 0$) o el modelo multivariante AR(1) heterocedástico ($a = 1$).

2. Salida: Para todo $\tau \in \{\beta_1^{(0)}, \beta_2^{(0)}, \sigma_{u1}^{2(0)}, \sigma_{u2}^{2(0)}, \beta_1^{(1)}, \beta_2^{(1)}, \sigma_{u1}^{2(1)}, \sigma_{u2}^{2(1)}, \rho^{(1)}\}$, calcular

$$MSE(\hat{\tau}) = \frac{1}{I} \sum_{i=1}^I (\hat{\tau}^{(i)} - \tau)^2, \quad BIAS(\hat{\tau}) = \frac{1}{I} \sum_{i=1}^I (\hat{\tau}^{(i)} - \tau),$$

$$MSE_{rrd}^{(a)} = \frac{1}{I} \sum_{i=1}^I (\hat{\mu}_{rd}^{(i,a)} - \mu_{rd}^{(i,a)})^2, \quad BIAS_{rrd}^{(a)} = \frac{1}{I} \sum_{i=1}^I (\hat{\mu}_{rd}^{(i,a)} - \mu_{rd}^{(i,a)}), \quad d = 1, D/2, D, r = 1, 2, a = 0, 1.$$

La tabla 5.1 presenta los errores cuadráticos medios y los sesgos empíricos de los estimadores REML de los parámetros de los dos modelos considerados. La tabla está ordenada por columnas y filas. La primera columna contiene el caso que se está considerando (modelo que genera los datos), la segunda columna especifica el estimador del parámetro del modelo, las cuatro columnas siguientes muestran el error cuadrático medio y las cuatro últimas el sesgo. Cada uno de los dos grupos de cuatro columnas que acabamos de nombrar se corresponde con un valor distinto del número de áreas; es decir, $D = 50, 100, 200, 400$. Las filas se disponen en grupos de nueve, un grupo para cada caso. En el primer caso, los datos se simulan del modelo con $\rho = \rho_e = 0$, mientras que en los casos 2, 3 y 4 los datos se generan del modelo multivariante con $\rho > 0$ o $\rho_e > 0$. Dentro de cada caso, las cuatro primeras filas se corresponden con las estimaciones para el modelo producto de univariantes ($a = 0$) y las cinco siguientes con las estimaciones para el modelo multivariante ($a = 1$).

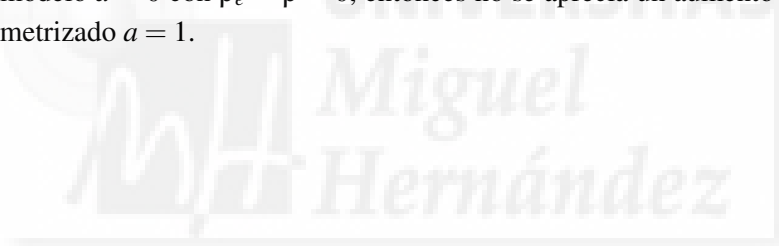
La tabla 5.1 muestra que los errores cuadráticos medios y sesgos de $\hat{\beta}_{E1}$ y $\hat{\beta}_{E2}$ son básicamente iguales para las estimaciones que se hacen asumiendo los modelos $a = 0$ o $a = 1$, independientemente del caso que se está considerando en la simulación de los datos. El error cuadrático medio de $\hat{\sigma}_{u2}^{2(1)}$ es menor que el de $\hat{\sigma}_{u2}^{2(0)}$ en los casos 3 y 4; es decir, cuando $\rho = \frac{1}{2}$. Los errores cuadráticos medios de $\hat{\sigma}_{u1}^{2(1)}$ y de $\hat{\sigma}_{u1}^{2(0)}$ tienden a ser muy similares a medida que aumenta el valor de D en los casos 1 y 2, es decir, cuando $\rho = 0$. En los dos casos restantes se observa un error cuadrático medio mayor para el modelo $a = 1$. En los cuatro casos considerados se observa siempre un error cuadrático medio para $\hat{\rho}^{(1)}$ inferior comparado con el resto de parámetros que intervienen en la varianza de los efectos. Los sesgos tienden a ser mayores en el modelo $a = 1$ en todos los casos considerados.

Las tablas 5.2 y 5.3 presentan los errores cuadráticos medios y los sesgos, $MSE_{drr}^{(a)}$ y $BIAS_{drr}^{(a)}$, $a = 0, 1$, de los EBLUPs de las medias de las componentes $r = 1$ y $r = 2$ respectivamente. En ambas tablas, las columnas se disponen de forma análoga a la tabla 5.1. La diferencia está en la segunda columna donde aparecen los valores de las áreas consideradas, $d = 1$, $d = D/2$ y $d = D$, y el valor medio a lo largo de ellas. El resto de columnas está estructurado de la misma forma que en la tabla 5.1, pero el error cuadrático medio y el sesgo son de los estimadores $\hat{\mu}_{d1}$ en la tabla 5.2 y $\hat{\mu}_{d2}$ en la tabla 5.3.

En las tablas 5.2 y 5.3 se observa en el primer caso ($\rho_e = 0, \rho = 0$) que los errores cuadráticos medios de los estimadores de $\hat{\mu}_{d1}$ y $\hat{\mu}_{d2}$ apenas se diferencian en ambos modelos; en los casos 2 y 3 la diferencia que se aprecia es considerable, y, en el caso 4, no se aprecia una diferencia notable debido, en parte, a la presencia de una excesiva correlación.

La figura 5.1 muestra las gráficas los valores de $MSE_{drr}^{(a)}$, $a = 0, 1$, $r = 1, 2$, $d = 1, \dots, D$, $D = 100$. La figura está dividida en 4 partes y tiene una disposición en forma de tabla con dos filas y dos columnas. Las filas 1 y 2 presentan los valores de $MSE_{drr}^{(a)}$ para $r = 1$ y $r = 2$ respectivamente. Las columnas 1 y 2 presentan los valores de $MSE_{drr}^{(a)}$ cuando los datos se generan del modelo teniendo en cuenta los casos primero ($\rho_e = 0, \rho = 0$) y tercero ($\rho_e = 0, \rho = 1/2$) respectivamente. Cada una de las cuatro sub-figuras muestran los valores de $MSE_{drr}^{(a)}$ para $a = 0$ y $a = 1$. En la figura 5.1 se observa que la diferencia de los errores cuadráticos medios entre el modelo producto de univariantes ($a = 0$) y el modelo multivariante ($a = 1$) cuando ρ cambia de 0 a $\frac{1}{2}$ es bastante pronunciada. También hay que destacar que si los datos se generan del modelo $a = 0$ con $\rho_e = \rho = 0$, entonces no se aprecia un aumento del error cuadrático medio al utilizar el modelo sobre-parametrizado $a = 1$.

La figura 5.2 muestra las gráficas los valores de $BIAS_{drr}^{(a)}$, $a = 0, 1$, $r = 1, 2$, $d = 1, \dots, D$, $D = 100$. La figura está estructurada de la misma forma que la figura 5.1. En la figura 5.2 se observa una leve diferencia en los sesgos de los modelos $a = 0$ y $a = 1$ cuando ρ cambia de 0 a $1/2$. También hay que destacar que si los datos se generan del modelo $a = 0$ con $\rho_e = \rho = 0$, entonces no se aprecia un aumento de sesgo al utilizar el modelo sobre-parametrizado $a = 1$.



(ρ_e, ρ)	D	MSE				BIAS			
		50	100	200	400	50	100	200	400
(0, 0)	$\hat{\beta}_{E1}^{(0)}$	0.0006	0.0003	0.0002	0.0001	0.0000	-0.0002	0.0000	0.0000
	$\hat{\beta}_{E2}^{(0)}$	0.0011	0.0006	0.0003	0.0001	0.0001	0.0002	0.0002	0.0000
	$\hat{\sigma}_{u1}^{2(0)}$	0.3697	0.1864	0.0918	0.0446	0.0004	-0.0044	-0.0008	0.0016
	$\hat{\sigma}_{u2}^{2(0)}$	1.4310	0.7129	0.3670	0.1801	-0.0160	-0.0120	0.0020	-0.0085
	$\hat{\beta}_{E1}^{(1)}$	0.0006	0.0003	0.0002	0.0001	0.0000	-0.0002	0.0000	0.0000
	$\hat{\beta}_{E2}^{(1)}$	0.0011	0.0006	0.0003	0.0001	0.0001	0.0002	0.0002	0.0000
	$\hat{\sigma}_{u1}^{2(1)}$	0.4936	0.2093	0.0969	0.0457	-0.1123	-0.0551	-0.0247	-0.0097
	$\hat{\sigma}_{u2}^{2(1)}$	1.4545	0.7201	0.3670	0.1806	-0.2063	-0.1061	-0.0443	-0.0306
$\hat{\rho}^{(1)}$	0.1127	0.0506	0.0239	0.0112	-0.0011	0.0032	-0.0031	0.0018	
$(\frac{1}{2}, 0)$	$\hat{\beta}_{E1}^{(0)}$	0.0006	0.0003	0.0002	0.0001	-0.0003	0.0004	0.0002	0.0000
	$\hat{\beta}_{E2}^{(0)}$	0.0011	0.0006	0.0003	0.0001	0.0001	0.0008	0.0000	0.0001
	$\hat{\sigma}_{u1}^{2(0)}$	0.3649	0.1766	0.0882	0.0459	0.0003	-0.0063	0.0024	0.0005
	$\hat{\sigma}_{u2}^{2(0)}$	1.4261	0.7408	0.3679	0.1801	-0.0122	-0.0032	0.0054	-0.0022
	$\hat{\beta}_{E1}^{(1)}$	0.0006	0.0003	0.0002	0.0001	-0.0003	0.0004	0.0002	0.0000
	$\hat{\beta}_{E2}^{(1)}$	0.0011	0.0006	0.0003	0.0001	0.0001	0.0008	0.0000	0.0001
	$\hat{\sigma}_{u1}^{2(1)}$	0.5070	0.1991	0.0933	0.0472	-0.1189	-0.0571	-0.0223	-0.0114
	$\hat{\sigma}_{u2}^{2(1)}$	1.4682	0.7464	0.3697	0.1804	-0.2117	-0.0977	-0.0425	-0.0257
$\hat{\rho}^{(1)}$	0.1191	0.0508	0.0247	0.0119	-0.0246	-0.0076	-0.0045	-0.0009	
$(0, \frac{1}{2})$	$\hat{\beta}_{E1}^{(0)}$	0.0007	0.0003	0.0002	0.0001	0.0000	0.0000	0.0001	0.0000
	$\hat{\beta}_{E2}^{(0)}$	0.0013	0.0006	0.0003	0.0002	-0.0005	-0.0001	0.0000	0.0000
	$\hat{\sigma}_{u1}^{2(0)}$	0.4328	0.2095	0.1071	0.0518	0.0041	0.0023	0.0033	0.0027
	$\hat{\sigma}_{u2}^{2(0)}$	1.7580	0.8938	0.4307	0.2161	0.0165	0.0116	-0.0007	-0.0016
	$\hat{\beta}_{E1}^{(1)}$	0.0007	0.0003	0.0002	0.0001	0.0000	0.0000	0.0001	0.0000
	$\hat{\beta}_{E2}^{(1)}$	0.0013	0.0006	0.0003	0.0002	-0.0005	-0.0001	0.0000	0.0000
	$\hat{\sigma}_{u1}^{2(1)}$	0.7105	0.3173	0.1516	0.0705	-0.1056	-0.0526	-0.0228	-0.0090
	$\hat{\sigma}_{u2}^{2(1)}$	1.5902	0.8104	0.4022	0.1964	-0.1715	-0.0878	-0.0499	-0.0248
$\hat{\rho}^{(1)}$	0.0984	0.0447	0.0212	0.0103	0.0113	0.0102	0.0049	0.0014	
$(\frac{1}{2}, \frac{1}{2})$	$\hat{\beta}_{E1}^{(0)}$	0.0007	0.0003	0.0002	0.0001	-0.0001	0.0001	0.0000	0.0001
	$\hat{\beta}_{E2}^{(0)}$	0.0012	0.0006	0.0003	0.0002	-0.0001	-0.0002	-0.0001	0.0000
	$\hat{\sigma}_{u1}^{2(0)}$	0.4312	0.2149	0.1067	0.0535	0.0070	0.0026	0.0003	0.0016
	$\hat{\sigma}_{u2}^{2(0)}$	1.7445	0.8540	0.4212	0.2105	0.0000	-0.0095	0.0046	-0.0006
	$\hat{\beta}_{E1}^{(1)}$	0.0007	0.0003	0.0002	0.0001	-0.0001	0.0001	0.0000	0.0001
	$\hat{\beta}_{E2}^{(1)}$	0.0012	0.0006	0.0003	0.0002	-0.0001	-0.0002	-0.0001	0.0000
	$\hat{\sigma}_{u1}^{2(1)}$	0.5570	0.2581	0.1224	0.0604	-0.0744	-0.0331	-0.0159	-0.0058
	$\hat{\sigma}_{u2}^{2(1)}$	1.2663	0.6059	0.3033	0.1533	-0.1742	-0.0903	-0.0336	-0.0188
$\hat{\rho}^{(1)}$	0.0858	0.0392	0.0182	0.0089	-0.0044	-0.0034	-0.0020	-0.0015	

Tabla 5.1: MSE (izquierda) y BIAS (derecha) para $\rho_x = 1/2$.

			MSE				BIAS			
(ρ_e, ρ)	a	d	50	100	200	400	50	100	200	400
$(0, 0)$	0	1	0.6845	0.6757	0.6718	0.6584	0.0039	0.0076	-0.0003	0.0065
		$D/2$	0.6820	0.6678	0.6704	0.6728	0.0104	-0.0049	0.0047	-0.0039
		D	0.6943	0.6793	0.6812	0.6740	-0.0015	0.0008	0.0050	0.0024
		mean	0.6878	0.6780	0.6718	0.6694	-0.0008	-0.0008	0.0004	-0.0002
	1	1	0.6926	0.6796	0.6741	0.6588	0.0035	0.0078	-0.0001	0.0066
		$D/2$	0.6871	0.6705	0.6718	0.6732	0.0098	-0.0052	0.0052	-0.0039
		D	0.7036	0.6824	0.6831	0.6744	-0.0024	0.0007	0.0048	0.0019
		mean	0.6953	0.6816	0.6734	0.6703	-0.0008	-0.0008	0.0004	-0.0002
$(\frac{1}{2}, 0)$	0	1	0.6751	0.6689	0.6824	0.6662	-0.0019	0.0112	0.0058	-0.0059
		$D/2$	0.6878	0.6805	0.6812	0.6725	-0.0031	-0.0093	-0.0028	0.0018
		D	0.6956	0.6970	0.6710	0.6836	0.0114	0.0127	-0.0052	0.0108
		mean	0.6871	0.6766	0.6721	0.6696	-0.0007	0.0008	0.0003	-0.0004
	1	1	0.6449	0.6374	0.6451	0.6289	-0.0005	0.0129	0.0031	-0.0022
		$D/2$	0.6638	0.6455	0.6421	0.6350	-0.0021	-0.0086	-0.0050	0.0044
		D	0.6723	0.6629	0.6336	0.6454	0.0108	0.0092	-0.0050	0.0066
		mean	0.6584	0.6439	0.6367	0.6324	-0.0007	0.0008	0.0003	-0.0004
$(0, \frac{1}{2})$	0	1	0.7016	0.6963	0.6964	0.6991	-0.0148	-0.0035	0.0130	-0.0251
		$D/2$	0.7078	0.6915	0.6925	0.6816	0.0102	-0.0059	-0.0074	0.0010
		D	0.6956	0.6974	0.6963	0.6966	0.0145	0.0062	-0.0023	0.0055
		mean	0.7119	0.7021	0.6969	0.6951	0.0011	0.0006	0.0011	-0.0002
	1	1	0.6890	0.6804	0.6809	0.6819	-0.0138	-0.0024	0.0106	-0.0242
		$D/2$	0.6970	0.6773	0.6753	0.6618	0.0069	-0.0044	-0.0083	0.0013
		D	0.6875	0.6811	0.6752	0.6758	0.0134	0.0051	-0.0015	0.0066
		mean	0.7004	0.6864	0.6793	0.6765	0.0011	0.0006	0.0011	-0.0002
$(\frac{1}{2}, \frac{1}{2})$	0	1	0.7005	0.6948	0.7066	0.6959	0.0019	0.0019	0.0153	0.0186
		$D/2$	0.7042	0.6912	0.6942	0.6933	0.0071	0.0025	-0.0122	0.0030
		D	0.7082	0.7138	0.6982	0.6955	0.0018	0.0031	0.0039	0.0031
		mean	0.7104	0.7020	0.6974	0.6950	0.0005	0.0001	0.0002	-0.0003
	1	1	0.7064	0.6944	0.7025	0.6935	0.0013	0.0021	0.0165	0.0184
		$D/2$	0.7082	0.6940	0.6923	0.6904	0.0054	0.0019	-0.0121	0.0032
		D	0.7132	0.7119	0.6969	0.6911	0.0026	0.0046	0.0034	0.0031
		mean	0.7139	0.7019	0.6952	0.6920	0.0005	0.0001	0.0002	-0.0003

Tabla 5.2: $MSE_{11d}^{(a)}$ (izquierda) y $BIAS_{11d}^{(a)}$ (derecha) para $\rho_x = 1/2$, $a = 0, 1$.

(ρ_e, ρ)	a	d	MSE				BIAS			
			50	100	200	400	50	100	200	400
(0, 0)	0	1	1.3605	1.3442	1.3557	1.3296	0.0040	-0.0006	0.0026	-0.0080
		$D/2$	1.3701	1.3598	1.3115	1.3850	0.0115	-0.0096	0.0089	-0.0103
		D	1.3620	1.3588	1.3297	1.3291	0.0129	0.0129	0.0009	0.0069
		mean	1.3736	1.3544	1.3422	1.3381	0.0004	-0.0004	0.0011	0.0005
	1	1	1.3759	1.3514	1.3603	1.3310	0.0051	-0.0020	0.0028	-0.0079
		$D/2$	1.3822	1.3639	1.3140	1.3860	0.0137	-0.0088	0.0092	-0.0101
		D	1.3804	1.3690	1.3345	1.3323	0.0133	0.0123	0.0012	0.0073
		mean	1.3881	1.3616	1.3455	1.3397	0.0004	-0.0004	0.0011	0.0005
$(\frac{1}{2}, 0)$	0	1	1.3351	1.3489	1.3608	1.3280	-0.0205	0.0210	-0.0038	0.0005
		$D/2$	1.3919	1.3223	1.3090	1.3395	-0.0034	0.0020	-0.0039	-0.0006
		D	1.3994	1.3618	1.3698	1.3828	0.0129	0.0112	-0.0089	0.0093
		mean	1.3740	1.3509	1.3449	1.3371	-0.0004	0.0019	0.0000	-0.0001
	1	1	1.2886	1.2734	1.2854	1.2483	-0.0224	0.0236	-0.0064	0.0023
		$D/2$	1.3526	1.2540	1.2485	1.2629	-0.0052	0.0006	-0.0059	0.0013
		D	1.3384	1.2874	1.2915	1.3068	0.0193	0.0074	-0.0066	0.0087
		mean	1.3186	1.2861	1.2737	1.2629	-0.0004	0.0019	0.0000	-0.0001
$(0, \frac{1}{2})$	0	1	1.4241	1.3834	1.3997	1.3888	-0.0083	-0.0083	-0.0190	0.0100
		$D/2$	1.4550	1.4172	1.4038	1.4158	0.0173	-0.0053	-0.0032	-0.0218
		D	1.4378	1.4229	1.4193	1.3783	0.0232	0.0162	-0.0056	-0.0026
		mean	1.4310	1.4083	1.4010	1.3949	-0.0019	0.0005	-0.0007	0.0002
	1	1	1.3999	1.3527	1.3637	1.3509	-0.0107	-0.0118	-0.0168	0.0079
		$D/2$	1.4357	1.3784	1.3726	1.3826	0.0177	-0.0071	-0.0014	-0.0203
		D	1.4161	1.3903	1.3854	1.3427	0.0250	0.0162	-0.0043	-0.0031
		mean	1.4084	1.3778	1.3670	1.3583	-0.0019	0.0004	-0.0007	0.0002
$(\frac{1}{2}, \frac{1}{2})$	0	1	1.4528	1.4151	1.3626	1.4175	-0.0066	0.0086	0.0126	0.0035
		$D/2$	1.4075	1.4286	1.4028	1.4026	0.0009	0.0046	-0.0142	-0.0059
		D	1.4245	1.4091	1.4127	1.4102	0.0045	-0.0080	0.0170	0.0106
		mean	1.4284	1.4038	1.3996	1.3953	0.0004	-0.0016	-0.0006	0.0001
	1	1	1.4587	1.4177	1.3558	1.4076	-0.0093	0.0080	0.0122	0.0028
		$D/2$	1.4195	1.4288	1.3988	1.3975	0.0016	0.0057	-0.0154	-0.0043
		D	1.4236	1.4064	1.4051	1.4027	0.0029	-0.0076	0.0180	0.0112
		mean	1.4344	1.4026	1.3946	1.3887	0.0004	-0.0016	-0.0006	0.0001

Tabla 5.3: $MSE_{22d}^{(a)}$ (izquierda) y $BIAS_{22d}^{(a)}$ (derecha) para $\rho_x = 1/2$, $a = 0, 1$.

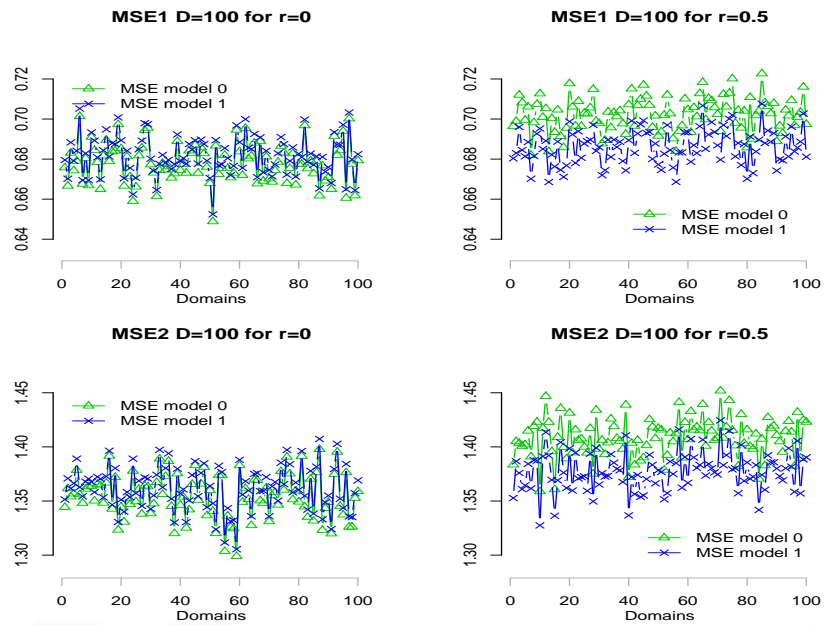


Figura 5.1: MSE_{drr} , para $a = 0, 1$, $r = 1, 2$, $\rho_x = 1/2$, $\rho_e = 0$, $D = 100$.

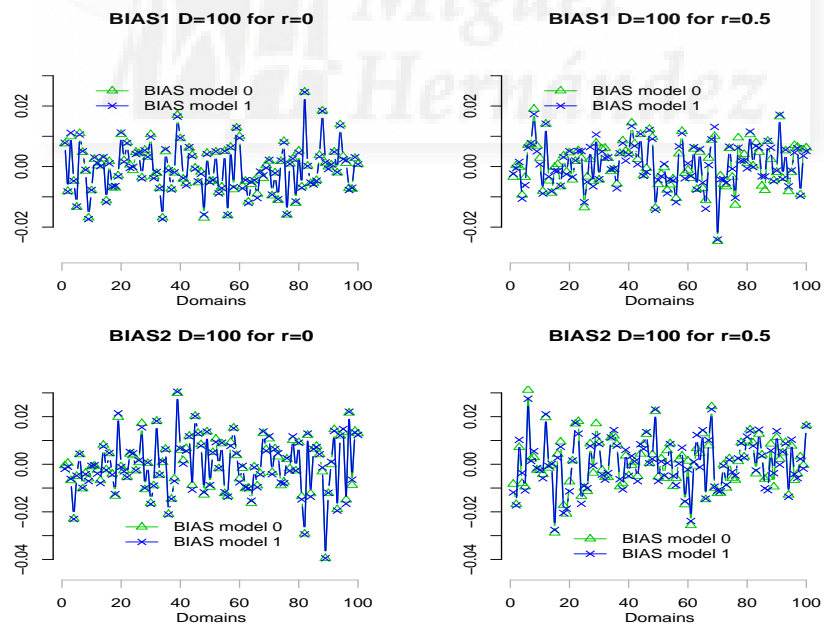


Figura 5.2: $BIAS_{drr}$, para $a = 0, 1$, $r = 1, 2$, $\rho_x = 1/2$, $\rho_e = 0$, $D = 100$.

5.2.2. Experimento de simulación 2

El objetivo de este experimento es investigar empíricamente la pérdida de eficiencia en las estimaciones cuando no se tiene en cuenta la naturaleza multivariante de los datos. Para ello se simulan los datos o bien del modelo multivariante ($a = 1$) o bien del modelo producto de marginales ($a = 0$), se estiman los parámetros y se calculan los EBLUP de ambos modelos. El experimento consiste en calcular por simulación Monte Carlo los sesgos y errores cuadráticos medios empíricos (B y E) de los estimadores analíticos (2.8) en la estimación de los errores cuadráticos medios del EBLUP de μ_{dr} .

En esta simulación se consideran los casos: (1) $\rho_e = 0, \rho = 0$, (2) $\rho_e = 1/2, \rho = 0$, (3) $\rho_e = 0, \rho = 1/2$ y (4) $\rho_e = 1/2, \rho = 1/2$. En el caso 1 los datos se simulan del modelo producto de modelos marginales. Los valores de los parámetros son los mismos que en la simulación 1; es decir, $\beta_1 = 1, \beta_2 = 1, \sigma_{u1}^2 = 2$ y $\sigma_{u1}^2 = 4, \sigma_{d11} = 1, \sigma_{d22} = 2$. Las variables auxiliares x_{dr} también se generan de la misma forma que en la simulación 1.

Los pasos del experimento de simulación son

1. Repetir $I = 500$ veces ($i = 1, \dots, 500$)

- 1.1. Generar una muestra $(y_{dr}^{(i)}, x_{dr}^{(i)})$, $d = 1, \dots, D$, $r = 1, 2$.
- 1.2. Calcular $\{\hat{\beta}_{E1}^{(i,0)}, \hat{\beta}_{E2}^{(i,0)}, \hat{\sigma}_{u1}^{2(i,0)}, \hat{\sigma}_{u2}^{2(i,0)}\}$, $\{\hat{\beta}_{E1}^{(i,1)}, \hat{\beta}_{E2}^{(i,1)}, \hat{\sigma}_{u1}^{2(i,1)}, \hat{\sigma}_{u2}^{2(i,1)}, \hat{\rho}^{(i,1)}\}$.
- 1.3. Para $d = 1, \dots, D$, $a = 0, 1$, $r = 1, 2$, calcular

$$mse_{drr}^{(i,a)} = g_{1drr}^{(i,a)}(\hat{\theta}^{(i,a)}) + g_{2drr}^{(i,a)}(\hat{\theta}^{(i,a)}) + 2g_{3drr}^{(i,a)}(\hat{\theta}^{(i,a)}),$$

$$\text{donde } \hat{\theta}^{(i,0)} = (\hat{\sigma}_{u1}^{2(i,0)}, \hat{\sigma}_{u2}^{2(i,0)}) \text{ y } \hat{\theta}^{(i,1)} = (\hat{\sigma}_{u1}^{2(i,1)}, \hat{\sigma}_{u2}^{2(i,1)}, \hat{\rho}^{(i,1)}).$$

2. Leer los valores $MSE_{drr}^{(a)}$ obtenidos en la simulación 1.
3. Salida:

$$B_{drr}^{(a)} = \frac{1}{I} \sum_{i=1}^I (mse_{drr}^{(i,a)} - MSE_{drr}^{(a)}), \quad E_{drr}^{(a)} = \frac{1}{I} \sum_{i=1}^I (mse_{drr}^{(i,a)} - MSE_{drr}^{(a)})^2, \quad r = 1, 2, a = 0, 1.$$

Las tablas 5.4 y 5.5 presentan los errores cuadráticos medios y los sesgos, $E_{drr}^{(a)}$ y $B_{drr}^{(a)}$, $a = 0, 1$, de los estimadores del error cuadrático medio de los EBLUP de las medias de las componentes $r = 1$ y $r = 2$ respectivamente. La tabla está ordenada por columnas y filas. La primera columna contiene cada uno de los casos que se están considerando, la segunda columna especifica el modelo producto de univariantes ($a = 0$) o multivariante ($a = 1$) bajo el cual se calcula el EBLUP, la tercera columna señala el área $d = 1$, $d = D/2$, $d = D$ y el valor medio de todas las áreas, las cuatro columnas siguientes muestran el error cuadrático medio $E_{drr}^{(a)}$ y las cuatro últimas el sesgo $B_{drr}^{(a)}$. Cada uno de los dos grupos de cuatro columnas que acabamos de nombrar se corresponde con un valor distinto del número de áreas; es decir, $D = 50, 100, 200, 400$. Las filas se disponen en grupos de ocho, un grupo para cada uno de los casos que se están considerando. En el primer caso, los datos se simulan del modelo con $\rho = \rho_e = 0$, mientras que en los casos 2, 3 y 4 los datos se generan

del modelo multivariante con $\rho > 0$ o $\rho_e > 0$. Dentro de cada grupo, las cuatro primeras filas se corresponden con las estimaciones para el modelo $a = 0$ y las cuatro últimas con las estimaciones para el modelo $a = 1$.

En las tablas 5.4 y 5.5 se observa, en el primer caso ($\rho_e = 0, \rho = 0$), unos errores cuadráticos medios para el modelo multivariante ($a = 1$) similares a los del modelo producto de marginales ($a = 0$), como cabía esperar, ya que ambos modelos son el mismo. En los casos segundo ($\rho_e = 1/2, \rho = 0$), tercero ($\rho_e = 0, \rho = 1/2$) y cuarto ($\rho_e = 1/2, \rho = 1/2$) se observa que los errores cuadráticos medios del modelo multivariante ($a = 1$) son ligeramente mayores que los del modelo producto de marginales ($a = 0$) siendo la diferencia entre ambos menor conforme aumenta el número de áreas consideradas. Debido a la presencia de tres parámetros desconocidos en la matriz de varianzas de los efectos ($\sigma_{u1}^2, \sigma_{u2}^2$ y ρ) y a la estructura funcional de la misma, que es más compleja que en el *modelo diagonal* y el *modelo AR(1)*, se pierde eficiencia como consecuencia directa en la estimación del término G_3 descrito en la fórmula (2.8). Por todo ello, se concluye que las ventajas obtenidas de aprovechar la naturaleza multivariante de los datos se pueden perder debido a las aproximaciones del término G_3 . En las tablas 5.4 y 5.5 se aprecia una mejor distribución de los sesgos alrededor del cero en el modelo multivariante ($a = 1$) a medida que aumenta el número de áreas consideradas.

La figura 5.3 muestra las gráficas los valores de $E_{drr}^{(a)}$, $a = 0, 1$, $r = 1, 2$, $d = 1, \dots, D$, $D = 100$. La figura está dividida en cuatro partes y tiene una disposición en forma de tabla con dos filas y dos columnas. La primera y segunda fila presentan los valores de $E_{drr}^{(a)}$ para $r = 1$ y $r = 2$ respectivamente. Las primera y segunda columna presentan los valores de $E_{drr}^{(a)}$ cuando los datos se generan del modelo definido por los casos 1 y 3 respectivamente. Cada una de las cuatro sub-figuras muestran los valores de $E_{drr}^{(a)}$ para $a = 0$ y $a = 1$. La figura 5.4 muestra las gráficas los valores de $B_{drr}^{(a)}$, $a = 0, 1$, $r = 1, 2$, $d = 1, \dots, D$, $D = 100$. Esta figura está estructurada de la misma forma que la figura 5.3. Las figuras 5.5 y 5.6 están estructuradas de la misma forma que las gráficas 5.3 y 5.4 sólo que las primeras contemplan el caso $D = 200$.

En la figura 5.3 se vuelve a observar lo mismo que se ha apuntado en el párrafo anterior para las tablas 5.4 y 5.5. Cabe añadir, para el caso $D = 100$, que se observa con claridad en la figura 5.4 que los sesgos para el modelo $a = 0$ aumentan y se distribuyen peor cuando se cambia del caso primero al caso tercero sobre para el estimador $\hat{\mu}_{d1}$.

En las figuras 5.5 y 5.6 se observa que todo lo que se ha apuntado en el párrafo anterior para las figuras 5.3 y 5.4 vuelve a suceder para $D = 200$, pero de forma más acusada. Por tanto, a la vista de las gráficas señaladas, cabe añadir que hay un problema de convergencia mayor que para los modelos estudiados con anterioridad (*modelo diagonal* y *modelo AR(1)*).

(ρ_e, ρ)	a	d	E				B			
			50	100	200	400	50	100	200	400
(0, 0)	0	1	0.0046	0.0020	0.0010	0.0007	0.0014	0.0040	0.0006	0.0110
		$D/2$	0.0046	0.0022	0.0010	0.0006	0.0045	0.0122	0.0022	-0.0033
		D	0.0046	0.0020	0.0011	0.0006	-0.0071	0.0012	-0.0085	-0.0044
		mean	0.0046	0.0021	0.0011	0.0007	-0.0013	0.0021	0.0007	0.0001
	1	1	0.0044	0.0020	0.0010	0.0007	0.0014	0.0033	-0.0001	0.0116
		$D/2$	0.0044	0.0022	0.0010	0.0006	0.0076	0.0127	0.0024	-0.0027
		D	0.0044	0.0020	0.0011	0.0006	-0.0081	0.0012	-0.0087	-0.0039
		mean	0.0044	0.0021	0.0011	0.0007	-0.0005	0.0017	0.0007	0.0002
$(\frac{1}{2}, 0)$	0	1	0.0048	0.0021	0.0012	0.0005	0.0108	0.0056	-0.0107	0.0014
		$D/2$	0.0046	0.0021	0.0012	0.0005	-0.0012	-0.0056	-0.0093	-0.0048
		D	0.0047	0.0026	0.0011	0.0008	-0.0083	-0.0218	0.0011	-0.0158
		mean	0.0047	0.0022	0.0012	0.0006	-0.0005	-0.0017	-0.0002	-0.0019
	1	1	0.0055	0.0026	0.0014	0.0007	0.0164	0.0020	-0.0085	0.0012
		$D/2$	0.0052	0.0026	0.0014	0.0007	-0.0018	-0.0057	-0.0053	-0.0048
		D	0.0053	0.0031	0.0014	0.0010	-0.0094	-0.0227	0.0035	-0.0151
		mean	0.0053	0.0026	0.0014	0.0008	0.0037	-0.0041	0.0001	-0.0022
$(0, \frac{1}{2})$	0	1	0.0041	0.0021	0.0009	0.0005	0.0084	0.0049	0.0011	-0.0052
		$D/2$	0.0040	0.0022	0.0010	0.0006	0.0029	0.0100	0.0051	0.0123
		D	0.0042	0.0021	0.0009	0.0005	0.0157	0.0045	0.0016	-0.0026
		mean	0.0041	0.0022	0.0010	0.0006	-0.0012	-0.0006	0.0008	-0.0011
	1	1	0.0044	0.0023	0.0011	0.0006	0.0106	0.0057	-0.0013	-0.0064
		$D/2$	0.0043	0.0024	0.0011	0.0008	0.0033	0.0090	0.0045	0.0138
		D	0.0044	0.0023	0.0011	0.0006	0.0135	0.0055	0.0048	-0.0001
		mean	0.0044	0.0024	0.0011	0.0007	-0.0001	-0.0001	0.0005	-0.0009
$(\frac{1}{2}, \frac{1}{2})$	0	1	0.0038	0.0017	0.0010	0.0005	0.0101	0.0076	-0.0101	0.0011
		$D/2$	0.0037	0.0018	0.0009	0.0005	0.0071	0.0115	0.0025	0.0037
		D	0.0036	0.0018	0.0009	0.0005	0.0037	-0.0107	-0.0014	0.0017
		mean	0.0037	0.0017	0.0009	0.0006	0.0009	0.0007	-0.0007	0.0020
	1	1	0.0037	0.0018	0.0009	0.0005	0.0070	0.0081	-0.0079	0.0007
		$D/2$	0.0037	0.0018	0.0009	0.0005	0.0059	0.0088	0.0024	0.0038
		D	0.0037	0.0018	0.0009	0.0005	0.0016	-0.0087	-0.0021	0.0032
		mean	0.0038	0.0018	0.0010	0.0006	0.0002	0.0010	-0.0005	0.0022

Tabla 5.4: $E_{d11}^{(a)}$ (izquierda) y $B_{d11}^{(a)}$ (derecha) para $\rho_x = 0,5$, $a = 0, 1$.

		E				B				
(ρ_e, ρ)	a	d	50	100	200	400	50	100	200	400
$(0, 0)$	0	1	0.0166	0.0079	0.0045	0.0023	0.0230	0.0073	-0.0119	0.0092
		$D/2$	0.0163	0.0079	0.0054	0.0043	0.0144	-0.0078	0.0325	-0.0461
		D	0.0166	0.0078	0.0045	0.0023	0.0247	-0.0058	0.0149	0.0102
		mean	0.0165	0.0082	0.0047	0.0026	0.0112	-0.0023	0.0019	0.0009
	1	1	0.0166	0.0076	0.0045	0.0023	0.0234	0.0066	-0.0132	0.0098
		$D/2$	0.0163	0.0076	0.0054	0.0042	0.0182	-0.0053	0.0334	-0.0451
		D	0.0163	0.0076	0.0045	0.0023	0.0222	-0.0094	0.0134	0.0088
		mean	0.0165	0.0079	0.0047	0.0026	0.0125	-0.0029	0.0019	0.0012
$(\frac{1}{2}, 0)$	0	1	0.0212	0.0096	0.0048	0.0023	0.0442	-0.0043	-0.0183	0.0108
		$D/2$	0.0193	0.0100	0.0056	0.0022	-0.0115	0.0228	0.0339	-0.0006
		D	0.0193	0.0097	0.0051	0.0041	-0.0168	-0.0155	-0.0264	-0.0436
		mean	0.0195	0.0100	0.0048	0.0026	0.0067	-0.0057	-0.0020	0.0018
	1	1	0.0239	0.0110	0.0057	0.0031	0.0399	0.0020	-0.0126	0.0145
		$D/2$	0.0227	0.0114	0.0062	0.0028	-0.0228	0.0221	0.0246	0.0001
		D	0.0221	0.0110	0.0059	0.0047	-0.0062	-0.0101	-0.0178	-0.0434
		mean	0.0226	0.0114	0.0059	0.0032	0.0114	-0.0099	-0.0005	0.0001
$(0, \frac{1}{2})$	0	1	0.0132	0.0084	0.0041	0.0019	0.0014	0.0239	-0.0028	0.0069
		$D/2$	0.0139	0.0079	0.0042	0.0022	-0.0285	-0.0094	-0.0066	-0.0200
		D	0.0131	0.0080	0.0046	0.0021	-0.0093	-0.0142	-0.0216	0.0178
		mean	0.0135	0.0081	0.0045	0.0022	-0.0043	-0.0004	-0.0038	0.0010
	1	1	0.0144	0.0092	0.0043	0.0020	0.0050	0.0244	-0.0022	0.0087
		$D/2$	0.0152	0.0086	0.0045	0.0025	-0.0297	-0.0007	-0.0109	-0.0229
		D	0.0143	0.0087	0.0049	0.0022	-0.0079	-0.0116	-0.0231	0.0173
		mean	0.0147	0.0089	0.0047	0.0023	-0.0021	0.0000	-0.0052	0.0015
$(\frac{1}{2}, \frac{1}{2})$	0	1	0.0135	0.0067	0.0053	0.0022	-0.0245	-0.0048	0.0357	-0.0190
		$D/2$	0.0134	0.0070	0.0040	0.0018	0.0218	-0.0178	-0.0043	-0.0040
		D	0.0128	0.0067	0.0042	0.0020	0.0068	0.0026	-0.0137	-0.0114
		mean	0.0132	0.0072	0.0045	0.0022	0.0012	0.0071	-0.0011	0.0033
	1	1	0.0142	0.0071	0.0055	0.0021	-0.0258	-0.0077	0.0379	-0.0152
		$D/2$	0.0137	0.0074	0.0041	0.0019	0.0145	-0.0184	-0.0049	-0.0050
		D	0.0136	0.0070	0.0042	0.0020	0.0124	0.0050	-0.0106	-0.0099
		mean	0.0139	0.0075	0.0045	0.0023	-0.0003	0.0080	-0.0006	0.0039

Tabla 5.5: $E_{d11}^{(a)}$ (izquierda) y $B_{d11}^{(a)}$ (derecha) para $\rho_x = 0,5$, $a = 0, 1$.

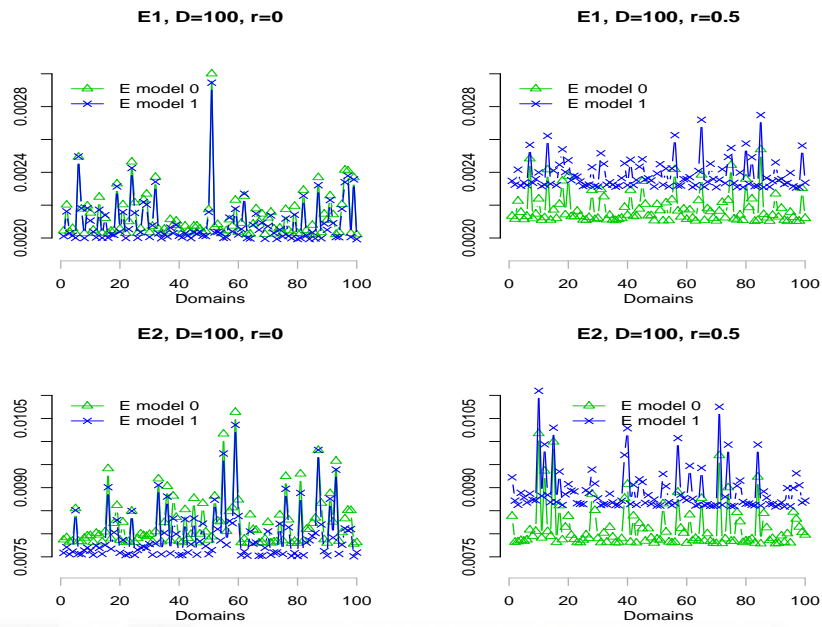


Figura 5.3: E_{drr} , para $a = 0, 1, r = 1, 2, \rho_x = 1/2, \rho_e = 0, D = 100$.

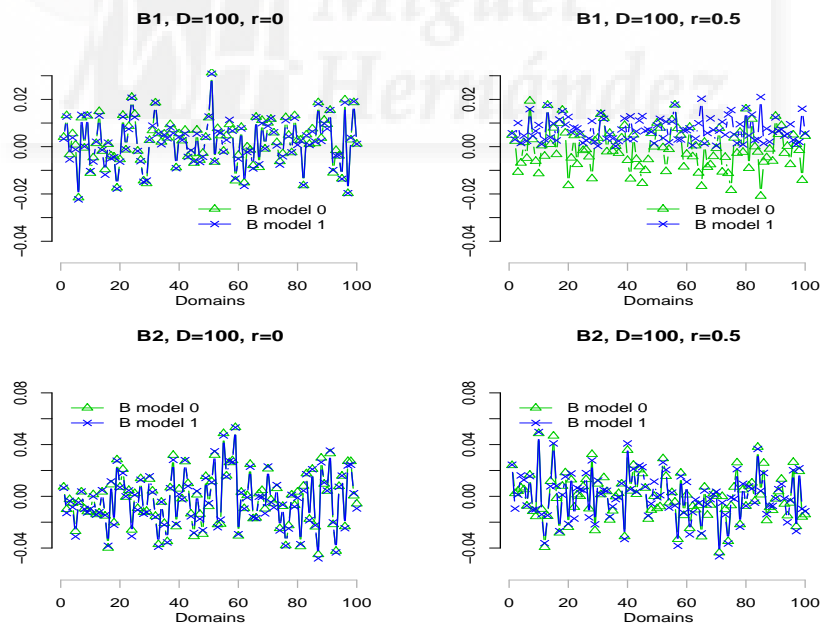


Figura 5.4: B_{drr} , para $a = 0, 1, r = 1, 2, \rho_x = 1/2, \rho_e = 0, D = 100$.

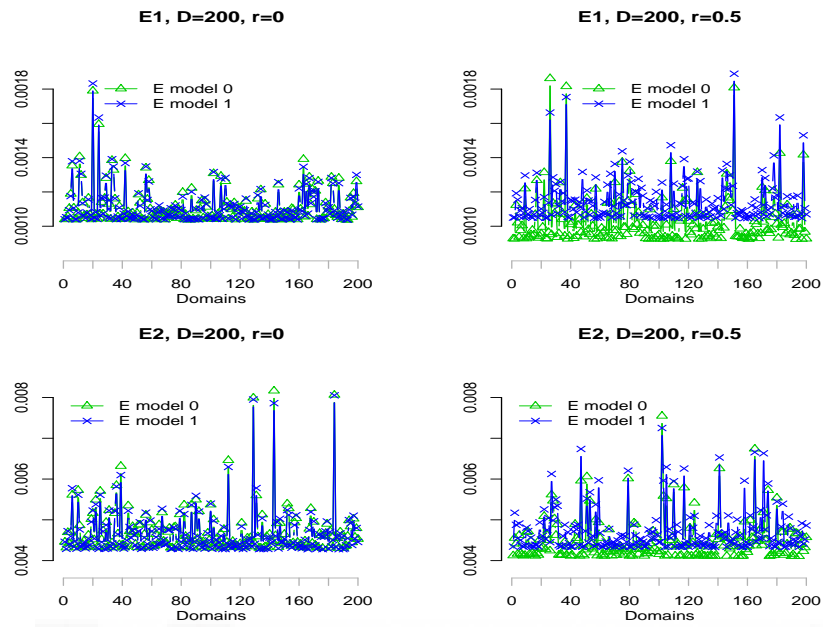


Figura 5.5: E_{drr} , para $a = 0, 1$, $r = 1, 2$, $\rho_x = 1/2$, $\rho_e = 0$, $D = 200$.

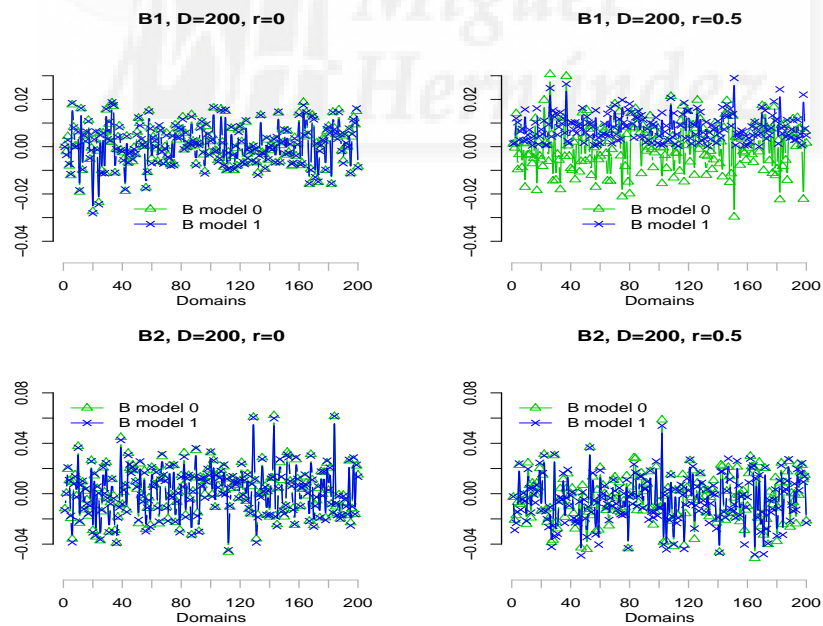


Figura 5.6: B_{drr} , para $a = 0, 1$, $r = 1, 2$, $\rho_x = 1/2$, $\rho_e = 0$, $D = 200$.

5.2.3. Experimento de simulación 3

El objetivo de este experimento es comprobar el funcionamiento del bootstrap paramétrico en la estimación de los errores cuadráticos medios del EBLUP de las medias poblacionales en un modelo multivariante AR(1) heterocedástico. En esta simulación se consideran las correlaciones $\rho_e = \rho = 1/2$ del caso 4. Los valores de los restantes parámetros son los mismos que en la simulación 1; es decir, $\beta_1 = 1, \beta_2 = 1, \sigma_{u_1}^2 = 2, \sigma_{u_2}^2 = 4, \sigma_{d11} = 1$ y $\sigma_{d22} = 2$. En este apartado se usa la notación $\theta = (\theta_1, \theta_2, \theta_3)$, donde $\theta_1 = \sigma_{u_1}^2, \theta_2 = \sigma_{u_2}^2$ y $\theta_3 = \rho$. Las variables auxiliares x_{dr} se generan de la misma forma que en la simulación 1.

Los pasos del experimento de simulación son

1. Repetir $I = 500$ veces ($i = 1, \dots, 500$)

1.1. Generar una muestra $(y_{dr}^{(i)}, x_{dr}^{(i)})$, $d = 1, \dots, D, r = 1, 2$ (cf. A-D en Sección 1).

1.2. Calcular $\mu_d^{(i)} = X_d^{(i)}\beta + I_2 u_d^{(i)}$.

1.3. Calcular $\hat{\theta}^{(i)}, \hat{\beta}_{E1}^{(i)}$ y $\hat{\beta}_{E2}^{(i)}$.

1.4. Para $d = 1, \dots, D$, calcular $\hat{u}_{Ed}^{(i)}$, usando $\hat{\theta}^{(i)}, \hat{\beta}_{Er}^{(i)}$. Calcular

$$\hat{\mu}_d^{(i)} = X_d^{(i)}\hat{\beta}_E^{(i)} + I_2 \hat{u}_d^{(i)}, \quad mse_d^{(i)} = G_{1d}^{(i)}(\hat{\theta}^{(i)}) + G_{2d}^{(i)}(\hat{\theta}^{(i)}) + 2G_{3d}^{(i)}(\hat{\theta}^{(i)})$$

1.5. Repetir $B = 200$ veces ($b = 1, \dots, B$)

1.5.1. Generar $u_d^{*(ib)}, e_{dr}^{*(ib)}$, $d = 1, \dots, D, r = 1, 2$ (cf. B-C en Sección 1), pero usando $\hat{\theta}^{(i)}$ en lugar de θ .

1.5.2. Generar una muestra bootstrap $(y_{dr}^{*(ib)}, x_{dr}^{(i)})$, $d = 1, \dots, D, r = 1, 2$, del modelo

$$y_{dr}^{*(ib)} = x_{dr}^{(i)}\hat{\beta}_{Rr}^{(i)} + u_{dr}^{*(ib)} + e_{dr}^{*(ib)}.$$

1.5.3. Calcular $\mu_d^{*(ib)} = X_d^{(i)}\hat{\beta}_E^{(i)} + u_d^{*(ib)}$.

1.5.4. Calcular $\hat{\theta}^{*(ib)}$ a partir de $\hat{\theta}$, reemplazando convenientemente los elementos de la muestra bootstrap.

1.5.5. Calcular $\hat{\beta}_{Br}^{*(ib)}$ y $\hat{\beta}_{Er}^{*(ib)}$, las versiones bootstrap $\hat{\beta}_{Br}$ y $\hat{\beta}_{Er}$, $r = 1, 2$, respectivamente. Calculados usando $\hat{V}_d^{(i)}$ e $y_d^{*(ib)}$ para el cálculo de $\hat{\beta}_{Br}^{*(ib)}$, y $\hat{V}_d^{*(ib)}$ e $y_d^{*(ib)}$ para el cálculo de $\hat{\beta}_{Er}^{*(ib)}$.

1.5.6. Para $d = 1, \dots, D$ y $r = 1, 2$; calcular $\hat{u}_{Bd}^{*(ib)}$ y $\hat{u}_d^{*(ib)}$, a partir de $\hat{\theta}^{(i)}$ y $\hat{\beta}_{Br}^{*(ib)}$, $\hat{\theta}^{*(ib)}$ y $\hat{\beta}_{Er}^{*(ib)}$, $r=1,2$ respectivamente.

1.5.7. Para $d = 1, \dots, D$, calcular

$$\hat{\mu}_{Bd}^{*(ib)} = X_d^{(i)}\hat{\beta}_B^{*(ib)} + I_2 \hat{u}_{Bd}^{*(ib)} \quad \text{y} \quad \hat{\mu}_{Ed}^{*(ib)} = X_d^{(i)}\hat{\beta}_E^{*(ib)} + I_2 \hat{u}_d^{*(ib)}.$$

1.5.8. Para $d = 1, \dots, D$, calcular

$$\delta_{Ed}^{*(ib)} = (\hat{\mu}_d^{*(ib)} - \mu_d^{*(ib)}), \quad \delta_{Bd}^{*(ib)} = (\hat{\mu}_{Bd}^{*(ib)} - \mu_d^{*(ib)}), \quad \delta_{EBd}^{*(ib)} = (\hat{\mu}_d^{*(ib)} - \hat{\mu}_{Bd}^{*(ib)}).$$

1.6 Para $d = 1, \dots, D$, calcular

$$\begin{aligned} mse_d^{*1(i)} &= \frac{1}{B} \sum_{b=1}^B \delta_{Ed}^{*(ib)} \delta_{Ed}^{*(ib)t} \\ mse_d^{*2(i)} &= G_{1d}^{(i)}(\hat{\theta}^{(i)}) + G_{2d}^{(i)}(\hat{\theta}^{(i)}) + \frac{1}{B} \sum_{b=1}^B \delta_{EBd}^{*(ib)} \delta_{EBd}^{*(ib)t} \\ mse_d^{*3(i)} &= 2[G_{1d}^{(i)}(\hat{\theta}^{(i)}) + G_{2d}^{(i)}(\hat{\theta}^{(i)})] - \frac{1}{B} \sum_{b=1}^B [G_1(\hat{\theta}^{*(ib)}) + G_2(\hat{\theta}^{*(ib)})] + \frac{1}{B} \sum_{b=1}^B \delta_{EBd}^{*(ib)} \delta_{EBd}^{*(ib)t}. \end{aligned}$$

2. Salida:

$$mse_d = \frac{1}{I} \sum_{i=1}^I mse_d^{(i)}, \quad mse_d^{*\ell} = \frac{1}{I} \sum_{i=1}^I mse_d^{*\ell(i)}, \quad \ell = 1, 2, 3.$$

3. Leer los MSE_{drr} obtenidos en la simulación 1 para el caso $\rho = \rho_e = \frac{1}{2}$ y hacer

$$\begin{aligned} B_{drr}^0 &= \frac{1}{I} \sum_{i=1}^I (mse_{dr}^{(i)} - MSE_{drr}), \quad B_{drr}^{*\ell} = \frac{1}{I} \sum_{i=1}^I (mse_{dr}^{*\ell(i)} - MSE_{drr}), \quad \ell = 1, 2, 3, r = 1, 2, \\ E_{drr}^0 &= \frac{1}{I} \sum_{i=1}^I (mse_{dr}^{(i)} - MSE_{drr})^2, \quad E_{drr}^{*\ell} = \frac{1}{I} \sum_{i=1}^I (mse_{dr}^{*\ell(i)} - MSE_{drr})^2, \quad \ell = 1, 2, 3, r = 1, 2, \\ B_{rr}^0 &= \frac{1}{D} \sum_{d=1}^D B_{drr}^0, \quad B_{rr}^{*\ell} = \frac{1}{D} \sum_{d=1}^D B_{drr}^{*\ell}, \quad E_{rr}^0 = \frac{1}{D} \sum_{d=1}^D E_{drr}^0, \quad E_{rr}^{*\ell} = \frac{1}{D} \sum_{d=1}^D E_{drr}^{*\ell}, \quad \ell = 1, 2, 3, r = 1, 2. \end{aligned}$$

D	E_{11}^0	E_{11}^{*1}	E_{11}^{*2}	E_{11}^{*3}	E_{22}^0	E_{22}^{*1}	E_{22}^{*2}	E_{22}^{*3}
50	0.00255	0.00767	0.00280	0.00311	0.01142	0.03415	0.01454	0.01391
100	0.00178	0.00683	0.00200	0.00206	0.00797	0.02817	0.00859	0.00912
200	0.00104	0.00591	0.00110	0.00107	0.00398	0.02369	0.00408	0.00416
400	0.00059	0.00535	0.00061	0.00059	0.00232	0.02156	0.00235	0.00232

Tabla 5.6: $E_{rr}^0, E_{rr}^{*\ell}, \ell = 1, 2, 3$.

D	B_{11}^0	B_{11}^{*1}	B_{11}^{*2}	B_{11}^{*3}	B_{22}^0	B_{22}^{*1}	B_{22}^{*2}	B_{22}^{*3}
50	0.01665	-0.00532	-0.00546	0.00355	0.00750	-0.03562	-0.03558	0.00050
100	0.00219	-0.00765	-0.00812	-0.00193	0.00910	-0.01090	-0.01100	0.00585
200	0.00029	-0.00424	-0.00452	-0.00024	0.00570	-0.00369	-0.00373	0.00526
400	-0.00130	-0.00391	-0.00365	-0.00127	0.00177	-0.00311	-0.00289	0.00192

Tabla 5.7: $B_{rr}^0, B_{rr}^{*\ell}, \ell = 1, 2, 3$.

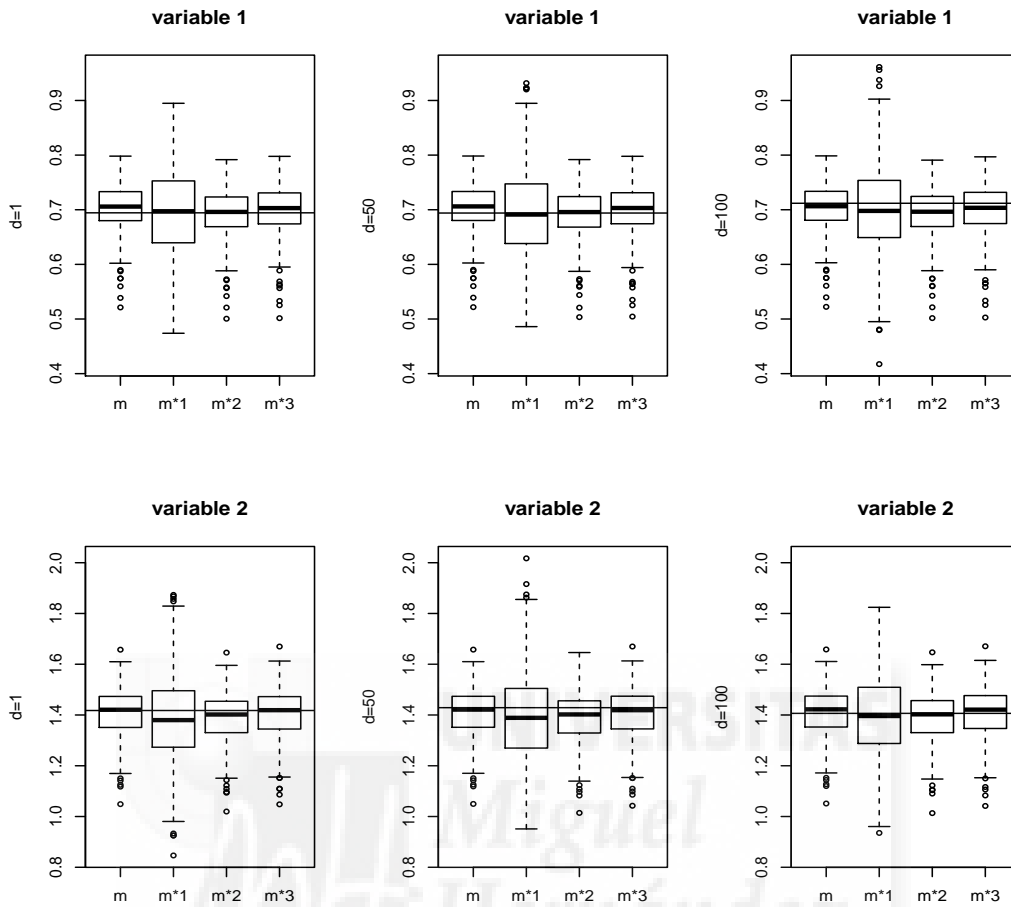


Figura 5.7: Diagrama de cajas de $msed_{dr}$'s, $msed_{dr}^{*1}$'s, $msed_{dr}^{*2}$'s y $msed_{dr}^{*3}$ para $D = 100$

En la tabla 5.6 se apunta en la primera columna el número de áreas consideradas en la simulación; es decir, $D = 50$, $D = 100$, $D = 200$ y $D = 400$. En las cuatro columnas siguientes se muestra el error cuadrático medio de los estimadores $msed$, $msed^{*1}$, $msed^{*2}$ y $msed^{*3}$ del error cuadrático medio de $\hat{\mu}_{d1}$, donde el valor teórico considerado es el obtenido en la simulación 1. En las cuatro columnas siguientes se muestra lo mismo, pero esta vez para el estimador $\hat{\mu}_{d2}$. La disposición de las columnas en la tabla 5.7 es la misma pero esta vez para el sesgo.

En la tabla 5.6 se observa claramente que el estimador $msed^{*1}$ produce los mayores errores y los estimadores $msed$ y $msed^{*3}$ los menores. Asimismo se observa que el error disminuye de forma considerable en los cuatro estimadores al aumentar el número de áreas consideradas. En la tabla 5.7 se observa que los sesgos para los estimadores $msed^{*1}$ y $msed^{*2}$ son negativos, lo cual era de esperar debido a que les afecta que el valor esperado del término $G_1(\hat{\theta})$, que es de forma aproximada $G_1(\theta) - G_3(\theta)$. También se aprecia una presencia mayor del sesgo negativo en el estimador $msed^{*3}$ posiblemente debida a la naturaleza de la matriz de varianzas de los efectos aleatorios del modelo, ya que dicha matriz afecta considerablemente a la estimación del término G_3 y consecuentemente a sus correcciones. Por último, cabe apuntar también que se

ve con claridad la presencia de sesgo positivo para el estimador $msed$, ello es debido a que la estimación hecha a partir de las términos $G_1 - G_3$ tiende a ser mayor que los valores teóricos que se han obtenido en la simulación 1.

La figura 5.7 contiene los diagramas de cajas de los cuatro estimadores considerados. Se observa que la distribución de los estimadores $msed$ y $msed^{*3}$ es mejor que la de los dos restantes. Se nota de forma evidente la presencia de sesgo negativo para los estimadores $msed^{*1}$ y $msed^{*2}$ y se aprecia una ligera tendencia negativa para el sesgo del estimador $msed^{*3}$ sobre todo en la primera variable del modelo. Por último se aprecia la presencia de sesgo positivo para el estimador $msed$.





6

Estimación bivalente de indicadores de pobreza

Los capítulos anteriores describen algunos de los modelos de área multivariantes que se pueden plantear a partir del modelo general presentado en el capítulo 2. Ahora se presenta una aplicación práctica al estudio de indicadores de pobreza. Las estimaciones se obtienen usando los predictores EBLUP basados en los modelos estudiados. Los errores cuadráticos medios se estiman aplicando el estimador (2.8) basado en la metodología de Prasad y Rao (1990).

6.1. Datos y modelo

Consideremos una población finita P (en el caso del presente estudio se trata de España) cuyas unidades j son individuos. Suponemos que la población P está dividida en D dominios (provincias cruzadas con sexo) que se denotan por P_d . El subíndice d (provincia-sexo) se determina según la provincia donde está establecida la familia a la que pertenece el individuo j y el sexo del mismo. El tamaño de P se denota por N y el de cada P_d por N_d . Se tiene que

$$P = \bigcup_{d=1}^D P_d \quad \text{y} \quad N = \sum_{d=1}^D N_d.$$

Sea z_{dj} el ingreso normalizado neto del individuo j . El Instituto Nacional de Estadística (INE) calcula el valor de cada z_{dj} sumando los ingresos netos anuales de los miembros de la familia a la que pertenece el individuo j y dividiendo el resultado obtenido por el tamaño familiar normalizado. El resultado de la división se asigna a cada uno de los miembros de la familia en cuestión. Así pues, el valor z_{dj} es el mismo para todos los miembros de una unidad familiar. El propósito que se persigue al normalizar los ingresos familiares es recoger la variabilidad existente en tamaño y composición de las familias.

El tamaño familiar normalizado se calcula con la escala modificada OECD que emplea EUROSTAT. Esa escala asigna un peso igual a 1 al primer adulto, 0,5 al segundo y al resto de miembros que tengan una

edad igual o superior a catorce años, y, por último 0,3 a cada miembro que tenga una edad inferior a catorce años. Lo anterior se denota de la forma que se indica a continuación

$$H_{dh} = 1 + 0,5(H_{dh \geq 14} - 1) + 0,3H_{dh < 14}$$

donde $H_{dh \geq 14}$ es el número de personas con edad igual o superior a catorce años en la familia h del dominio d y $H_{dh < 14}$ es el número de personas que tienen una edad inferior a catorce años.

Sea z el umbral de pobreza, de modo que los individuos j cuyos ingresos normalizados netos están por debajo del umbral de pobreza, $z_{dj} < z$, se dice que están en riesgo de pobreza. Teniendo en cuenta las directrices que marca la Oficina Estadística Europea (EUROSTAT) el umbral de pobreza se establece en el 60% de la mediana de los ingresos familiares netos anuales de los hogares españoles y se calcula usando los datos de la encuesta de condiciones de vida. En los años que se han tenido en cuenta la medida anterior resultó ser $z_{2005} = 6160$ y $z_{2006} = 6556$ y son las que se utilizan para determinar los estimadores directos de los indicadores de pobreza.

El objetivo de la aplicación práctica que se desarrolla en el presente capítulo es la estimación de la proporción de pobreza y la brecha de pobreza por provincia y sexo. Estos indicadores de pobreza son

$$y_{d1} = \frac{1}{N_d} \sum_{j=1}^{N_d} y_{d1j} \quad \text{e} \quad y_{d2} = \frac{1}{N_d} \sum_{j=1}^{N_d} y_{d2j},$$

respectivamente, donde

$$y_{d1j} = I(z_{dj} < z) \quad \text{e} \quad y_{d2j} = \left(\frac{z - z_{dj}}{z} \right) I(z_{dj} < z).$$

En las expresiones que se utilizan para las variables y_{d1} e y_{d2} se tiene que $I(z_{dj} < z) = 1$ si $z_{dj} < z$ y $I(z_{dj} < z) = 0$ en caso contrario. El indicador y_{d1} da la condición de pobreza y el indicador y_{d2} mide la distancia relativa de un individuo pobre al umbral de pobreza.

Para las tareas de ajuste de los modelos estadísticos de área se utilizan los datos de la encuesta de condiciones de vida (ECV) de los años 2005 y 2006 con tamaños muestrales 37491 y 34694 respectivamente. La ECV es la versión española de la "European Survey on Income and Living Conditions" (EU-SILC). La ECV comenzó a hacerse en 2004 y a partir de esa fecha se realiza con periodicidad anual. Su propósito principal es conseguir información para comparar estadísticamente la distribución de los ingresos familiares en el entorno europeo.

La ECV no proporciona estimaciones oficiales en cada dominio, no obstante el estimador directo del total es $Y_d = \sum_{j=1}^{N_d} y_{dj}$ es

$$\hat{Y}_d^{dir} = \sum_{j \in S_d} w_{dj} y_{dj},$$

donde S_d es la muestra observada en el dominio d y w_{dj} es el peso de diseño corregido por falta de respuesta y calibrado (factor de elevación). En el caso particular en el que $y_{dj} = 1$ para todo $j \in P_d$ se obtiene una estimación del tamaño del dominio d , es decir,

$$\hat{N}_d^{dir} = \sum_{j \in S_d} w_{dj}.$$

Utilizando la estimación del total anterior se puede construir un estimador directo para la media del dominio d , en concreto es $\bar{y}_d = \hat{Y}_d^{dir} / \hat{N}_d^{dir}$. Las estimaciones directas de las medias de los dominios se utilizan como observaciones de la variable de interés en el modelo de área. Las varianzas de estas estimaciones pueden aproximarse mediante la expresión que sigue

$$\hat{V}_\pi(\hat{Y}_d^{dir}) = \sum_{j \in S_d} w_{dj}(w_{dj} - 1)(y_{dj} - \bar{y}_d)^2 \quad \text{y} \quad \sigma_{\pi,d}^2 = \hat{V}_\pi(\bar{y}_d) = \hat{V}_\pi(\hat{Y}_d^{dir}) / \hat{N}_d^2$$

Las fórmulas precedentes se basan en las que aparecen en Särndal et al. (1992), pp. 43, 185 y 391, empleando las simplificaciones $w_{dj} = \frac{1}{\pi_{dj}}$, $\pi_{dtj,dtj} = \pi_{dj}$ y $\pi_{dj,dtj} = \pi_{dj}\pi_{dj}$, $i \neq j$ en las probabilidades de inclusión de segundo orden.

En esta sección se utilizan datos de 2005 y 2006 de la encuesta de condiciones de vida para la estimación de los indicadores de pobreza. Los dominios de interés son el resultado de combinar las distintas provincias con las dos categorías de la variable sexo. Hay $D = 104$ dominios, obtenidos a partir de las 52 provincias (incluyendo a Ceuta y a Melilla) y de las subpoblaciones de hombres y de mujeres. Los cuartiles de la distribución de los tamaños muestrales de los dominios son 13, 149, 251, 530, 1494 en el año 2005 son. En el año 2006 son 18, 129, 233, 481, 1494. Como puede apreciarse los tamaños muestrales son demasiado pequeños para emplear estimadores directos en las estimaciones de los parámetros de interés en todas los dominios.

A continuación, se expone la codificación de las distintas provincias que se han considerado. Las provincias españolas se codifican de la forma siguientes: 1 Álava, 2 Albacete, 3 Alicante, 4 Almería, 5 Ávila, 6 Badajoz, 7 Baleares, 8 Barcelona, 9 Burgos, 10 Cáceres, 11 Cádiz, 12 Castellón, 13 Ciudad Real, 14 Córdoba, 15 Coruña, La, 16 Cuenca, 17 Gerona, 18 Granada, 19 Guadalajara, 20 Guipúzcoa, 21 Huelva, 22 Huesca, 23 Jaén, 24 León, 25 Lérida, 26 La Rioja, 27 Lugo, 28 Madrid, 29 Málaga, 30 Murcia, 31 Navarra, 32 Orense, 33 Asturias (Oviedo), 34 Palencia, 35 Palmas Las, 36 Pontevedra, 37 Salamanca, 38 Santa Cruz de Tenerife, 39 Cantabria (Santander), 40 Segovia, 41 Sevilla, 42 Soria, 43 Tarragona, 44 Teruel, 45 Toledo, 46 Valencia, 47 Valladolid, 48 Vizcaya, 49 Zamora, 50 Zaragoza, 51 Ceuta, 52 Melilla

A partir de la información auxiliar disponible, se construyen las siguientes variables auxiliares:

1. El total poblacional del dominio.
2. Los grupos de edad son cinco y se corresponden con los intervalos: ≤ 15 (age1), $16 - 24$ (age2), $25 - 49$ (age3), $50 - 64$ (age4) y ≥ 65 (age5).
3. El nivel de estudios se contempla formando cuatro categorías distintas: la primera para nivel de estudios inferior a la educación primaria (edu0), la segunda para la educación primaria (edu1), la tercera para la educación secundaria (edu2) y la cuarta para nivel de estudios universitario (edu3).
4. La nacionalidad comprende dos categorías: la primera para la nacionalidad española (cit1) y la segunda para nacionalidades no españolas (cit2).
5. La situación laboral comprende cuatro categorías: la primera para menores de 16 años (lab0), la segunda para empleados (lab1), la tercera para parados (lab2) y la cuarta para inactivos (lab3).

En los modelos que se van a plantear se consideran variables explicativas extraídas de las que acabamos de explicar. Para poder expresar el modelo de una forma más clara, se adopta la notación

$$x_{d1} = (x_{d11}, x_{d12}, \dots, x_{d1p_1}) \quad \text{y} \quad x_{d2} = (x_{d21}, x_{d22}, \dots, x_{d2p_2}),$$

donde $d = 1, \dots, 104$. La primera variable es constante; en concreto $x_{dr1} = 1$, $r = 1, 2$. El resto de variables auxiliares, x_{drj} para $j > 1$, son las proporciones que resultan al dividir el total poblacional de la categoría considerada por el total poblacional del dominio. Cada una de las variables anteriores tiene asociado un parámetro de regresión en el modelo de área y por tal motivo se agrupan de la forma

$$\beta = \begin{pmatrix} \beta_1 \\ \beta_2 \end{pmatrix}, \quad \beta'_1 = (\beta_{11}, \beta_{12}, \dots, \beta_{1p_1}), \quad \beta'_2 = (\beta_{21}, \beta_{22}, \dots, \beta_{2p_2}).$$

Se definen además los vectores y matrices

$$y_d = \begin{pmatrix} y_{d1} \\ y_{d2} \end{pmatrix}, \quad X_d = \begin{pmatrix} x_{d1} & 0 \\ 0 & x_{d2} \end{pmatrix}, \quad u_d = \begin{pmatrix} u_{d1} \\ u_{d2} \end{pmatrix}, \quad e_d = \begin{pmatrix} e_{d1} \\ e_{d2} \end{pmatrix},$$

$$y = \underset{1 \leq d \leq 104}{\text{col}}(y_d), \quad X = \underset{1 \leq d \leq 104}{\text{col}}(X_d), \quad u = \underset{1 \leq d \leq 104}{\text{col}}(u_d), \quad e = \underset{1 \leq d \leq 104}{\text{col}}(e_d),$$

$$Z_d = \underset{1 \leq \ell \leq 104}{\text{col}}(\delta_{\ell d} I_2), \quad Z = \underset{1 \leq d \leq 104}{\text{col}'}(Z_d) = I_{208}.$$

En los modelos de área considerados intervienen todos los elementos expuestos y consecuentemente admiten la representación lineal

$$y = X\beta + Zu + e = X\beta + Z_1u_1 + \dots + Z_Du_D + e,$$

donde e y u son vectores aleatorios independientes que verifican

$$u \sim N(0, V_u), \quad V_u = \underset{1 \leq d \leq 104}{\text{diag}}(V_{ud}), \quad e \sim N(0, V_e), \quad V_e = \underset{1 \leq d \leq 104}{\text{diag}}(W_d^{-1}),$$

donde, para $d = 1, \dots, 104$, las matrices V_{ud} dependen de m parámetros desconocidos y las matrices W_d^{-1} son conocidas. Finalmente, la forma matricial del modelo es

$$\begin{pmatrix} y_{11} \\ y_{12} \\ y_{21} \\ y_{22} \\ \vdots \\ y_{D1} \\ y_{D2} \end{pmatrix} = \begin{pmatrix} x_{11} & 0 \\ 0 & x_{12} \\ x_{21} & 0 \\ 0 & x_{22} \\ \vdots & \\ x_{D1} & 0 \\ 0 & x_{D2} \end{pmatrix} \begin{pmatrix} \beta_1 \\ \beta_2 \end{pmatrix} + \begin{pmatrix} u_{11} \\ u_{12} \\ u_{21} \\ u_{22} \\ \vdots \\ u_{D1} \\ u_{D2} \end{pmatrix} + \begin{pmatrix} e_{11} \\ e_{12} \\ e_{21} \\ e_{22} \\ \vdots \\ e_{D1} \\ e_{D2} \end{pmatrix}.$$

6.2. Estimación basada en modelos de área (enfoque multivariante)

Como se ha apuntado al inicio del capítulo, en esta sección se estiman indicadores de pobreza utilizando cada uno de los modelos que se han estudiado. Para ello se presentan las variables explicativas que se incluyen y las hipótesis que se asumen sobre los efectos aleatorios. Todo ello complementa la descripción del modelo de área multivariante dada en la sección 6.1. Se presentan también las estimaciones de los parámetros y las interpretaciones que resulten oportunas en cada caso.

6.2.1. Modelo con varianza de los efectos diagonal

Para la variable y_{d1} se consideran las variables explicativas constante, age1, age2, edu1, cit1 y lab2, y para y_{d2} se consideran las variables constante, edu0, edu1, edu2, cit1 y lab1. En esta sección se supone que los efectos que corresponden a las distintos dominios verifican

$$u \sim N(0, V_u), \quad V_u = \text{diag}_{1 \leq d \leq 104} (V_{ud}), \quad V_{ud} = \begin{pmatrix} \sigma_{u1}^2 & 0 \\ 0 & \sigma_{u2}^2 \end{pmatrix}.$$

Con el objeto de obtener una estimación inicial de los parámetros σ_{u1}^2 y σ_{u2}^2 , se consideran los modelos univariantes que se deducen a partir del modelo multivariante planteado en la sección 6.1; es decir,

$$y_{dr} = x_{dr}\beta_r + u_{dr} + e_{dr}, \quad d = 1, \dots, D, \quad r = 1, 2,$$

donde

$$u_{.r} \sim N_D(0, \sigma_{ur}^2 I_D), \quad e_{.r} \sim N_D(0, W_r^{-1}), \quad W_r^{-1} = \text{diag}_{1 \leq d \leq D} (\sigma_{edr}^{-2}).$$

Se hace una primera estimación de las componentes de la varianza utilizando la fórmula del método de Henderson 3; es decir,

$$\hat{\sigma}_{urH}^2 = \frac{y_{.r}' P_2 y_{.r} - D + p_r}{\text{tr}\{P_2\}}, \quad r = 1, 2.$$

Las estimaciones $\hat{\sigma}_{u1H}^2$ y $\hat{\sigma}_{u2H}^2$ se utilizan como semillas en el algoritmo de Fisher-scoring. A partir de los datos de la muestra, las estimaciones que se obtienen de las semillas son

$$\hat{\sigma}_{u1H}^2 = 0,00095 \quad \text{y} \quad \hat{\sigma}_{u2H}^2 = 0,00188.$$

Partiendo de los valores anteriores, se ejecuta el algoritmo Fisher-scoring para obtener las estimaciones REML de σ_{u1}^2 y σ_{u2}^2 . Después de seis iteraciones se obtiene

$$\hat{\sigma}_{u1}^2 = 0,00138 \quad \text{y} \quad \hat{\sigma}_{u2}^2 = 0,00037.$$

A partir de lo anterior se construyen las estimaciones para los parámetros β_{rj} , $r = 1, 2$. Utilizando también las distribuciones asintóticas para los estimadores $\hat{\beta}_{rj}$ presentadas en el capítulo 2, se determinan los p -valores correspondientes a los contrastes $H_0 : \beta_{rj} = 0$. Todo ello se presenta en las tablas 6.1-6.2.

Variabes	<i>constante</i>	<i>age1</i>	<i>age2</i>	<i>edu1</i>	<i>cit1</i>	<i>lab2</i>
β	-0.74414	0.96641	1.65996	0.76452	0.33312	1.86921
<i>p</i> -valor	0.00000	0.00057	0.00033	0.00000	0.00055	0.00000

Tabla 6.1. Parámetros de regresión y *p*-valores para $\alpha = 0$.

Variabes	<i>constante</i>	<i>edu0</i>	<i>edu1</i>	<i>edu2</i>	<i>cit1</i>	<i>lab1</i>
β	-0.38295	0.99395	0.34896	0.17409	0.15597	-0.06853
<i>p</i> -valor	0.00000	0.00000	0.00000	0.09457	0.00077	0.01613

Tabla 6.2. Parámetros de regresión y *p*-valores para $\alpha = 1$.

A continuación se construyen los intervalos de confianza correspondientes, donde el nivel de significación considerado es $\alpha = 0,1$. Para el presente caso, adoptan la forma $\hat{\beta}_{rj} \pm \sqrt{q}z_{\alpha/2}$, donde q es el elemento de la diagonal principal de la matriz Q , definida en la sección 2.4, que se corresponde con $\hat{\beta}_{rj}$. En la última columna de la tabla se incluye “V” o “F” según 0 pertenezca al intervalo de confianza o no. Lo anterior se resume en las tablas 6.3-6.4.

Variabes	$\hat{\beta}_{1j}$	$\hat{\beta}_{1j} - \sqrt{q}z_{\alpha/2}$	$\hat{\beta}_{1j} + \sqrt{q}z_{\alpha/2}$	$0 \in IC$
<i>constante</i>	-0.74414	-0.92148	-0.56680	<i>F</i>
<i>age1</i>	0.96641	0.50492	1.42800	<i>F</i>
<i>age2</i>	1.65997	0.89901	2.42091	<i>F</i>
<i>edu1</i>	0.76452	0.62102	0.90802	<i>F</i>
<i>cit1</i>	0.33312	0.17448	0.49176	<i>F</i>
<i>lab2</i>	1.86921	1.25457	2.48385	<i>F</i>

Tabla 6.3. Intervalos de confianza para $\alpha = 0$.

Variabes	$\hat{\beta}_{2j}$	$\hat{\beta}_{2j} - \sqrt{q}z_{\alpha/2}$	$\hat{\beta}_{2j} + \sqrt{q}z_{\alpha/2}$	$0 \in IC$
<i>constante</i>	-0.38295	-0.51819	-0.24771	<i>F</i>
<i>edu0</i>	0.99395	0.81769	1.17022	<i>F</i>
<i>edu1</i>	0.34896	0.21953	0.47839	<i>F</i>
<i>edu2</i>	0.17409	0.00280	0.34537	<i>F</i>
<i>cit1</i>	0.15597	0.07971	0.23222	<i>F</i>
<i>lab1</i>	-0.06853	-0.11539	-0.02168	<i>F</i>

Tabla 6.4. Intervalos de confianza para $\alpha = 1$.

En las tablas 6.1-6.4 se observa, viendo los *p*-valores o que el 0 no pertenece a ninguno de los intervalos de confianza, que los parámetros de regresión son significativos. Por otra parte, observando la magnitud y

los signos de los mismos, se deduce que al aumentar la población formada por los individuos con nivel de estudios primarios o inferiores aumenta el nivel de pobreza, siendo el aumento considerablemente mayor en el caso de nivel de estudios inferior a la educación primaria. También se observa que al aumentar la población de empleados (*lab1*) la brecha de pobreza disminuye y que al aumentar la población que se encuentra en paro (*lab2*) aumenta de forma considerable la proporción de pobreza.

En la figura 6.1 se muestran las gráficas para los pares $(y_{dr}, y_{dr} - x_{dr}\hat{\beta}_r - \hat{u}_{dr})$. En la gráfica de la izquierda se representan los residuos asociados a la proporción de pobreza ($\alpha = 0, r = 1$) y en la gráfica de la derecha se representan los residuos correspondientes a la brecha de pobreza ($\alpha = 1, r = 2$). La dispersión de ambas gráficas no parece decir nada en contra de la hipótesis de inesgadez del modelo ajustado. Asimismo, en la parte derecha de ambas gráficas se observa una tendencia a presentar residuos positivos que se corresponden con los valores mayores de los estimadores directos. Este hecho se considera una propiedad interesante, ya que significa que el modelo ajustado tiende a disminuir aquellos valores del estimador directo que superan cierto nivel, y con ello se consigue evitar la presencia de estimaciones extremas.

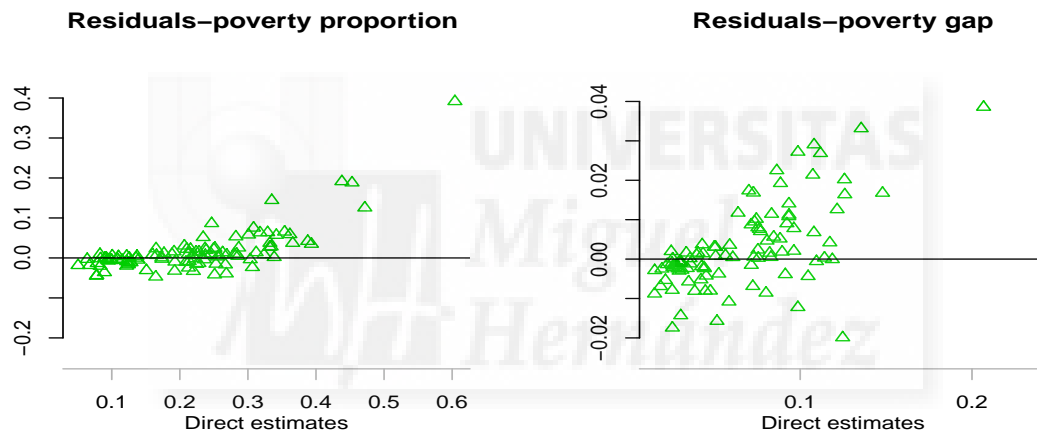


Figura 6.1: Residuos frente a estimadores directos.

Se tienen en cuenta dos estimadores para las dos variables que son objeto de estudio (proporción de pobreza y brecha de pobreza): estimador directo y EBLUP1, donde el estimador directo es conocido y el EBLUP1 se obtiene a partir del modelo multivariante que se ha ajustado. En la figura 6.2 se presentan las gráficas de los dos estimadores para la variable proporción de pobreza; en la parte izquierda se contemplan los dominios que corresponden a los hombres, y, en la parte derecha los que corresponden a las mujeres. En la figura 6.3 se presenta lo mismo para la variable brecha de pobreza. De forma análoga se presenta en las gráficas 6.4 y 6.5 la información que corresponde a la raíz cuadrada de la varianza del estimador directo y del error cuadrático medio del estimador EBLUP1.

En las gráficas 6.2 y 6.3 se puede apreciar con claridad que el estimador directo alcanza valores superiores respecto del estimador EBLUP1 para tamaños muestrales inferiores, a medida que aumentan éstos la diferencia tiende a reducirse. Por otra parte, en las gráficas que corresponden a la raíz cuadrada de la varianza de los dos estimadores (gráficas 6.4 y 6.5), se observan mayores valores en lo que respecta al estimador directo, la diferencia es más acusada para tamaños muestrales inferiores.

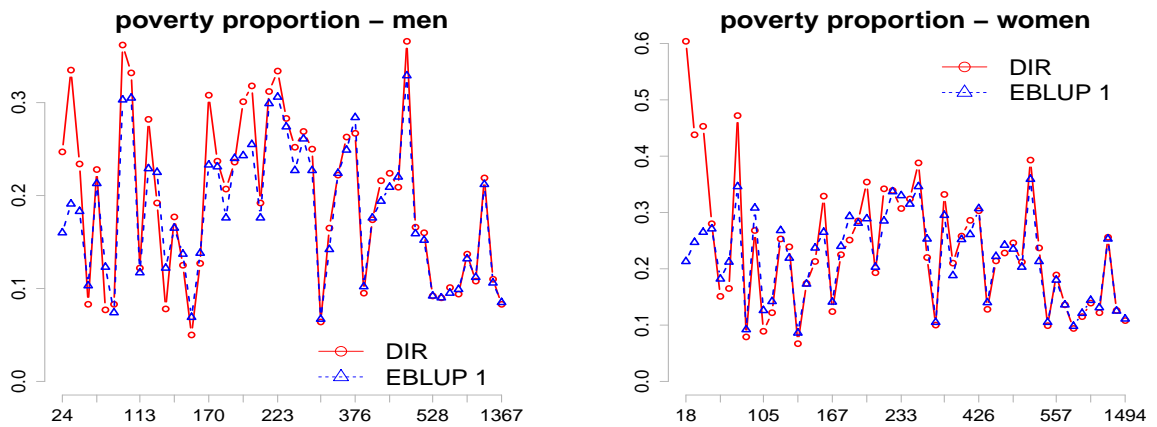


Figura 6.2: Estimaciones EBLUP1 y DIR de proporciones de pobreza por provincias en 2006.

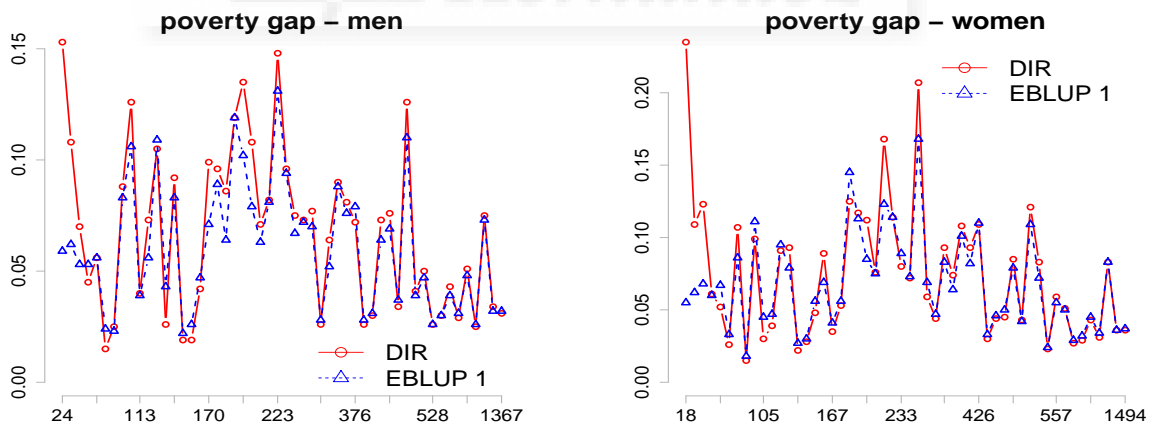


Figura 6.3: Estimaciones EBLUP1 y DIR de brechas de pobreza por provincias en 2006.

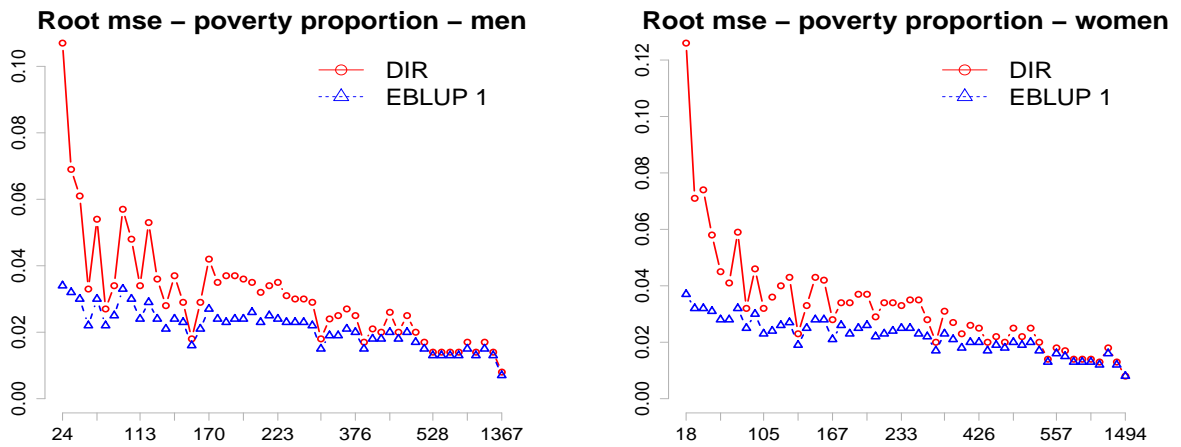


Figura 6.4: RMSEs de estimadores EBLUP1 y DIR de proporciones de pobreza por provincias en 2006.

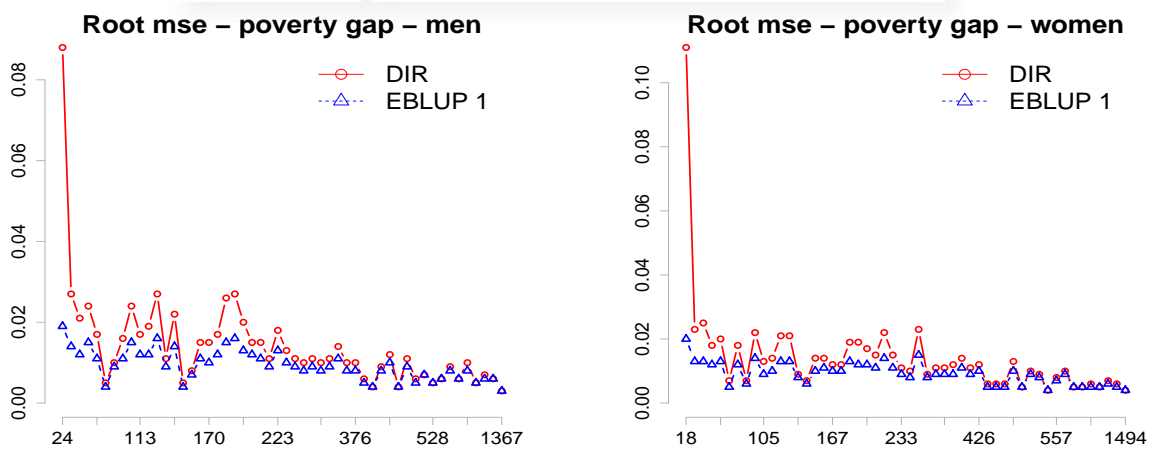


Figura 6.5: RMSEs de estimadores EBLUP1 y DIR de proporciones de pobreza por provincias en 2006.

6.2.2. Modelo con varianza de los efectos AR(1)

En esta sección se consideran las variables explicativas constante, age1, age2, edu1, cit1 y lab2 para la variable y_{d1} y las variables explicativas constante, edu0, edu1, edu2, cit1 y lab1 para la variable y_{d2} . Se supone que los efectos que corresponden a los distintos dominios se distribuyen según un proceso estocástico AR(1). Por tanto

$$u_{dr} = \rho u_{dr-1} + a_{dr}, \quad r = 1, 2, \quad u_{d0} \sim N(0, 1),$$

$$u \sim N(0, V_u), \quad V_u = \text{diag}_{1 \leq d \leq 104} (V_{ud}), \quad V_{ud} = \sigma_u^2 \frac{1}{1 - \rho^2} \begin{pmatrix} 1 & \rho \\ \rho & 1 \end{pmatrix}.$$

Con el objeto de obtener una estimación inicial de los parámetros σ_u^2 y ρ se considerarán los modelos univariantes que se deducen a partir del modelo multivariante planteado en la sección 6.1; es decir,

$$y_{dr} = x_{dr} \beta_r + u_{dr} + e_{dr}, \quad d = 1, \dots, D, \quad r = 1, 2,$$

donde

$$u_{.r} \sim N_D \left(0, \frac{\sigma_u^2}{1 - \rho^2} I_D \right), \quad e_{.r} \sim N_D(0, W_r^{-1}), \quad W_r^{-1} = \text{diag}_{1 \leq d \leq D} (\sigma_{edr}^{-2}).$$

A partir de ellos se determina la estimación utilizando la fórmula del método de Henderson 3

$$\hat{\sigma}_{urH}^2 = \frac{y_{.r}^t P_2 y_{.r} - D + p_r}{\text{tr}\{P_2\}}, \quad r = 1, 2,$$

y así las estimaciones que resultan se utilizan como semillas en el algoritmo de Fisher-scoring en la forma

$$\hat{\sigma}_u^{2(0)} = \frac{\hat{\sigma}_{u1H}^2 + \hat{\sigma}_{u2H}^2}{2} \quad \text{y} \quad \hat{\rho}^{(0)} = 0.$$

A partir de los datos de la muestra se obtienen las estimaciones iniciales

$$\hat{\sigma}_u^{2(0)} = 0,002213 \quad \text{y} \quad \hat{\rho}^{(0)} = 0.$$

Partiendo de los valores anteriores se ejecuta el algoritmo de Fisher-scoring para obtener las estimaciones REML de σ_u^2 y ρ . Después de trece iteraciones se tiene

$$\hat{\sigma}_u^2 = 0,00051 \quad \text{y} \quad \hat{\rho} = 0,53342.$$

El siguiente paso es calcular las estimaciones y los intervalos de confianza para los parámetros β_{rj} , $r = 1, 2$. Utilizando las distribuciones asintóticas de los estimadores $\hat{\beta}_{rj}$, se calculan los p -valores correspondientes a los contrastes $H_0 : \beta_{rj} = 0$.

Variables	constante	age1	age2	edu1	cit1	lab2
β	-0.75534	1.07209	1.53816	0.74652	0.34529	1.89741
p -valor	0.00000	0.00000	0.00000	0.00000	0.00001	0.00000

Tabla 6.5. Parámetros de regresión y p -valores para $\alpha = 0$.

Variabes	constante	edu0	edu1	edu2	cit1	lab1
β	-0.34779	1.008	0.32081	0.13462	0.14926	-0.07925
<i>p</i> -valor	0.00041	0.00000	0.00051	0.25762	0.01287	0.01795

Tabla 6.6. Parámetros de regresión y *p*-valores para $\alpha = 1$.

A continuación se construyen los intervalos de confianza correspondientes, donde el nivel de significación considerado es $\alpha = 0,1$. En este caso adoptan la forma $\hat{\beta}_{rj} \pm \sqrt{q}z_{\alpha/2}$, donde q es el elemento de la diagonal principal de la matriz Q , definida en la sección 2.4, que se corresponde con $\hat{\beta}_{rj}$. En la última columna de la tabla se incluye 'V' o 'F' según 0 pertenezca al intervalo de confianza o no. Lo anterior se resume en las tablas 6.7 y 6.8.

Variabes	$\hat{\beta}_{rj}$	$\hat{\beta}_{rj} - \sqrt{q}z_{\alpha/2}$	$\hat{\beta}_{rj} + \sqrt{q}z_{\alpha/2}$	$0 \in IC$
constante	-0.75534	-0.89874	-0.61195	F
age1	1.07209	0.70708	1.43709	F
age2	1.53816	0.98879	2.08752	F
edu1	0.74652	0.62631	0.86672	F
cit1	0.34529	0.21623	0.47435	F
lab2	1.89741	1.42702	2.36781	F

Tabla 6.7. Intervalos de confianza para $\alpha = 0$.

Variabes	$\hat{\beta}_{rj}$	$\hat{\beta}_{rj} - \sqrt{q}z_{\alpha/2}$	$\hat{\beta}_{rj} + \sqrt{q}z_{\alpha/2}$	$0 \in IC$
constante	-0.34779	-0.50964	-0.18595	F
edu0	1.00800	0.78511	1.23089	F
edu1	0.32081	0.16892	0.47271	F
edu2	0.13462	-0.06099	0.33023	V
cit1	0.14926	0.05056	0.24796	F
lab1	-0.07925	-0.13433	-0.02417	F

Tabla 6.8. Intervalos de confianza para $\alpha = 1$.

En las tablas 6.5-6.6 se observa que todos los *p*-valores son inferiores a 0.05, excepto en el caso de β_{24} . En las tablas 6.7-6.8 se observa que el 0 no pertenece a ninguno de los intervalos de confianza, excepto en el caso de β_{24} . Por tanto se concluye que todas las variables explicativas son significativas, salvo *edu2* para la brecha de pobreza. No obstante, se mantiene la variable en el modelo por el interés que tiene la comparación con el modelo multivariante diagonal de la sección 6.2.1. Por otra parte, observando la magnitud y los signos de los mismos, se deduce que al aumentar la población formada por los individuos con nivel de estudios primarios o inferiores aumenta el nivel de pobreza siendo el aumento considerablemente mayor en el caso de nivel de estudios inferior a la educación primaria. También se observa que al aumentar la población empleada (*lab1*) la brecha de pobreza disminuye y aumenta la proporción de pobreza de forma considerable cuando aumenta la población que se encuentra en paro (*lab2*).

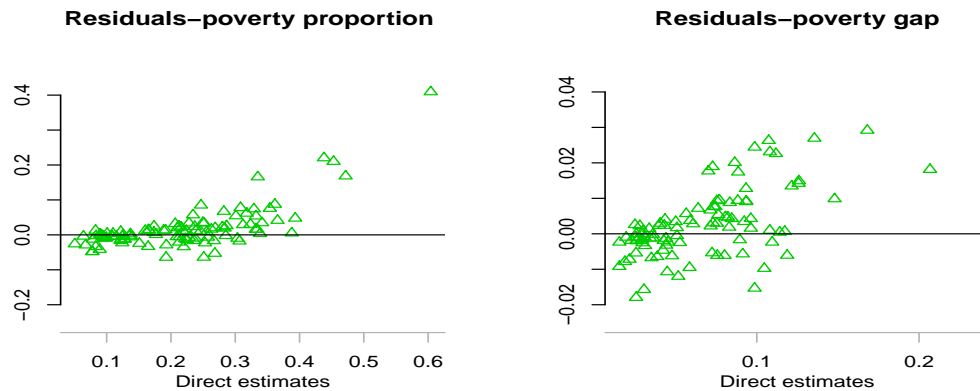


Figura 6.6: Residuos frente a estimadores directos.

En la figura 6.6 se muestran las gráficas para los pares $(y_{dr}, y_{dr} - x_{dr}\hat{\beta}_r - \hat{u}_{dr})$. En la gráfica de la izquierda se representan los residuos asociados a la proporción de pobreza ($\alpha = 0, r = 1$) y en la gráfica de la derecha se representan los residuos correspondientes a la brecha de pobreza ($\alpha = 1, r = 2$). La dispersión de ambas gráficas no parece decir nada en contra de la hipótesis de insesgaredad del modelo ajustado. Asimismo, en la parte derecha de ambas gráficas se observa una tendencia a presentar residuos positivos que se corresponden con los valores mayores de los estimadores directos. Este hecho se considera una propiedad interesante, ya que significa que el modelo ajustado tiende a disminuir aquellos valores que superan cierto nivel, y con ello se consigue evitar la presencia de outliers.

La disposición de los gráficos que se presentan en lo que sigue es análoga a la dispuesta en la sección 6.2.1

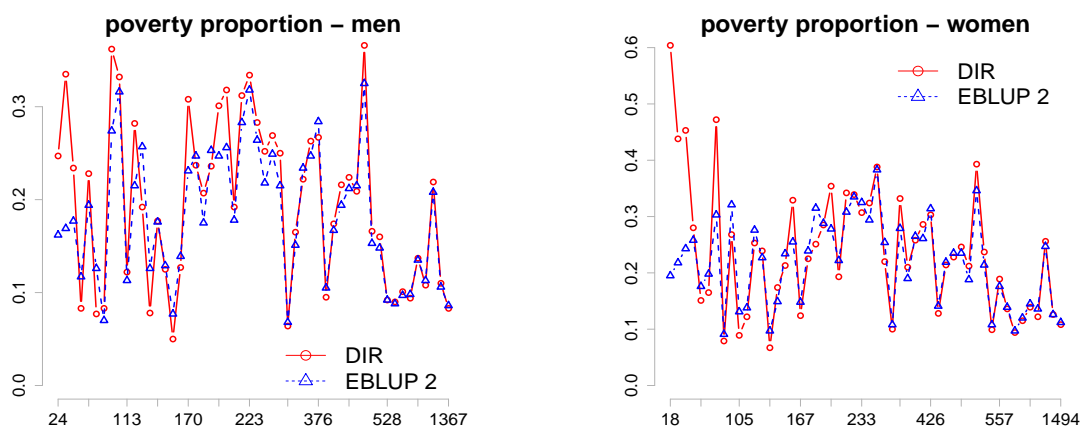


Figura 6.7: Estimaciones EBLUP2 y DIR de proporciones de pobreza por provincias en 2006.

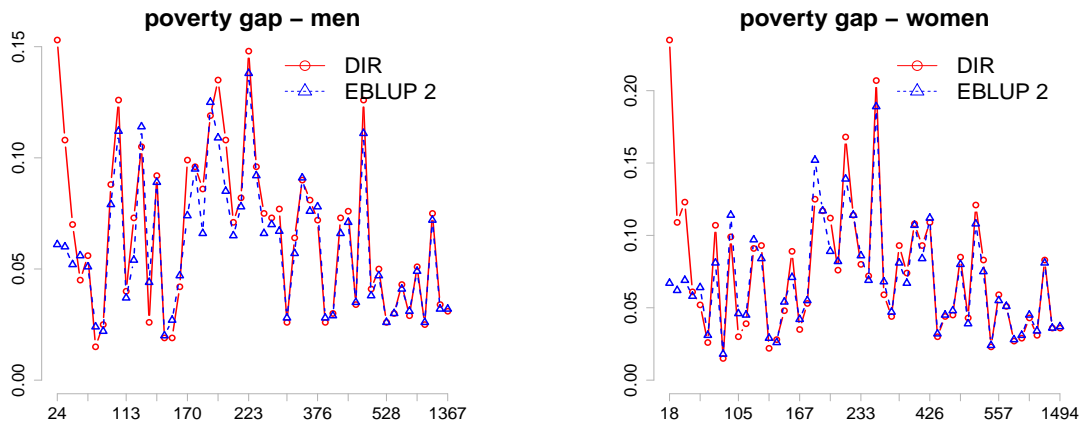


Figura 6.8: Estimaciones EBLUP2 y DIR de brechas de pobreza por provincias en 2006.

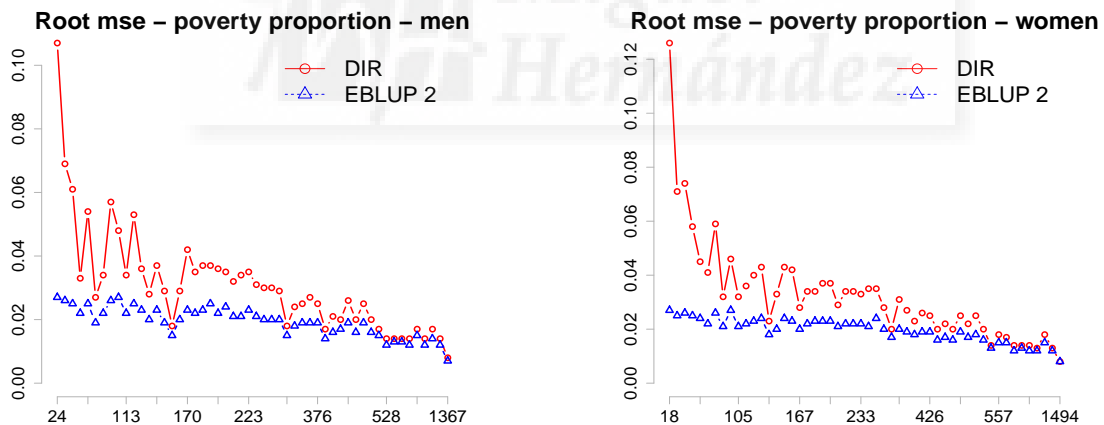


Figura 6.9: RMSEs de estimadores EBLUP2 y DIR de proporciones de pobreza por provincias en 2006.

En las gráficas 6.7 y 6.8 se puede apreciar con claridad que el estimador directo alcanza valores superiores respecto del estimador EBLUP2 para tamaños muestrales inferiores, a medida que aumentan éstos la diferencia tiende a reducirse. Por otra parte, en las gráficas que corresponden a la raíz cuadrada de la varianza de los dos estimadores (gráficas 6.9 y 6.10), se observan mayores valores en lo que respecta al estimador directo, la diferencia es más acusada para tamaños muestrales inferiores.

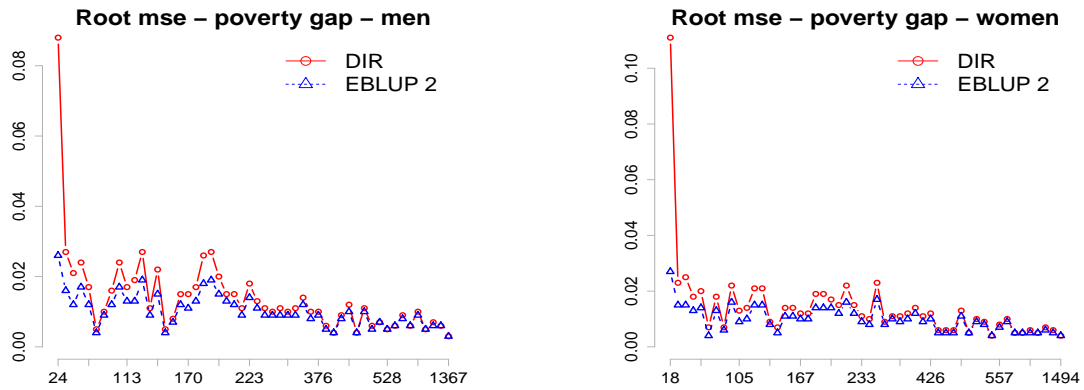


Figura 6.10: RMSEs de estimadores EBLUP2 y DIR de proporciones de pobreza por provincias en 2006.

6.2.3. Modelo con varianza de los efectos AR(1) heterocedástico

En esta sección se consideran las variables explicativas constante, age1, age2, edu1, cit1 y lab2 para la variable y_{d1} y las variables explicativas constante, edu0, edu1, edu2, cit1 y lab1 para la variable y_{d2} . Se supone que los efectos que corresponden a los distintos dominios se distribuyen según un proceso estocástico AR(1) heterocedástico. Por tanto

$$u_{dr} = \rho u_{dr-1} + a_{dr}, \quad u_{d0} \sim N(0, 1), \quad a_{dr} \sim N(0, \sigma_r^2), \quad r = 1, 2.$$

Se supone que a_{d1} y a_{d2} son independientes y que

$$u \sim N(0, V_u), \quad V_u = \text{diag}_{1 \leq d \leq 104} (V_{ud}), \quad \text{donde} \quad V_{ud} = \begin{pmatrix} \sigma_1^2 + \rho^2 & \rho\sigma_1^2 + \rho^3 \\ \rho\sigma_1^2 + \rho^3 & \sigma_2^2 + \rho^2\sigma_1^2 + \rho^4 \end{pmatrix}.$$

Con el objeto de obtener una estimación inicial para los parámetros σ_1^2 , σ_2^2 y ρ , se considerarán los modelos univariantes que se deducen a partir del modelo multivariante planteado en la sección 6.1; es decir,

$$y_{dr} = x_{dr}\beta_r + u_{dr} + e_{dr}, \quad d = 1, \dots, D, \quad r = 1, 2,$$

donde

$$u_{\cdot r} \sim N_D(0, \sigma_{ur}^2 I_D), \quad e_{\cdot r} \sim N_D(0, W_r^{-1}), \quad W_r^{-1} = \text{diag}_{1 \leq d \leq D} (\sigma_{edr}^{-2}).$$

A partir de ellos se estiman las varianzas aplicando la fórmula del método de Henderson 3; es decir,

$$\hat{\sigma}_{urH}^2 = \frac{y_{\cdot r}' P_2 y_{\cdot r} - D + p_r}{\text{tr}\{P_2\}}, \quad r = 1, 2.$$

Se obtiene

$$\hat{\sigma}_{u1H}^2 = 0,001507583 \quad \text{y} \quad \hat{\sigma}_{u2H}^2 = 0,002920324.$$

En las primeras pruebas que se han realizado para estimar los parámetros por el método de Fisher-scoring los resultados obtenidos no han sido satisfactorios debido a que, a partir de distintas semillas consideradas, siempre se obtiene una estimación para σ_u^2 negativa. Por ello, se sospecha que el *modelo AR(1) heterocedástico* no se ajusta bien a los datos. De todas formas, se propone una alternativa al ajuste directo que consiste en distinguir la estimación en dos fases que se detallan a continuación.

1. Se estiman los parámetros σ_{u1}^2 y σ_{u2}^2 utilizando el *modelo diagonal* considerando como semillas los valores obtenidos por el método Henderson 3.
2. Las estimaciones $\hat{\sigma}_{u1}^2$ y $\hat{\sigma}_{u2}^2$ obtenidas en la Fase 1 se consideran como verdaderos valores de los parámetros σ_{u1}^2 y σ_{u2}^2 y así se procede a estimar el parámetro ρ por Fisher-scoring tomando como semilla para el mismo $\rho^{(0)} = 0$.

Para realizar la primera parte del procedimiento que se propone se aprovecha lo que se ha realizado en el apartado 6.2.1 y se apunta que las estimaciones para σ_{u1}^2 y σ_{u2}^2 son

$$\hat{\sigma}_{u1}^2 = 0,00138 \quad \text{y} \quad \hat{\sigma}_{u2}^2 = 0,00037.$$

A partir de las estimaciones anteriores se tiene que la matriz de varianzas de los efectos es

$$V_{ud} = \begin{pmatrix} \sigma_1^2 + \rho^2 & \rho\sigma_1^2 + \rho^3 \\ \rho\sigma_1^2 + \rho^3 & \sigma_2^2 + \rho^2\sigma_1^2 + \rho^4 \end{pmatrix} = \begin{pmatrix} 0,00138 + \rho^2 & 0,00138\rho + \rho^3 \\ 0,00138\rho + \rho^3 & 0,00037 + 0,00138\rho^2 + \rho^4 \end{pmatrix}.$$

Ahora, se procede a realizar la segunda fase del procedimiento propuesto y se tiene que el algoritmo Fisher-scoring converge. Tras veintitrés iteraciones se llega a la estimación $\hat{\rho} = 0,018588$.

El siguiente paso es calcular las estimaciones y los intervalos de confianza para los parámetros β_{rj} , $r = 1, 2$. Utilizando las distribuciones asintóticas de los estimadores $\hat{\beta}_{rj}$, se calculan los p -valores correspondientes a los contrastes $H_0 : \beta_{rj} = 0$.

Variables	constante	age1	age2	edu1	cit1	lab2
β	-0.70357	0.95489	1.45541	0.74745	0.30873	1.50049
p -valor	0.00000	0.00066	0.00165	0.00000	0.00136	0.00006

Tabla 6.9. Parámetros de regresión y p -valores para $\alpha = 0$.

Variables	constante	edu0	edu1	edu2	cit1	lab1
β	-0.37458	0.97049	0.34255	0.16550	0.15203	-0.06384
p -valor	0.00001	0.00000	0.00001	0.11197	0.00104	0.02502

Tabla 6.10. Parámetros de regresión y p -valores para $\alpha = 1$.

A continuación se construyen los intervalos de confianza correspondientes, donde el nivel de significación considerado es $\alpha = 0,1$. Para el presente caso, adoptan la forma $\hat{\beta}_{rj} \pm \sqrt{q}z_{\alpha/2}$, donde q es el elemento de la diagonal principal de la matriz Q , definida en la sección 2.4, que se corresponde con $\hat{\beta}_{rj}$. En la última columna de la tabla se incluye 'V' o 'F' según 0 pertenezca al intervalo de confianza o no. Lo anterior se resume en las tablas 6.11 y 6.12.

Variabes	$\hat{\beta}_{rj}$	$\hat{\beta}_{rj} - \sqrt{q}z_{\alpha/2}$	$\hat{\beta}_{rj} + \sqrt{q}z_{\alpha/2}$	$0 \in IC$
<i>constante</i>	-0.70356	-0.88090	-0.52623	F
<i>age1</i>	0.95489	0.49342	1.41637	F
<i>age2</i>	1.45541	0.69448	2.21633	F
<i>edu1</i>	0.74745	0.60395	0.89094	F
<i>cit1</i>	0.30873	0.15009	0.46736	F
<i>lab2</i>	1.50049	0.88587	2.11512	F

Tabla 6.11. Intervalos de confianza para $\alpha = 0$.

Variabes	$\hat{\beta}_{rj}$	$\hat{\beta}_{rj} - \sqrt{q}z_{\alpha/2}$	$\hat{\beta}_{rj} + \sqrt{q}z_{\alpha/2}$	$0 \in IC$
<i>constante</i>	-0.37458	-0.50982	-0.23934	F
<i>edu0</i>	0.97049	0.79423	1.14675	F
<i>edu1</i>	0.34254	0.21312	0.47197	F
<i>edu2</i>	0.16550	-0.00577	0.33678	V
<i>cit1</i>	0.15203	0.07578	0.22828	F
<i>lab1</i>	-0.06384	-0.11069	-0.01698	F

Tabla 6.12. Intervalos de confianza para $\alpha = 1$.

En las tablas 6.9-6.10 se observa que todos los p -valores son inferiores a 0.1, excepto en el caso de β_{24} . En las tablas 6.7-6.8 se observa que el 0 no pertenece a ninguno de los intervalos de confianza, excepto también en el caso de β_{24} . Por tanto, se concluye que todas las variables explicativas son significativas, salvo *edu2* para la brecha de pobreza. No obstante, se mantiene la variable en el modelo por el interés que tiene la comparación con el modelo multivariante diagonal de la sección 6.2.1. Por otra parte, observando la magnitud y los signos de los mismos, se deduce que al aumentar la población formada por los individuos con nivel de estudios primarios o inferiores aumenta el nivel de pobreza siendo el aumento considerablemente mayor en el caso de nivel de estudios inferior a la educación primaria. También se observa que al aumentar la población de empleados (*lab1*) la brecha de pobreza disminuye y que al aumentar la población que se encuentra en paro (*lab2*) aumenta de forma considerable la proporción de pobreza.

La figura 6.11 muestra las gráficas para los pares $(y_{dr}, y_{dr} - x_{dr}\hat{\beta}_r - \hat{u}_{dr})$. En la gráfica de la izquierda se representan los residuos asociados a la proporción de pobreza ($\alpha = 0, r = 1$) y en la gráfica de la derecha se representan los residuos correspondientes a la brecha de pobreza ($\alpha = 1, r = 2$). La dispersión de ambas

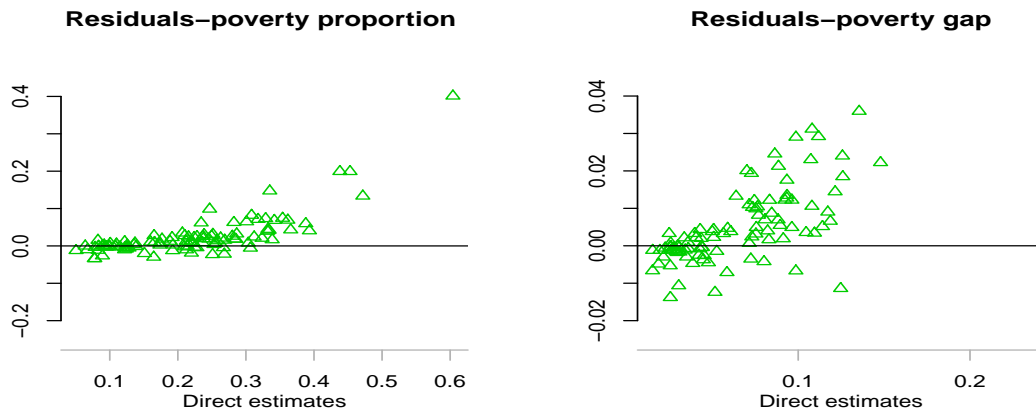


Figura 6.11: Residuos frente a estimadores directos.

gráficas no parece decir nada en contra de la hipótesis de insesgadez del modelo ajustado. Asimismo, en la parte derecha de ambas gráficas se observa una tendencia a presentar residuos positivos que se corresponden con los valores mayores de los estimadores directos. Este hecho se considera una propiedad interesante, ya que significa que el modelo ajustado tiende a disminuir aquellos valores que superan cierto nivel, y con ello se consigue evitar la presencia de outliers.

La figura 6.12 presenta los diagramas de cajas de los residuos $y_{dr} - x_{dr}\hat{\beta}_r - \hat{u}_{dr}$ estandarizados. Los diagramas muestran los residuos positivos que se observan en la Figura 6.11 para valores grandes de y_d donde el ajuste del modelo es peor. El modelo multivariante seleccionado presenta un mejor ajuste para la primera variable (proporción de pobreza) que para la segunda (brecha de pobreza). La disposición de las figuras 6.13-6.16 es análoga a la dispuesta en las secciones anteriores.

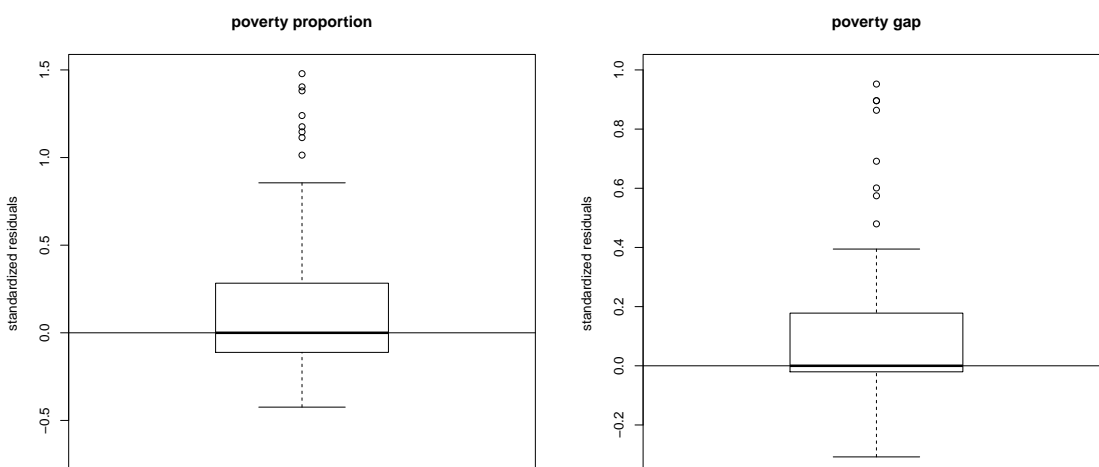


Figura 6.12: Diagramas de cajas de residuos estandarizados.

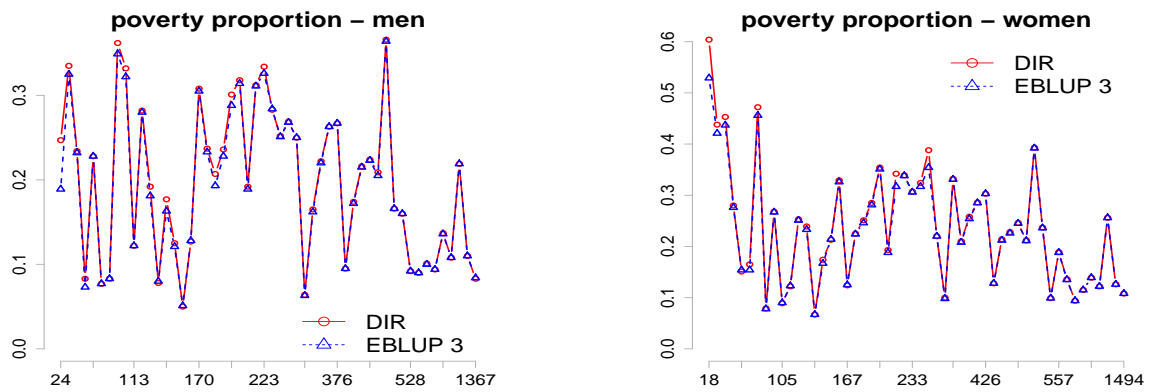


Figura 6.13: Estimaciones EBLUP3 y DIR de proporciones de pobreza por provincias en 2006.

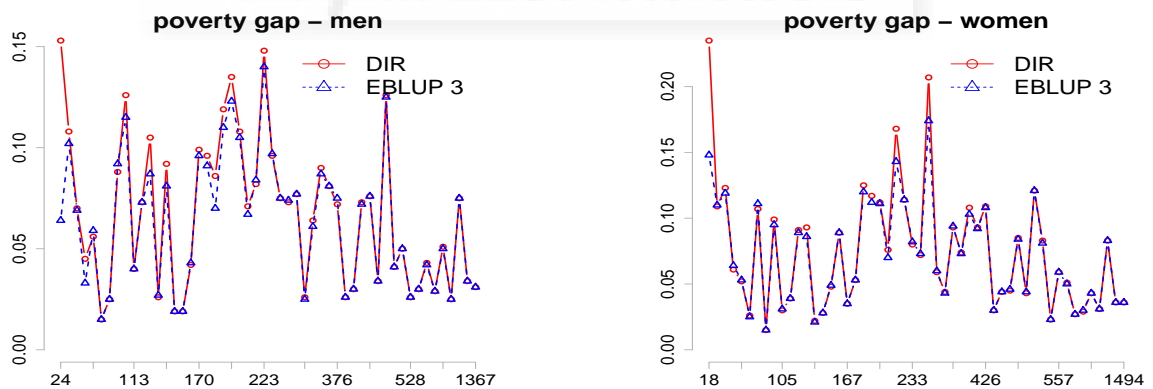


Figura 6.14: Estimaciones EBLUP3 y DIR de brechas de pobreza por provincias en 2006.

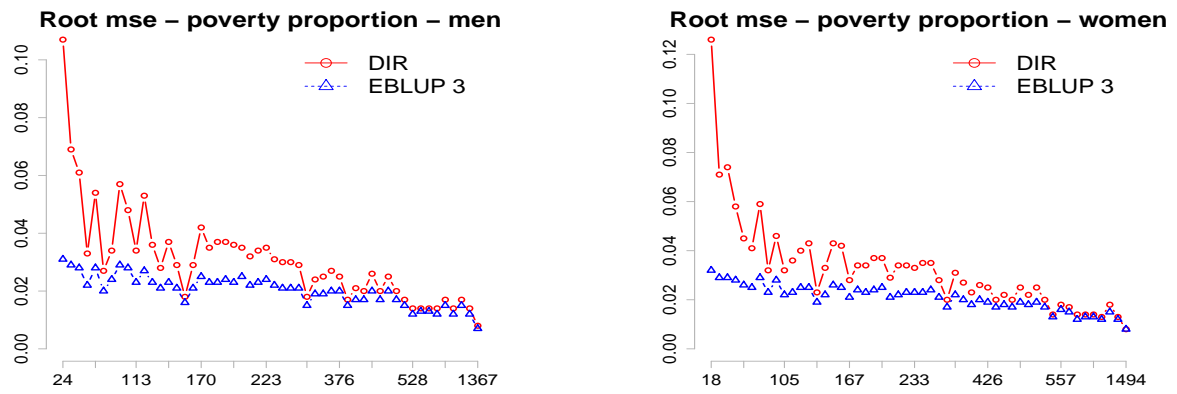


Figura 6.15: RMSEs de estimadores EBLUP3 y DIR de proporciones de pobreza por provincias en 2006.

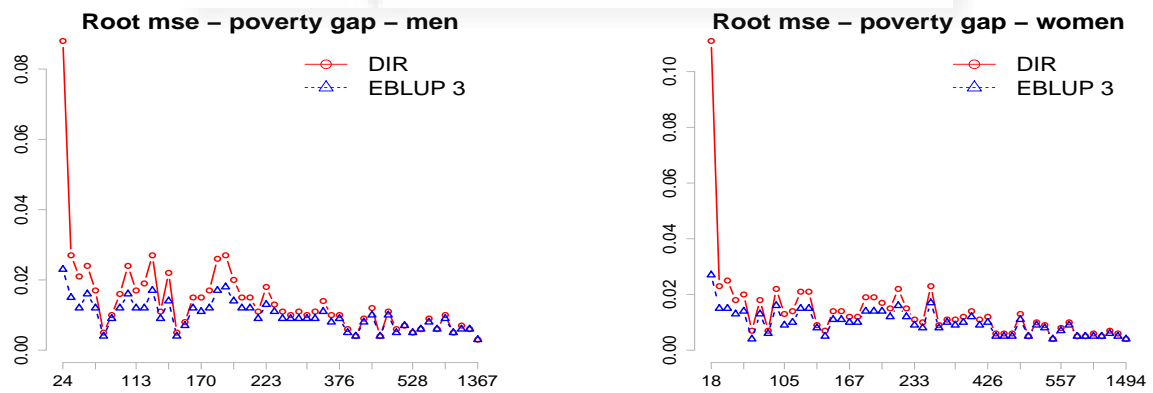


Figura 6.16: RMSEs de estimadores EBLUP3 y DIR de proporciones de pobreza por provincias en 2006.

En las gráficas 6.12 y 6.13 se puede apreciar con claridad que el estimador directo alcanza valores superiores respecto del estimador EBLUP3 para tamaños muestrales inferiores, a medida que aumentan éstos la diferencia tiende a reducirse. Por otra parte, en las gráficas que corresponden a la raíz cuadrada de la varianza de los dos estimadores (gráficas 6.14 y 6.15), se observan mayores valores en lo que respecta al estimador directo, la diferencia es más acusada para tamaños muestrales inferiores.

Contrastamos la hipótesis $H_0 : \sigma_{u1}^2 = \sigma_{u2}^2$. El estadístico del contraste es

$$T_{12} = \frac{\hat{\sigma}_{u1}^2 - \hat{\sigma}_{u2}^2}{\sqrt{v_{11} + v_{22} - 2v_{12}}} = 3,34588,$$

donde v_{ij} , $i, j = 1, 2, 3$ son los elementos de la inversa de la matriz REML de información de Fisher del modelo AR(1) heteroscedástico evaluada en $\hat{\theta} = (\hat{\sigma}_1^2, \hat{\sigma}_2^2, \hat{\rho})$. Como T_{12} tiene distribución asintótica normal estándar bajo H_0 , el p -valor es 0.00082. Se concluye que las varianzas de los efectos aleatorios son distintas, de modo que se prefiere el modelo AR(1) heteroscedástico al modelo AR(1). Contrastamos también la hipótesis $H_0 : \rho = 0$. El estadístico del contraste es

$$T_{\rho} = \frac{\hat{\rho}}{\sqrt{v_{33}}} = 1,96464.$$

Como T_{ρ} tiene distribución asintótica normal estándar bajo H_0 , el p -valor es 0.049456. Entonces, se concluye que ambas componentes (proporción y brecha de pobreza) están positivamente correladas y preferimos el modelo AR(1) heteroscedástico al modelo diagonal.

6.2.4. Conclusiones

En los apartados 6.2.1, 6.2.2 y 6.2.3 se han aplicado los distintos modelos estudiados a los datos de la muestra y finalmente se ha seleccionado el modelo con varianza de los efectos AR(1) heteroscedástico. En este apartado, se presentan los resultados finales.

En la tabla 6.13. se presenta una clasificación de las provincias españolas en cuatro categorías según los valores del EBLUP del modelo estimado de la proporción de pobreza y de la brecha de pobreza, es decir, $p_d = 100 \cdot \hat{Y}_{0,d,2006}^{eblup2}$ y $g_d = 100 \cdot \hat{Y}_{1,d,2006}^{eblup2}$.

men	$p_d < 10$	1 7 8 17 19 20 22 31 39 44 48 50
	$10 < p_d < 20$	3 9 12 21 24 25 26 27 28 33 36 42 43 46 47
	$20 < p_d < 30$	2 10 11 13 15 18 23 29 30 32 34 35 38 40 41 45 52
	$p_d > 30$	4 5 6 14 16 37 49 51
women	$p_d < 10$	1 17 20 22 31 48
	$10 < p_d < 20$	3 7 8 9 12 19 24 28 33 39 43 44 46 50
	$20 < p_d < 30$	2 15 21 25 26 27 29 30 32 34 35 36 38 41 45 47 49 52
	$p_d > 30$	4 5 6 10 11 13 14 16 18 23 37 40 42 51
men	$g_d < 3$	1 7 17 19 20 22 31 33 36 39 43 48
	$3 < g_d < 6$	3 8 9 12 26 28 34 41 44 46 50
	$6 < g_d < 10$	2 10 11 13 14 15 16 21 23 24 25 27 29 30 32 35 37 38 40 42 45 47
	$g_d > 10$	4 5 6 18 49 51 52
women	$g_d < 3$	1 7 17 19 31 39 43 48
	$3 < g_d < 6$	3 8 9 12 20 22 26 27 28 32 33 36 41 44 45 46 50
	$6 < g_d < 10$	10 13 14 15 21 24 25 30 34 35 37 38 47 49
	$g_d > 10$	2 4 5 6 11 16 18 23 29 40 42 51 52

Tabla 6.13. Provincias españolas clasificadas por proporción (arriba) y brecha (abajo) de pobreza.

La figura 6.17 contiene mapas de España en los que las provincias se colorean según los niveles de proporción de pobreza y de brecha de pobreza definidos en la tabla 6.13. Se observa que la proporción de la población por debajo de la línea de pobreza es menor en las provincias del noreste como Cataluña, Aragón, Navarra, Catabria. Por otra parte, se observa que las provincias españolas con mayor proporción de pobreza se encuentran situadas en el centro y en el sur como Andalucía, Extremadura, Castilla la Mancha y Canarias. En una posición intermedia se encuentran algunas provincias españolas del centro norte como Galicia, algunas provincias de Castilla León, Madrid y Comunidad Valenciana.

La tabla 6.14 presenta las estimaciones de la proporción de pobreza bajo el *modelo AR(1) heterocedástico*. La primera columna contiene la provincia, las tres siguientes presentan el estimador directo, el EBLUP0 y el EBLUP3 para la subpoblación de hombres, las tres siguientes muestran lo mismo para la subpoblación mujeres. Las seis últimas columnas se disponen de la misma forma para presentar la raíz cuadrada de los errores cuadráticos medios. La tabla 6.15 presenta las estimaciones de la brecha de pobreza bajo el *modelo AR(1) heterocedástico*. La estructura de sus columnas es la misma que la de la tabla 6.14.

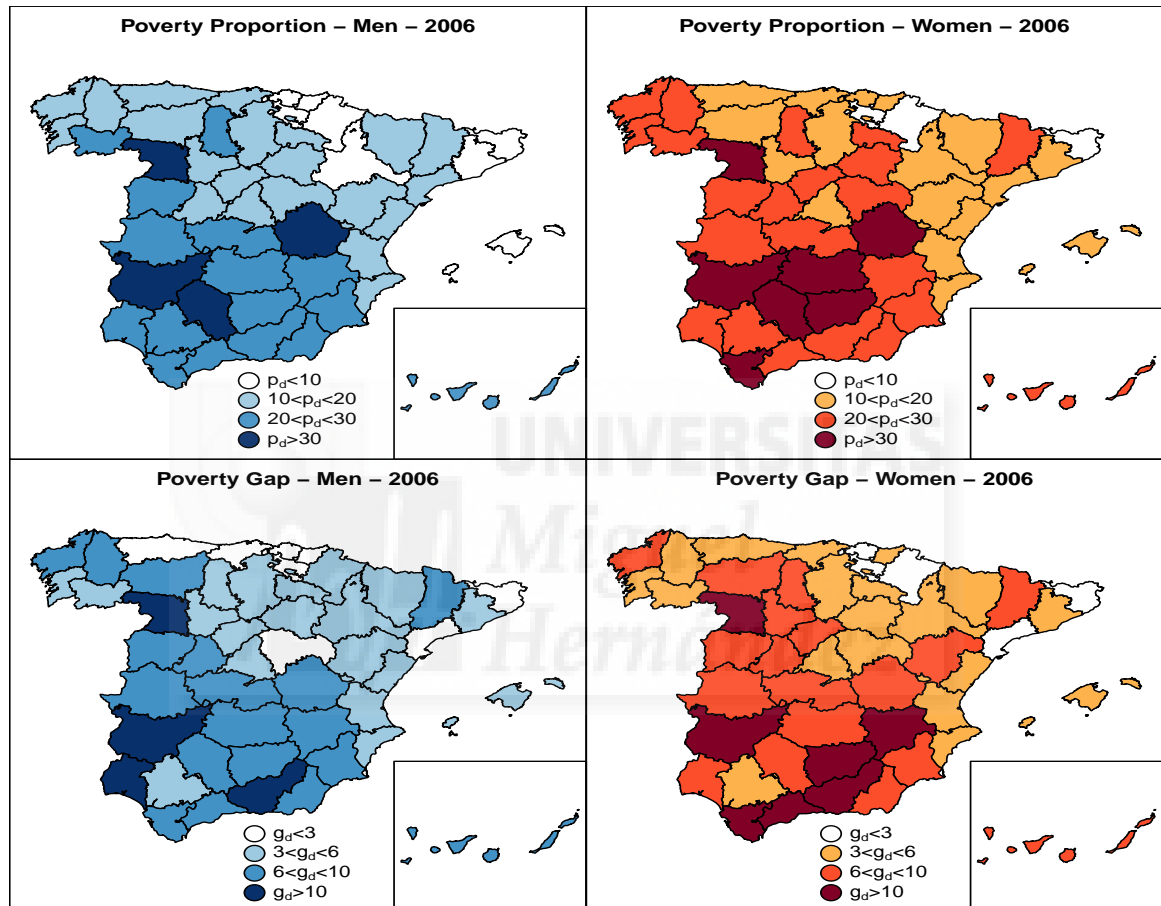


Figura 6.17: Estimaciones de las proporciones (arriba) y brechas (abajo) de pobreza para hombres (izquierda) y mujeres (derecha).

<i>d</i>	men/poverty proportions/women						men/sqrt.mse/women					
	dir	eb0	eb3	dir	eb0	eb3	dir	eb0	eb3	dir	eb0	eb3
1	0.083	0.074	0.083	0.079	0.087	0.078	0.034	0.028	0.024	0.032	0.027	0.023
2	0.237	0.246	0.233	0.285	0.289	0.281	0.035	0.028	0.023	0.037	0.029	0.024
3	0.160	0.155	0.160	0.189	0.184	0.188	0.017	0.016	0.015	0.018	0.017	0.016
4	0.318	0.286	0.314	0.354	0.313	0.351	0.035	0.029	0.025	0.037	0.030	0.025
5	0.335	0.202	0.325	0.453	0.289	0.437	0.069	0.039	0.029	0.074	0.039	0.029
6	0.366	0.346	0.364	0.393	0.372	0.392	0.025	0.022	0.020	0.025	0.022	0.019
7	0.094	0.096	0.094	0.115	0.118	0.115	0.014	0.013	0.012	0.014	0.014	0.013
8	0.083	0.084	0.084	0.108	0.110	0.108	0.008	0.007	0.007	0.008	0.008	0.008
9	0.127	0.134	0.128	0.124	0.139	0.125	0.029	0.025	0.021	0.028	0.024	0.021
10	0.252	0.235	0.251	0.332	0.306	0.331	0.030	0.025	0.021	0.031	0.026	0.022
11	0.267	0.278	0.267	0.303	0.311	0.303	0.025	0.022	0.020	0.025	0.022	0.019
12	0.122	0.117	0.122	0.122	0.135	0.123	0.034	0.028	0.023	0.036	0.028	0.023
13	0.269	0.261	0.268	0.324	0.315	0.317	0.030	0.025	0.021	0.035	0.028	0.023
14	0.312	0.299	0.311	0.307	0.323	0.306	0.034	0.027	0.023	0.033	0.027	0.023
15	0.216	0.206	0.215	0.237	0.226	0.236	0.020	0.019	0.017	0.020	0.018	0.017
16	0.362	0.307	0.349	0.472	0.358	0.456	0.057	0.037	0.029	0.059	0.037	0.029
17	0.050	0.063	0.051	0.067	0.083	0.067	0.018	0.017	0.016	0.023	0.021	0.019
18	0.301	0.267	0.288	0.342	0.326	0.317	0.036	0.029	0.023	0.034	0.028	0.022
19	0.077	0.105	0.077	0.165	0.198	0.154	0.027	0.023	0.020	0.041	0.031	0.025
20	0.064	0.065	0.063	0.100	0.103	0.098	0.018	0.017	0.015	0.020	0.019	0.017
21	0.192	0.237	0.181	0.253	0.273	0.251	0.036	0.029	0.023	0.040	0.030	0.025
22	0.078	0.105	0.080	0.089	0.115	0.090	0.028	0.024	0.021	0.032	0.027	0.022
23	0.283	0.273	0.284	0.339	0.341	0.338	0.031	0.026	0.022	0.034	0.027	0.023
24	0.192	0.185	0.189	0.193	0.213	0.188	0.032	0.026	0.022	0.029	0.025	0.021
25	0.177	0.176	0.163	0.239	0.235	0.233	0.037	0.030	0.023	0.043	0.032	0.025
26	0.166	0.161	0.166	0.212	0.204	0.211	0.020	0.018	0.017	0.022	0.020	0.018
27	0.207	0.188	0.193	0.225	0.241	0.224	0.037	0.029	0.023	0.034	0.028	0.024
28	0.110	0.108	0.110	0.126	0.126	0.126	0.014	0.013	0.012	0.013	0.013	0.012
29	0.222	0.230	0.220	0.258	0.263	0.254	0.025	0.022	0.019	0.023	0.021	0.018
30	0.219	0.215	0.219	0.256	0.253	0.256	0.017	0.016	0.015	0.018	0.017	0.015
31	0.090	0.089	0.090	0.094	0.096	0.094	0.014	0.014	0.013	0.014	0.013	0.012
32	0.282	0.241	0.280	0.213	0.235	0.214	0.053	0.035	0.027	0.043	0.032	0.026
33	0.108	0.111	0.108	0.122	0.128	0.122	0.014	0.013	0.012	0.013	0.013	0.012
34	0.228	0.209	0.228	0.280	0.276	0.276	0.054	0.035	0.028	0.058	0.036	0.028
35	0.224	0.219	0.223	0.246	0.242	0.245	0.026	0.023	0.020	0.025	0.022	0.019
36	0.174	0.173	0.172	0.214	0.220	0.212	0.021	0.019	0.017	0.022	0.020	0.018
37	0.308	0.259	0.305	0.329	0.284	0.326	0.042	0.031	0.025	0.042	0.031	0.025
38	0.263	0.256	0.263	0.286	0.276	0.285	0.027	0.023	0.020	0.026	0.023	0.020
39	0.095	0.099	0.095	0.128	0.136	0.128	0.017	0.016	0.015	0.020	0.018	0.017
40	0.234	0.195	0.232	0.438	0.265	0.421	0.061	0.037	0.028	0.071	0.038	0.029
41	0.209	0.215	0.205	0.228	0.235	0.226	0.020	0.018	0.017	0.020	0.019	0.017
42	0.247	0.174	0.189	0.604	0.229	0.529	0.107	0.042	0.031	0.126	0.043	0.032
43	0.125	0.132	0.121	0.174	0.167	0.167	0.029	0.025	0.021	0.033	0.027	0.022
44	0.083	0.108	0.073	0.151	0.170	0.154	0.033	0.028	0.022	0.045	0.033	0.026
45	0.250	0.232	0.250	0.220	0.244	0.220	0.029	0.025	0.021	0.028	0.024	0.021
46	0.137	0.136	0.136	0.139	0.142	0.139	0.017	0.016	0.015	0.014	0.013	0.013
47	0.165	0.158	0.162	0.210	0.200	0.208	0.024	0.021	0.019	0.027	0.023	0.020
48	0.092	0.092	0.092	0.099	0.103	0.099	0.014	0.013	0.012	0.014	0.014	0.013
49	0.332	0.329	0.322	0.268	0.312	0.267	0.048	0.035	0.028	0.046	0.035	0.028
50	0.101	0.099	0.100	0.136	0.137	0.135	0.014	0.014	0.013	0.017	0.016	0.015
51	0.334	0.323	0.326	0.388	0.389	0.354	0.035	0.029	0.024	0.035	0.029	0.024
52	0.236	0.245	0.228	0.251	0.291	0.246	0.037	0.030	0.024	0.034	0.028	0.023

Tabla 6.14. Estimaciones de la proporción de pobreza en 2006 y de sus raíces de errores cuadráticos medios.

<i>d</i>	men/poverty gaps/women						men/sqrt.mse/women					
	dir	eb0	eb3	dir	eb0	eb3	dir	eb0	eb3	dir	eb0	eb3
1	0.025	0.025	0.025	0.015	0.017	0.015	0.010	0.010	0.009	0.007	0.007	0.006
2	0.096	0.088	0.091	0.117	0.113	0.112	0.017	0.014	0.012	0.019	0.015	0.013
3	0.050	0.049	0.050	0.059	0.058	0.059	0.007	0.007	0.007	0.008	0.007	0.007
4	0.108	0.090	0.105	0.112	0.097	0.111	0.015	0.013	0.012	0.017	0.014	0.013
5	0.108	0.083	0.102	0.123	0.091	0.119	0.027	0.018	0.015	0.025	0.017	0.014
6	0.126	0.118	0.125	0.121	0.117	0.121	0.011	0.010	0.010	0.010	0.009	0.009
7	0.029	0.030	0.029	0.029	0.031	0.030	0.006	0.006	0.006	0.005	0.005	0.005
8	0.031	0.031	0.031	0.036	0.036	0.036	0.003	0.003	0.003	0.004	0.004	0.004
9	0.042	0.046	0.043	0.035	0.037	0.035	0.015	0.013	0.012	0.012	0.011	0.010
10	0.075	0.073	0.075	0.093	0.090	0.094	0.011	0.010	0.009	0.011	0.010	0.009
11	0.072	0.076	0.075	0.109	0.109	0.108	0.010	0.009	0.009	0.012	0.011	0.010
12	0.040	0.041	0.040	0.039	0.044	0.039	0.017	0.014	0.012	0.014	0.012	0.010
13	0.073	0.075	0.074	0.072	0.075	0.073	0.010	0.009	0.009	0.010	0.009	0.008
14	0.082	0.084	0.084	0.080	0.086	0.082	0.011	0.010	0.009	0.011	0.010	0.009
15	0.073	0.069	0.072	0.083	0.077	0.081	0.009	0.008	0.008	0.009	0.009	0.008
16	0.088	0.092	0.092	0.107	0.105	0.111	0.016	0.013	0.012	0.018	0.014	0.012
17	0.019	0.021	0.019	0.022	0.023	0.021	0.008	0.007	0.007	0.009	0.008	0.008
18	0.135	0.113	0.123	0.168	0.129	0.143	0.020	0.015	0.014	0.022	0.016	0.015
19	0.015	0.018	0.015	0.026	0.029	0.025	0.005	0.005	0.004	0.007	0.007	0.004
20	0.026	0.027	0.025	0.044	0.045	0.043	0.010	0.009	0.009	0.011	0.010	0.010
21	0.105	0.098	0.087	0.091	0.092	0.089	0.027	0.018	0.017	0.021	0.016	0.014
22	0.026	0.034	0.027	0.030	0.037	0.031	0.011	0.010	0.009	0.013	0.011	0.009
23	0.096	0.098	0.097	0.114	0.114	0.114	0.013	0.012	0.011	0.015	0.013	0.011
24	0.071	0.066	0.067	0.076	0.071	0.070	0.015	0.013	0.012	0.015	0.013	0.012
25	0.092	0.084	0.081	0.093	0.082	0.086	0.022	0.016	0.014	0.021	0.016	0.014
26	0.041	0.041	0.041	0.043	0.044	0.044	0.006	0.005	0.005	0.005	0.005	0.005
27	0.086	0.071	0.070	0.053	0.055	0.053	0.026	0.018	0.017	0.012	0.011	0.010
28	0.034	0.033	0.034	0.036	0.036	0.036	0.006	0.006	0.006	0.006	0.006	0.005
29	0.090	0.086	0.087	0.108	0.101	0.103	0.014	0.012	0.011	0.014	0.012	0.011
30	0.075	0.075	0.075	0.083	0.084	0.083	0.007	0.006	0.006	0.007	0.006	0.006
31	0.030	0.030	0.030	0.027	0.028	0.027	0.006	0.006	0.006	0.005	0.005	0.005
32	0.073	0.066	0.073	0.048	0.053	0.049	0.019	0.015	0.012	0.014	0.012	0.011
33	0.025	0.026	0.025	0.031	0.032	0.031	0.005	0.005	0.005	0.005	0.005	0.005
34	0.056	0.061	0.059	0.061	0.064	0.064	0.017	0.014	0.012	0.018	0.014	0.012
35	0.076	0.072	0.076	0.085	0.082	0.084	0.012	0.011	0.010	0.013	0.011	0.011
36	0.030	0.031	0.030	0.044	0.045	0.044	0.004	0.004	0.004	0.006	0.005	0.005
37	0.099	0.084	0.096	0.089	0.079	0.089	0.015	0.013	0.011	0.014	0.012	0.011
38	0.081	0.079	0.081	0.093	0.088	0.092	0.010	0.009	0.008	0.011	0.010	0.009
39	0.026	0.027	0.026	0.030	0.031	0.030	0.006	0.005	0.005	0.006	0.006	0.005
40	0.070	0.061	0.069	0.109	0.086	0.110	0.021	0.016	0.012	0.023	0.017	0.014
41	0.034	0.035	0.034	0.045	0.048	0.046	0.004	0.004	0.004	0.006	0.005	0.005
42	0.153	0.064	0.064	0.235	0.064	0.148	0.088	0.023	0.023	0.111	0.023	0.023
43	0.019	0.021	0.019	0.028	0.031	0.028	0.005	0.005	0.004	0.007	0.006	0.005
44	0.045	0.046	0.033	0.052	0.064	0.053	0.024	0.017	0.016	0.020	0.015	0.013
45	0.077	0.076	0.077	0.059	0.064	0.060	0.011	0.010	0.009	0.009	0.009	0.008
46	0.051	0.049	0.050	0.043	0.043	0.043	0.010	0.009	0.009	0.006	0.005	0.005
47	0.064	0.056	0.061	0.074	0.069	0.073	0.011	0.010	0.009	0.012	0.011	0.010
48	0.026	0.026	0.026	0.023	0.023	0.023	0.005	0.005	0.005	0.004	0.004	0.004
49	0.126	0.109	0.115	0.099	0.104	0.095	0.024	0.018	0.016	0.022	0.017	0.015
50	0.043	0.041	0.042	0.051	0.050	0.050	0.009	0.008	0.008	0.010	0.009	0.009
51	0.148	0.135	0.140	0.207	0.167	0.174	0.018	0.015	0.013	0.023	0.017	0.016
52	0.119	0.114	0.110	0.125	0.131	0.120	0.027	0.019	0.018	0.019	0.015	0.013

Tabla 6.15. Estimaciones de la brecha de pobreza en 2006 y de sus raíces de errores cuadráticos medios.

6.3. Estimación basada en modelos de área (enfoque temporal)

De la forma en que ya se ha apuntado al inicio del capítulo se va a realizar un enfoque temporal del modelo de área estudiado en el capítulo 2. En esta sección se presenta una nueva aplicación del modelo multivariante de área (2.1) utilizando datos de la misma encuesta, solo que para el caso se utilizarán los datos que corresponden a los años 2005 y 2006. A diferencia del estudio realizado en la sección anterior 6, se considera un enfoque temporal; es decir, se estudia una sola variable objetivo en dos instantes diferentes de tiempo (años 2005 y 2006). Se considera oportuno destacar el papel que pueden realizar el *modelo AR(1)* y el *modelo AR(1) heterocedástico* para realizar estimaciones. No obstante, se estudia también para completar el estudio el *modelo diagonal*. Se presentan en cada una de las siguientes secciones las variables explicativas que se van a incluir en el modelo y las hipótesis sobre los efectos aleatorios del modelo que hay que añadir a lo descrito en en la sección 6.2.1. Se presentan también las estimaciones de los parámetros y las interpretaciones que resulten oportunas en cada caso.

En esta sección se consideran las variables explicativas *constante*, *age1*, *age2*, *edu1*, *cit1* y *lab2* para las variables y_{d1} e y_{d2} , que son las proporciones de pobreza en los años 2005 y 2006 respectivamente.

6.3.1. Modelo con varianza de los efectos diagonal

En esta sección se supone que los efectos aleatorios verifican

$$u \sim N(0, V_u), \quad V_u = \text{diag} (V_{ud}), \quad V_{ud} = \begin{pmatrix} \sigma_{u1}^2 & 0 \\ 0 & \sigma_{u2}^2 \end{pmatrix}.$$

Con el objeto de obtener una estimación inicial de los parámetros σ_{u1}^2 y σ_{u2}^2 se considerarán los modelos univariantes que se deducen a partir del modelo multivariante planteado en la sección 6.1; es decir,

$$y_{dr} = x_{dr}\beta_r + u_{dr} + e_{dr}, \quad d = 1, \dots, D, \quad r = 1, 2,$$

donde

$$u_{.r} \sim N_D(0, \sigma_{ur}^2 I_D), \quad e_{.r} \sim N_D(0, W_r^{-1}), \quad W_r^{-1} = \text{diag}(\sigma_{edr}^{-2}).$$

Se hace una primera estimación de las componentes de la varianza utilizando la fórmula del método de Henderson 3; es decir,

$$\hat{\sigma}_{wrH}^2 = \frac{y'_{.r} P_2 y_{.r} - D + p_r}{\text{tr}\{P_2\}}, \quad r = 1, 2.$$

Las estimaciones $\hat{\sigma}_{u1H}^2$ y $\hat{\sigma}_{u2H}^2$ se utilizan como semillas en el algoritmo de Fisher-scoring. A partir de los datos de la muestra, las estimaciones que se obtienen de las semillas son

$$\hat{\sigma}_{u1H}^2 = 0,001803664 \quad \text{y} \quad \hat{\sigma}_{u2H}^2 = 0,034170970.$$

Partiendo de los valores anteriores, se ejecuta el algoritmo Fisher-scoring para obtener las estimaciones REML de σ_{u1}^2 y σ_{u2}^2 . Después de siete iteraciones se obtiene

$$\hat{\sigma}_{u1}^2 = 0,002597476 \quad \text{y} \quad \hat{\sigma}_{u2}^2 = 0,001928316.$$

A continuación, se realiza el contraste de igualdad de varianzas $H_0 : \sigma_{u1}^2 = \sigma_{u2}^2$ al nivel de significación $\alpha = 0,1$. Para ello se utiliza la distribución asintótica de $\hat{\theta} = (\hat{\theta}_1 = \hat{\sigma}_{u1}^2, \hat{\theta}_2 = \hat{\sigma}_{u2}^2)$ presentada en el capítulo 2. A partir de la distribución asintótica se obtiene que

$$\hat{\sigma}_{u1}^2 - \hat{\sigma}_{u2}^2 \sim N(\sigma_{u1}^2 - \sigma_{u2}^2, v_{11} + v_{22} - 2v_{12}),$$

donde v_{ij} es el elemento situado en la fila i y la columna j en la matriz F^{-1} definida en la sección 2.3. Teniendo en cuenta lo anterior, y aplicándolo al caso que nos ocupa, se obtiene el estadístico del contraste

$$\frac{\hat{\sigma}_{u1}^2 - \hat{\sigma}_{u2}^2}{\sqrt{v_{11} + v_{22} - 2v_{12}}} = 1,0756$$

y el p -valor=0.2820787. Por tanto, teniendo en cuenta los datos observados y el nivel de significación que se ha fijado, se concluye que no hay evidencia suficiente para rechazar la hipótesis nula.

A partir de lo anterior se construyen las estimaciones para los parámetros β_{rj} , $r = 2$. Utilizando también las distribuciones asintóticas para los estimadores $\hat{\beta}_{rj}$ presentadas en el capítulo primero, se determinan los p -valores correspondientes a los contrastes $H_0 : \beta_{rj} = 0$. Todo ello se presenta en las tablas 6.17 y 6.18.

Variabes	constante	age1	age2	edu1	cit1	lab2
β	-0.65933	0.69445	2.42186	0.71191	0.25932	0.71777
p -valor	0.00003	0.05177	0.00017	0.00000	0.07097	0.12555

Tabla 6.16. Parámetros de regresión y p -valores para $\alpha = 0$ y $r = 1$ (año 2005).

Variabes	constante	age1	age2	edu1	cit1	lab2
β	-0.75221	0.88471	1.89549	0.79795	0.31376	2.05462
p -valor	0.00000	0.00609	0.00048	0.00000	0.00425	0.00000

Tabla 6.17. Parámetros de regresión y p -valores para $\alpha = 1$.

A continuación se construyen los intervalos de confianza correspondientes, donde el nivel de significación considerado es $\alpha = 0,1$. Para el presente caso, adoptan la forma $\hat{\beta}_{rj} \pm \sqrt{q}z_{\alpha/2}$, donde q es el elemento de la diagonal principal de la matriz Q , definida en la sección 2.4, que se corresponde con $\hat{\beta}_{rj}$. En la última columna de la tabla se incluye “V” o “F” según 0 pertenezca al intervalo de confianza o no. Lo anterior se resume en las tablas 6.18 y 6.19.

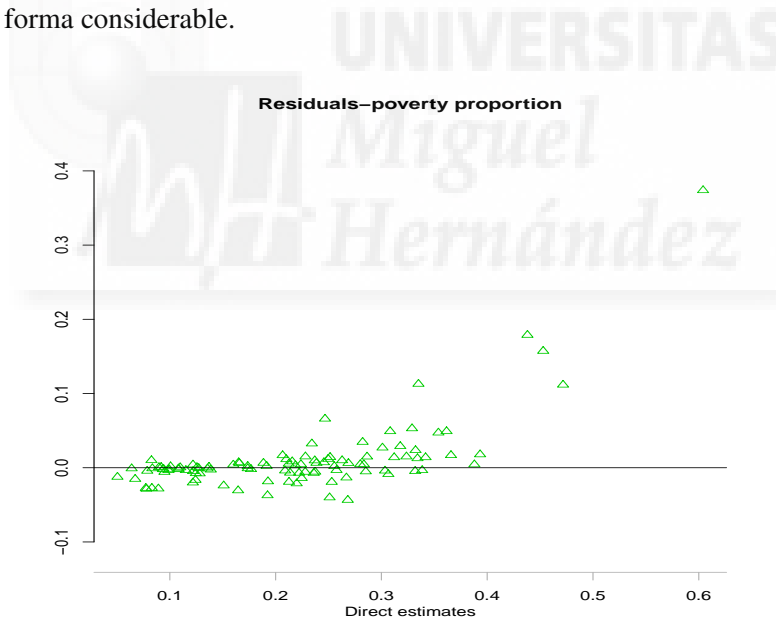
Variabes	$\hat{\beta}_{1j}$	$\hat{\beta}_{1j} - \sqrt{q}z_{\alpha/2}$	$\hat{\beta}_{1j} + \sqrt{q}z_{\alpha/2}$	$0 \in IC$
constante	-0.65933	-0.91968	-0.39897	F
age1	0.69445	0.10715	1.28174	F
age2	2.42186	1.36130	3.48241	F
edu1	0.71191	0.54378	0.88003	F
cit1	0.25932	0.02309	0.49554	F
lab2	0.71777	-0.05294	1.48849	V

Tabla 6.18. Intervalos de confianza para $\alpha = 0$ y $r = 1$ (año 2005).

Variabes	$\hat{\beta}_{1j}$	$\hat{\beta}_{1j} - \sqrt{q}z_{\alpha/2}$	$\hat{\beta}_{1j} + \sqrt{q}z_{\alpha/2}$	$0 \in IC$
constante	-0.75221	-0.95438	-0.55004	F
age1	0.88471	0.35409	1.41532	F
age2	1.89549	1.00278	2.78821	F
edu1	0.79794	0.63689	0.95900	F
cit1	0.31376	0.13325	0.49428	F
lab2	2.05462	1.34508	2.76417	F

Tabla 6.19. Intervalos de confianza para $\alpha = 0$ y $r = 2$ (año 2006).

En las tablas 6.16 y 6.17 se observa, viendo los p -valores, que los parámetros de regresión son significativos, excepto el que se corresponde con la variable *lab2* para $r = 1$. No obstante, se mantiene la variable en el modelo por el interés que tiene su interpretación. Por otra parte, se observa en las tablas 6.18 y 6.19 que el 0 no pertenece a ninguno de los intervalos de confianza, excepto al intervalo correspondiente a la variable *lab2* para $r = 1$. También se puede decir, observando la magnitud y los signos de los mismos, que al aumentar la población formada por los individuos con nivel de estudios primarios aumenta el nivel de pobreza. También se observa que al aumentar la población que se encuentra en situación de paro el nivel de pobreza aumenta de forma considerable.

Figura 6.18: Residuos frente a estimadores directos para $\alpha = 0$ y $r = 2$ (año 2006).

La figura 6.18 muestra la gráfica de los pares $(y_{dr}, y_{dr} - x_{dr}\hat{\beta}_r - \hat{u}_{dr})$, donde la variable objetivo es la proporción de pobreza en el año 2006 ($\alpha = 0$, $r = 2$). La dispersión del gráfico no parece decir nada en contra de la hipótesis de insesgadez del modelo ajustado. Asimismo, en la parte derecha de la gráfica se observa una tendencia a presentar residuos positivos que se corresponden con los valores mayores de los estimadores directos. Este hecho se considera una propiedad interesante, ya que significa que el modelo ajustado tiende a disminuir aquellos valores que superan cierto nivel, y con ello se consigue evitar la presencia de outliers.

Se tienen en cuenta dos estimadores para la variable proporción de pobreza en el año 2006 ($\alpha = 0, r = 2$): estimador directo y EBLUP1, donde el estimador directo es conocido. El EBLUP1 se obtiene a partir del modelo multivariante ajustado (*modelo diagonal*) para la variable proporción de pobreza en los años 2005 y 2006. La figura 6.18 presenta las estimaciones EBLUP1 y DIR de proporciones de pobreza por provincias en 2006. La figura 6.19 presenta las estimaciones de las raíces cuadradas de los errores cuadráticos medios de estimadores EBLUP1 y DIR de proporciones de pobreza por provincias en 2006.

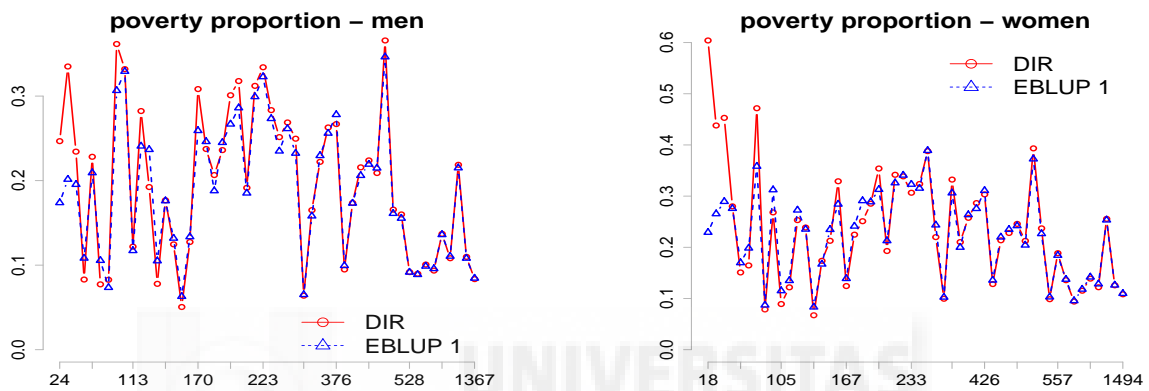


Figura 6.19: Estimaciones EBLUP1 y DIR de proporciones de pobreza por provincias en 2006.

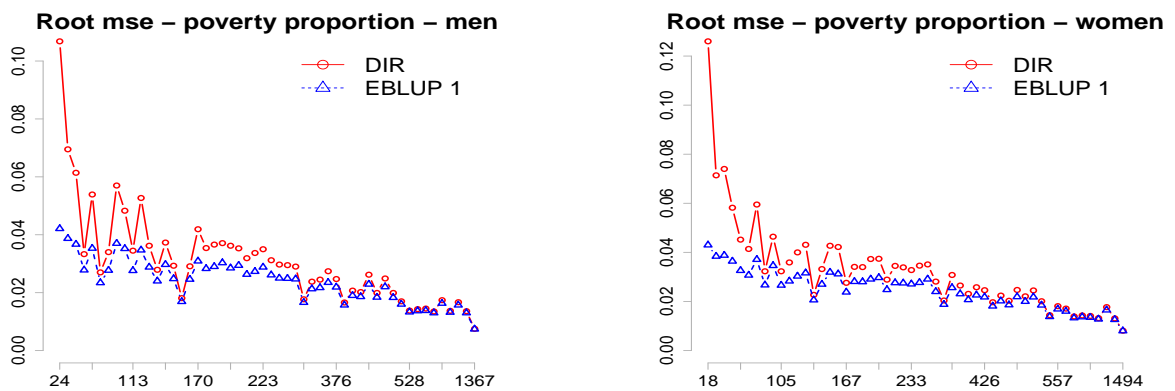


Figura 6.20: RMSEs de estimadores EBLUP1 y DIR de proporciones de pobreza por provincias en 2006.

En las gráfica 6.18 se puede apreciar con claridad que el estimador directo alcanza valores superiores respecto del estimador EBLUP1 para tamaños muestrales inferiores, a medida que aumentan éstos la diferencia tiende a reducirse. Por otra parte, en las gráficas que corresponden a la raíz cuadrada de la varianza de los dos estimadores (gráfica 6.19), se observan mayores valores en lo que respecta al estimador directo, la diferencia es más acusada para tamaños muestrales inferiores.

6.3.2. Modelo con varianza de los efectos AR(1)

Se supone que los efectos que corresponden a los distintos dominios se distribuyen según un proceso estocástico AR(1). Por tanto,

$$u_{dr} = \rho u_{dr-1} + a_{dr}, \quad r = 1, 2, \quad u_{d0} \sim N(0, 1),$$

$$u \sim N(0, V_u), \quad V_u = \text{diag}_{1 \leq d \leq 104} (V_{ud}), \quad V_{ud} = \sigma_u^2 \frac{1}{1 - \rho^2} \begin{pmatrix} 1 & \rho \\ \rho & 1 \end{pmatrix}.$$

Con el objeto de obtener una estimación inicial de los parámetros σ_u^2 y ρ se considerarán los modelos univariantes que se deducen a partir del modelo multivariante planteado en la sección 6.1; es decir,

$$y_{dr} = x_{dr} \beta_r + u_{dr} + e_{dr}, \quad d = 1, \dots, D, \quad r = 1, 2,$$

donde

$$u_{.r} \sim N_D \left(0, \frac{\sigma_u^2}{1 - \rho^2} I_D \right), \quad e_{.r} \sim N_D (0, W_r^{-1}), \quad W_r^{-1} = \text{diag}_{1 \leq d \leq D} (\sigma_{edr}^{-2}).$$

A partir de ellos se determina la estimación utilizando la fórmula del método de Henderson 3

$$\hat{\sigma}_{urH}^2 = \frac{y'_{.r} P_2 y_{.r} - D + p_r}{\text{tr}\{P_2\}}, \quad r = 1, 2,$$

y así las estimaciones que resultan se utilizan como semillas en el algoritmo de Fisher-scoring en la forma

$$\hat{\sigma}_u^{2(0)} = \frac{\hat{\sigma}_{u1H}^2 + \hat{\sigma}_{u2H}^2}{2} \quad \text{y} \quad \hat{\rho}^{(0)} = 0.$$

Las estimaciones que se obtienen de los datos son

$$\hat{\sigma}_{u1H}^2 = 0,001803664 \quad \text{y} \quad \hat{\sigma}_{u2H}^2 = 0,034132248.$$

En consecuencia, las estimaciones iniciales son

$$\hat{\sigma}_u^{2(0)} = 0,01796796 \quad \text{y} \quad \hat{\rho}^{(0)} = 0.$$

Partiendo de los valores anteriores se ejecuta el algoritmo de Fisher-scoring para obtener las estimaciones REML de σ_u^2 y ρ y se obtiene, después de doce iteraciones,

$$\hat{\sigma}_u^2 = 0,0006959078 \quad \text{y} \quad \hat{\rho} = 0,8608996891.$$

Para contrastar la hipótesis $H_0 : \rho = 0$ al nivel de significación $\alpha = 0,1$ se utiliza la distribución asintótica de $\hat{\theta} = (\hat{\theta}_1 = \hat{\sigma}_u^2, \hat{\theta}_2 = \hat{\rho})$, presentada en la sección 4.1. A partir de la distribución asintótica se obtiene que

$$\hat{\rho} \sim N(\rho, v_{22}),$$

donde v_{22} es el elemento situado en la fila 2 y la columna 2 en la matriz F^{-1} definida en la sección 2.3. Teniendo en cuenta lo anterior y aplicándolo al caso que nos ocupa, el estadístico del contraste es

$$\frac{\hat{\rho}}{\sqrt{v_{22}}} = 16,72633$$

y el p -valor es 0. Por tanto, teniendo en cuenta los datos observados y el nivel de significación que se ha fijado, se rechaza la hipótesis nula y preferimos usar el modelo AR(1), en lugar del modelo diagonal, para dar estimaciones EBLUP. El siguiente paso, en el estudio del modelo AR(1), es calcular las estimaciones y los intervalos de confianza para los parámetros β_{rj} , $r = 1, 2$. Utilizando las distribuciones asintóticas de los estimadores $\hat{\beta}_{rj}$, se calculan los p -valores correspondientes a los contrastes $H_0 : \beta_{rj} = 0$. Los resultados se presentan en las tablas 6.20 y 6.21.

Variabes	constante	age1	age2	edu1	cit1	lab2
β	-0.53822	0.67365	1.74785	0.60288	0.23672	0.99025
p -valor	0.00040	0.03876	0.00209	0.00000	0.08997	0.02350

Tabla 6.20. Parámetros de regresión y p -valores para proporción de pobreza $r = 1$ (año 2005).

Variabes	constante	age1	age2	edu1	cit1	lab2
β	-0.74082	0.90128	1.69006	0.68293	0.37468	1.78575
p -valor	0.00000	0.00595	0.00127	0.00000	0.00163	0.00006

Tabla 6.21. Parámetros de regresión y p -valores para proporción de pobreza $r = 2$ (año 2006).

A continuación se construyen los intervalos de confianza correspondientes, donde el nivel de significación considerado es $\alpha = 0,1$. En este caso adoptan la forma $\hat{\beta}_{rj} \pm \sqrt{q}z_{\alpha/2}$, donde q es el elemento de la diagonal principal de la matriz Q , definida en la sección 2.4, que se corresponde con $\hat{\beta}_{rj}$. En la última columna de la tabla se incluye 'V' o 'F' según 0 pertenezca al intervalo de confianza o no. Lo anterior se resume en las tablas 6.22 y 6.23

Variabes	$\hat{\beta}_{rj}$	$\hat{\beta}_{rj} - \sqrt{q}z_{\alpha/2}$	$\hat{\beta}_{rj} + \sqrt{q}z_{\alpha/2}$	$0 \in IC$
constante	-0.53822	-0.78842	-0.28801	F
age1	0.67365	0.13750	1.20979	F
age2	1.74785	0.81348	2.68221	F
edu1	0.60288	0.44500	0.76076	F
cit1	0.23671	0.00707	0.46636	F
lab2	0.99024	0.27115	1.70933	F

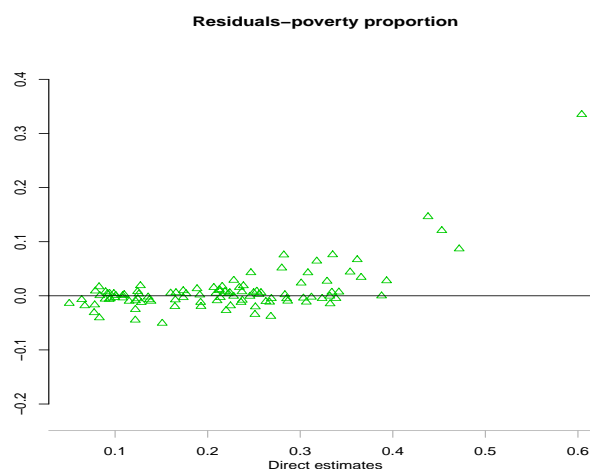
Tabla 6.22. Intervalos de confianza para la proporción de pobreza en 2005.

Variables	$\hat{\beta}_{rj}$	$\hat{\beta}_{rj} - \sqrt{q}z_{\alpha/2}$	$\hat{\beta}_{rj} + \sqrt{q}z_{\alpha/2}$	$0 \in IC$
constante	-0.74082	-0.95802	-0.52363	F
age1	0.90127	0.36224	1.44030	F
age2	1.69005	0.82734	2.55276	F
edu1	0.68293	0.51536	0.85050	F
cit1	0.37468	0.17901	0.57034	F
lab2	1.78575	1.04967	2.52182	F

Tabla 6.23. Intervalos de confianza para la proporción de pobreza en 2006.

En las tablas 6.20 y 6.21, tanto en la primera como en la segunda, se puede observar examinando los p -valores que los parámetros de regresión son significativos, excepto el que se corresponde con la variable *cit1* para $r = 1$. No obstante, se mantiene la variable en el modelo por el interés que tiene su interpretación. Por otra parte, se observa en las tablas 6.22 y 6.23 que el 0 no pertenece a ninguno de los intervalos de confianza, excepto en el intervalo que se corresponde con la variable *cit1* para $r = 1$. También se puede decir, observando la magnitud y los signos de los mismos, que al aumentar la población formada por los individuos con nivel de estudios primarios el nivel de pobreza aumenta considerablemente. Además se observa que al aumentar la población que se encuentra en situación de paro la proporción de pobreza aumenta.

En la figura 6.21 se muestran las gráficas de los pares $(y_{dr}, y_{dr} - x_{dr}\hat{\beta}_r - \hat{u}_{dr})$, donde la variable objetivo es la proporción de pobreza en 2006 ($\alpha = 0$, $r = 2$). La dispersión del gráfico no parece decir nada en contra de la hipótesis de insesgaredad del modelo ajustado. Asimismo, en la parte derecha de del gráfico, se observa una tendencia a presentar residuos positivos que se corresponden con los valores mayores de los estimadores directos. Este hecho se considera una propiedad interesante, ya que significa que el modelo ajustado tiende a disminuir aquellos valores que superan cierto nivel, y con ello se consigue evitar la presencia de outliers.

Figura 6.21: Residuos frente a estimadores directos para $\alpha = 0$, $r = 2$ (año 2006).

La figura 6.22 presenta los diagramas de cajas de los residuos $y_{dr} - x_{dr}\hat{\beta}_r - \hat{u}_{dr}$ estandarizados para 2005 y 2006. Los diagramas muestran una asimetría aceptable con una cola más amplia que la distribución normal por la parte positiva del eje real.

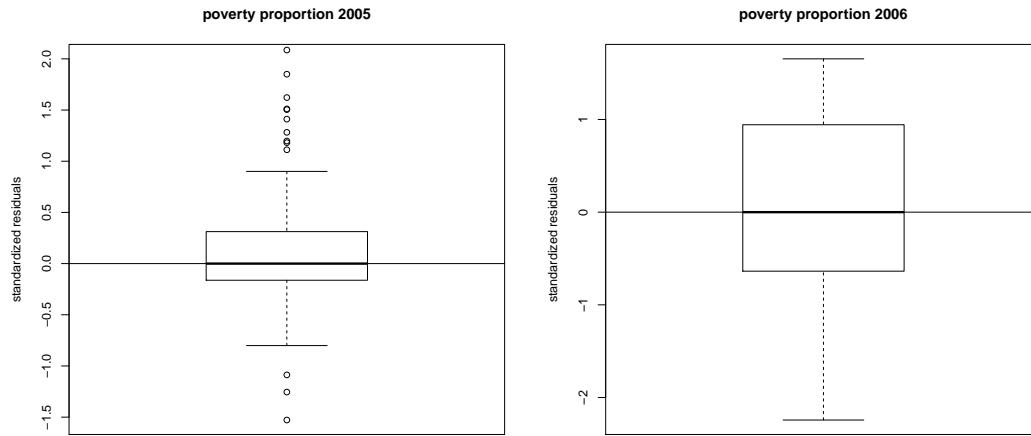


Figura 6.22: Diagramas de cajas de residuos estandarizados.

En las gráfica 6.23 se puede apreciar con claridad que el estimador directo alcanza valores superiores respecto del estimador EBLUP2 para tamaños muestrales inferiores, a medida que aumentan éstos la diferencia tiende a reducirse. Por otra parte, en las gráficas que corresponden a la raíz cuadrada de la varianza de los dos estimadores (gráfica 6.24, se observan mayores valores en lo que respecta al estimador directo, la diferencia es más acusada para tamaños muestrales inferiores.

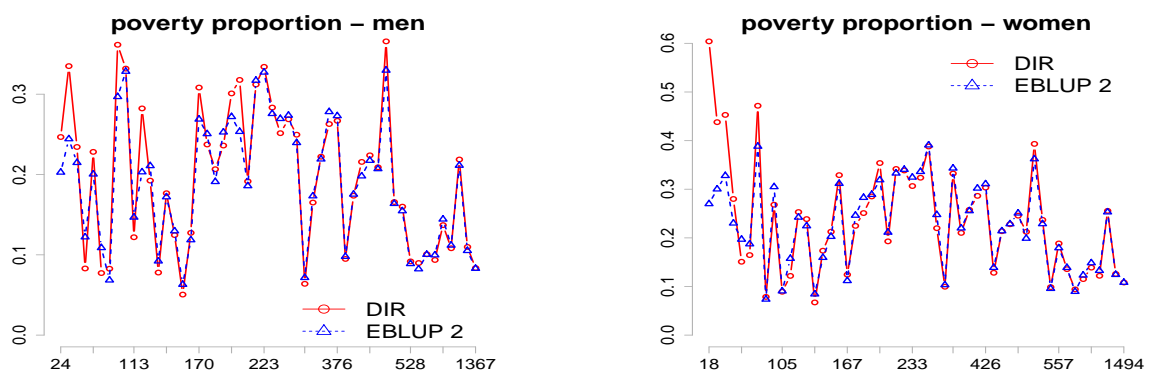


Figura 6.23: Estimaciones EBLUP2 y DIR de proporciones de pobreza por provincias en 2006.

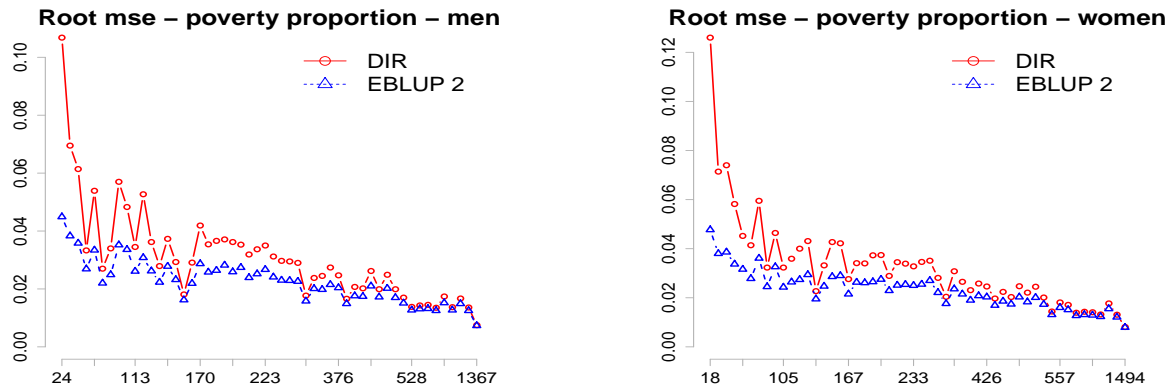


Figura 6.24: RMSEs de estimadores EBLUP2 y DIR de proporciones de pobreza por provincias en 2006.

6.3.3. Modelo con varianza de los efectos AR(1) heterocedástico

Se supone que los efectos que corresponden a los distintos dominios se distribuyen según un proceso estocástico AR(1) heterocedástico. Por tanto, se tiene que

$$u_{dr} = \rho u_{dr-1} + a_{dr}, \quad u_{d0} \sim N(0, 1), \quad a_{dr} \sim N(0, \sigma_r^2), \quad r = 1, 2.$$

Se supone que a_{d1} y a_{d2} son independientes y que

$$u \sim N(0, V_{ud}), \quad V_u = \text{diag}(V_{ud}), \quad \text{donde } V_{ud} = \begin{pmatrix} \sigma_1^2 + \rho^2 & \rho\sigma_1^2 + \rho^3 \\ \rho\sigma_1^2 + \rho^3 & \sigma_2^2 + \rho^2\sigma_1^2 + \rho^4 \end{pmatrix}.$$

Con el objeto de obtener una estimación inicial para los parámetros σ_1^2 , σ_2^2 y ρ , se considerarán los modelos univariantes que se deducen a partir del modelo multivariante planteado en la sección 6.1; es decir,

$$y_{dr} = x_{dr}\beta_r + u_{dr} + e_{dr}, \quad d = 1, \dots, D, \quad r = 1, 2,$$

donde

$$u_{.r} \sim N_D(0, \sigma_{ur}^2 I_D), \quad e_{.r} \sim N_D(0, W_r^{-1}), \quad W_r^{-1} = \text{diag}(\sigma_{edr}^{-2}).$$

A partir de ellos se estiman las varianzas aplicando la fórmula del método de Henderson 3; es decir,

$$\hat{\sigma}_{urH}^2 = \frac{y'_{.r} P_2 y_{.r} - D + p_r}{\text{tr}\{P_2\}}, \quad r = 1, 2.$$

Se obtiene

$$\hat{\sigma}_{u1H}^2 = 0,001803664 \quad \text{y} \quad \hat{\sigma}_{u2H}^2 = 0,03417097.$$

En las primeras pruebas que se han realizado para estimar los parámetros por el método de Fisher-scoring los resultados obtenidos no han sido satisfactorios debido a que, a partir de distintas semillas consideradas siempre se obtiene una estimación para σ_{u1}^2 negativa; por ello, se concluye que el *modelo AR(1) heteroscedástico* no se ajusta bien a los datos. Se propone una alternativa al ajuste directo que consiste en distinguir la estimación en dos fases que se detallan a continuación.

1. Se estiman los parámetros σ_{u1}^2 y σ_{u2}^2 utilizando el *modelo diagonal* considerando como semillas los valores obtenidos por el método Henderson 3.
2. Las estimaciones $\hat{\sigma}_{u1}^2$ y $\hat{\sigma}_{u2}^2$ obtenidas en la Fase 1 se consideran como verdaderos valores de los parámetros σ_{u1}^2 y σ_{u2}^2 y así se procede a estimar el parámetro ρ por Fisher-scoring tomando como semilla para el mismo $\rho^{(0)} = 0$.

Para ejecutar la primera parte del procedimiento propuesto se aprovecha lo que se ha realizado en el apartado 6.3.1 y se apunta que las estimaciones para σ_{u1}^2 y σ_{u2}^2 son

$$\hat{\sigma}_{u1}^2 = 0,002597 \quad \text{y} \quad \hat{\sigma}_{u2}^2 = 0,001928316.$$

A partir de las estimaciones anteriores se tiene que la matriz de varianzas de los efectos es

$$V_{ud} = \begin{pmatrix} \sigma_1^2 + \rho^2 & \rho\sigma_1^2 + \rho^3 \\ \rho\sigma_1^2 + \rho^3 & \sigma_2^2 + \rho^2\sigma_1^2 + \rho^4 \end{pmatrix} = \begin{pmatrix} 0,002597 + \rho^2 & 0,002597\rho + \rho^3 \\ 0,002597\rho + \rho^3 & 0,001928 + 0,002597\rho^2 + \rho^4 \end{pmatrix}.$$

Ahora, se procede a realizar la segunda fase del procedimiento propuesto y se tiene que el algoritmo Fisher-scoring converge. Después de veintitrés iteraciones se llega a la estimación $\hat{\rho} = 0,02105169$.

Contrastamos $H_0 : \sigma_{u1}^2 = \sigma_{u2}^2$. El estadístico del contraste es

$$T_{12} = \frac{\hat{\sigma}_{u1}^2 - \hat{\sigma}_{u2}^2}{\sqrt{v_{11} + v_{22} - 2v_{12}}} = 1,0756,$$

donde v_{ij} , $i, j = 1, 2, 3$ son los elementos de la inversa de la matriz REML de información de Fisher del modelo AR(1) heteroscedástico evaluada en $\hat{\theta} = (\hat{\sigma}_1^2, \hat{\sigma}_2^2, \hat{\rho})$. Como T_{12} tiene distribución asintótica normal estándar bajo H_0 , el p -valor es 0.28208. No podemos concluir que las varianzas de los efectos aleatorios sean distintas. Por tanto, preferimos utilizar el modelo AR(1), en lugar del modelo AR(1) heteroscedástico, para dar estimaciones EBLUP.

El siguiente paso, en el estudio del modelo AR(1) heteroscedástico, es calcular las estimaciones y los intervalos de confianza para los parámetros β_{rj} , $r = 1, 2$. Utilizando las distribuciones asintóticas de los estimadores $\hat{\beta}_{rj}$, se calculan los p -valores correspondientes a los contrastes $H_0 : \beta_{rj} = 0$. Los resultados se presentan en la tablas 6.24 y 6.25.

Variabes	constante	age1	age2	edu1	cit1	lab2
β	-0.65428	0.69799	2.38239	0.71074	0.25923	0.71268
<i>p</i> -valor	0.00010	0.06539	0.00048	0.00000	0.08960	0.15129

Tabla 6.24. Parámetros de regresión y *p*-valores para $r = 1$ (año 2005).

Variabes	constante	age1	age2	edu1	cit1	lab2
β	-0.75277	0.88496	1.89752	0.79734	0.31471	2.04599
<i>p</i> -valor	0.00000	0.00608	0.00047	0.00000	0.00414	0.00000

Tabla 6.25. Parámetros de regresión y *p*-valores para $r = 2$ (año 2006).

A continuación se construyen los intervalos de confianza correspondientes, donde el nivel de significación considerado es $\alpha = 0,1$. En este caso adoptan la forma $\hat{\beta}_{rj} \pm \sqrt{q}z_{\alpha/2}$, donde q es el elemento de la diagonal principal de la matriz Q , definida en la sección 2.4, que se corresponde con $\hat{\beta}_{rj}$. En la última columna de la tabla se incluye 'V' o 'F' según 0 pertenezca al intervalo de confianza o no. Lo anterior se resume en las tablas 6.26 y 6.27.

Variabes	$\hat{\beta}_{rj}$	$\hat{\beta}_{rj} - \sqrt{q}z_{\alpha/2}$	$\hat{\beta}_{rj} + \sqrt{q}z_{\alpha/2}$	$0 \in IC$
constante	-0.65428	-0.93115	-0.37741	F
age1	0.69799	0.07488	1.32111	F
age2	2.38239	1.25840	3.50638	F
edu1	0.71074	0.53256	0.88892	F
cit1	0.25923	0.00803	0.51043	F
lab2	0.71268	-0.10423	1.52959	V

Tabla 6.26. Intervalos de confianza para $\alpha = 0, r = 1$ (año 2005).

Variabes	$\hat{\beta}_{rj}$	$\hat{\beta}_{rj} - \sqrt{q}z_{\alpha/2}$	$\hat{\beta}_{rj} + \sqrt{q}z_{\alpha/2}$	$0 \in IC$
constante	-0.75277	-0.95497	-0.55058	F
age1	0.88496	0.35430	1.41562	F
age2	1.89752	1.00476	2.79028	F
edu0	0.79734	0.63627	0.95841	F
cit1	0.31471	0.13417	0.49525	F
lab2	2.04599	1.33638	2.75561	F

Tabla 6.27. Intervalos de confianza para $\alpha = 0, r = 2$ (año 2006).

En las tablas 6.26-6.27 se observa que el 0 no pertenece a ninguno de los intervalos de confianza, excepto en el caso de β_{24} . Por tanto se concluye que todas las variables explicativas son significativas, salvo *edu2* para $r = 2$. No obstante, se mantiene la variable en el modelo por el interés que tiene la comparación con el modelo multivariante diagonal de la sección 6.3.1. Por otra parte, observando la magnitud y los signos de los mismos, se puede observar que al aumentar la población formada por los individuos con nivel de estudios primarios o inferiores aumenta la proporción de pobreza, siendo el aumento considerablemente mayor en el caso de nivel de estudios inferior a la educación primaria. Además se observa que al aumentar la población que se encuentra en situación de paro la proporción de pobreza aumenta.

En la figura 6.25 se muestra la gráfica de los pares $(y_{dr}, y_{dr} - x_{dr}\hat{\beta}_r - \hat{u}_{dr})$ que se corresponden con la variable proporción de pobreza en el año 2006 ($\alpha = 0$, $r = 2$). La dispersión que se observa en el gráfico no parece decir nada en contra de la hipótesis de insesgadez del modelo ajustado. Asimismo, en la parte derecha del gráfico se observa una tendencia a presentar residuos positivos que se corresponden con los valores mayores de los estimadores directos. Este hecho se considera una propiedad interesante, ya que significa que el modelo ajustado tiende a disminuir aquellos valores que superan cierto nivel, y con ello se consigue evitar la presencia de outliers.

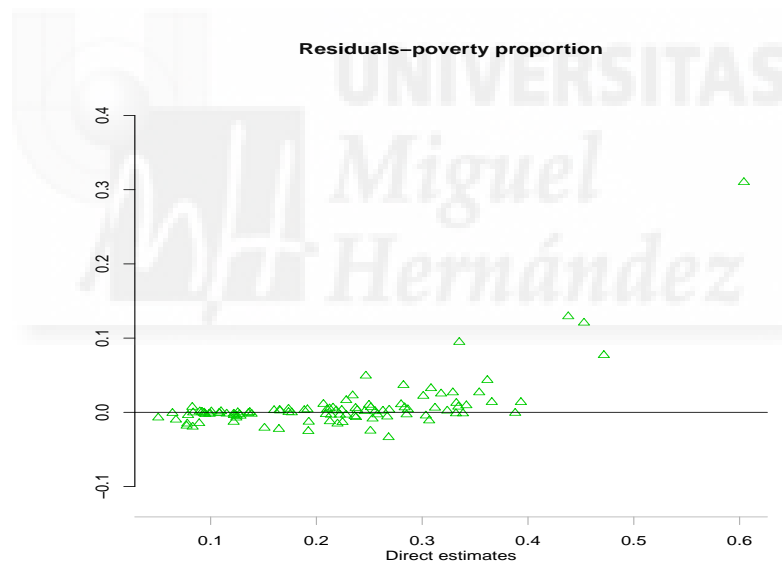


Figura 6.25: Residuos frente a estimadores directos para $\alpha = 0$, $r = 2$ (año 2006).

En la gráfica 6.28 se puede apreciar con claridad que el estimador directo alcanza valores superiores respecto del estimador EBLUP3 para tamaños muestrales inferiores, a medida que aumentan éstos la diferencia tiende a reducirse. Por otra parte, en las gráficas que corresponden a la raíz cuadrada de la varianza de los dos estimadores (gráfica 6.27), se observan mayores valores en lo que respecta al estimador directo, la diferencia es más acusada para tamaños muestrales inferiores.

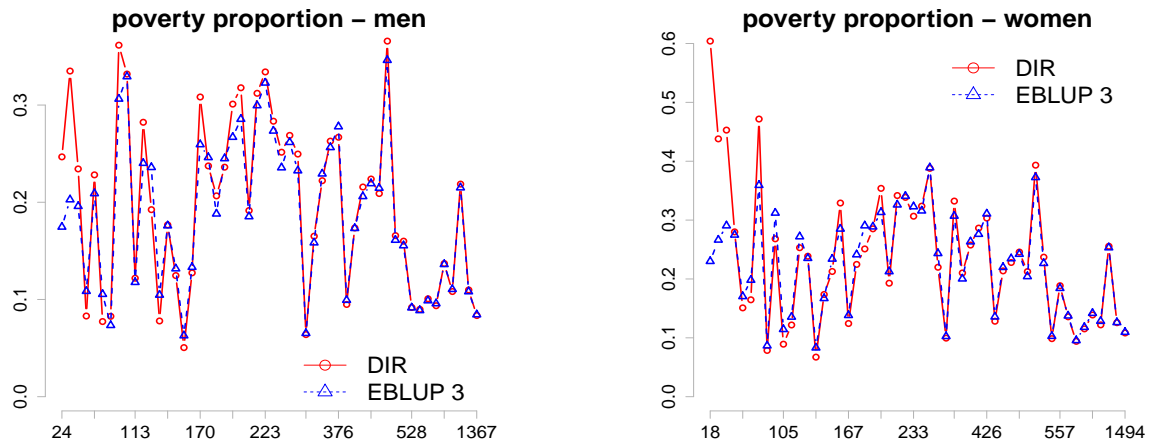


Figura 6.26: Estimaciones EBLUP3 y DIR de proporciones de pobreza por provincias en 2006.

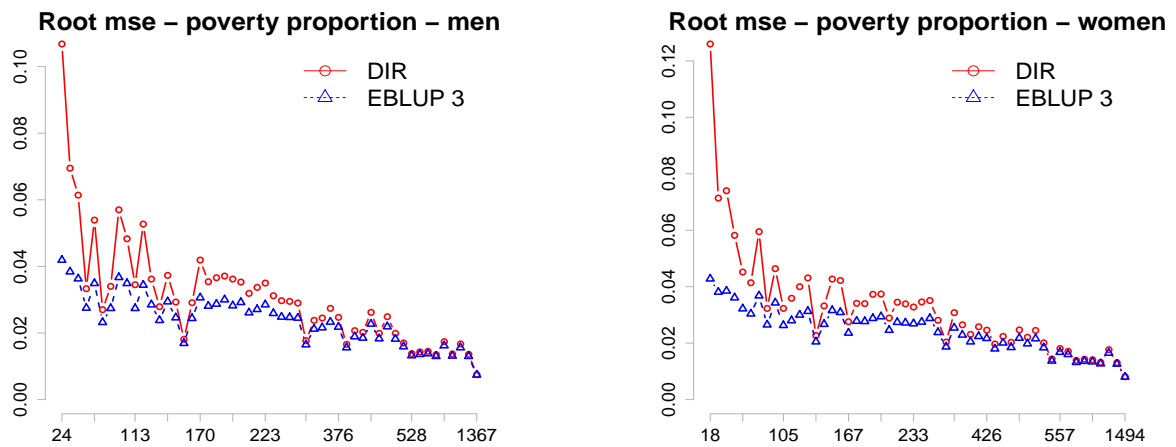


Figura 6.27: RMSEs de estimadores EBLUP3 y DIR de proporciones de pobreza por provincias en 2006.

6.3.4. Conclusiones

En los apartados 6.3.1, 6.3.2 y 6.3.3 se han utilizado el *modelo diagonal*, *modelo AR(1)* y el *modelo AR(1) heterocedástico* y se selecciona finalmente el modelo *modelo AR(1)*. En este apartado se dan los resultados obtenidos usando este modelo. Se expone un resumen de las estimaciones que se han obtenido al utilizar el *modelo AR(1)*, debido a que es el modelo mediante el cual se obtienen errores cuadráticos medios inferiores.

La figura 6.28 contiene mapas de España en los que las provincias se colorean según los niveles de proporción de pobreza y de brecha de pobreza definidos en la tabla 6.28. Se observa que la proporción de la población por debajo de la línea de pobreza es menor en las provincias del noreste como Cataluña, Aragón, Navarra, Catabria. Por otra parte, se observa que las provincias españolas con mayor proporción de pobreza se encuentran situadas en el centro y en el sur como Andalucía, Extremadura, Castilla la Mancha y Canarias. En una posición intermedia se encuentran algunas provincias españolas del centro norte como Galicia, algunas provincias de Castilla León, Madrid y Comunidad Valenciana.

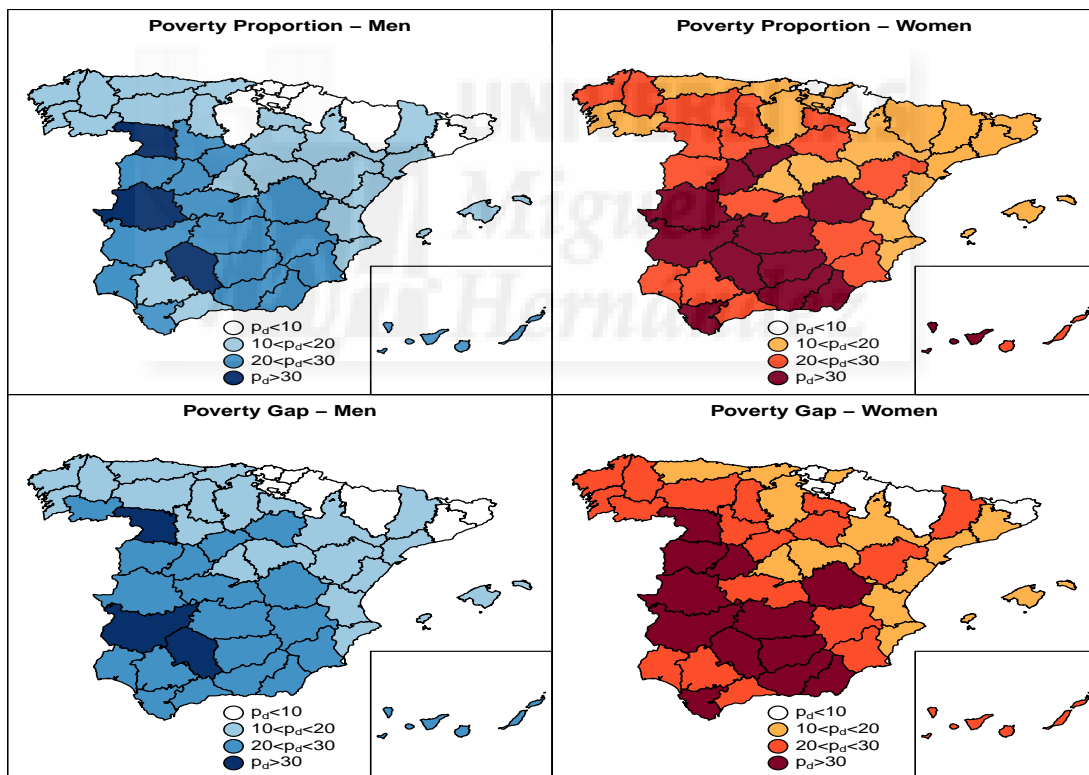


Figura 6.28: Estimaciones de las proporciones de pobreza en 2005 (arriba) y 2006 (abajo) para hombres (izquierda) y mujeres (derecha).

En la tabla 6.28 se presenta una clasificación de las provincias españolas en cuatro categorías según los valores del EBLUP del modelo estimado de la proporción de pobreza, es decir, $p_d = 100 \cdot \hat{Y}_{0,d,2006}^{eblup}$. Los mismos resultados se presentan en la figura 6.28. Se observa que la proporción de la población por debajo de la línea de pobreza es menor en las provincias del noreste como Cataluña, Aragón, Navarra, Catabria. Por otra parte, se observa que las provincias españolas con mayor proporción de pobreza se encuentran situadas en el centro y en el sur como Andalucía, Extremadura, Castilla la Mancha y Canarias. En una posición intermedia se encuentran algunas provincias españolas del centro norte como Galicia, algunas provincias de Castilla León, Madrid y Comunidad Valenciana.

men	$p_d < 10$	1 8 17 20 22 31 48
	$10 < p_d < 20$	3 7 9 12 15 19 24 25 26 27 28 33 34 36 39 43 44 46 47 50
	$20 < p_d < 30$	2 4 5 10 11 13 16 18 21 23 29 30 32 35 37 38 40 41 42 45 52
	$p_d > 30$	6 14 49 51
women	$p_d < 10$	1 17 22 31 48
	$10 < p_d < 20$	3 7 8 9 12 19 20 28 33 39 43 46 50
	$20 < p_d < 30$	2 15 21 24 25 26 27 29 30 32 34 35 36 38 40 41 42 44 45 47 52
	$p_d > 30$	4 5 6 10 11 13 14 16 18 23 37 49 51

Tabla 6.28. Provincias españolas clasificadas por proporción de pobreza.

La tabla 6.29 presenta las estimaciones de la proporción de pobreza en 2006 al utilizar el *modelo AR(1)*. La primera columna contiene la provincia, las tres siguientes muestran el estimador directo, el EBLUP0 y el EBLUP2 correspondientes a la subpoblación de hombres, las tres siguientes muestran lo mismo para las mujeres. Las seis últimas columnas se disponen de la misma forma para presentar la raíz cuadrada de los errores cuadráticos medios.

<i>d</i>	men/poverty proportions/women						men/sqrt.mse/women					
	dir	eb0	eb2	dir	eb0	eb2	dir	eb0	eb2	dir	eb0	eb2
1	0.083	0.073	0.066	0.079	0.083	0.070	0.034	0.028	0.025	0.032	0.027	0.024
2	0.237	0.243	0.246	0.285	0.290	0.291	0.035	0.029	0.026	0.037	0.030	0.026
3	0.160	0.156	0.155	0.189	0.182	0.175	0.017	0.016	0.015	0.018	0.017	0.016
4	0.318	0.289	0.254	0.354	0.307	0.310	0.035	0.030	0.027	0.037	0.030	0.027
5	0.335	0.222	0.259	0.453	0.296	0.332	0.069	0.040	0.038	0.074	0.040	0.039
6	0.366	0.349	0.332	0.393	0.375	0.366	0.025	0.022	0.020	0.025	0.022	0.020
7	0.094	0.096	0.100	0.115	0.119	0.125	0.014	0.013	0.013	0.014	0.014	0.013
8	0.083	0.085	0.083	0.108	0.109	0.108	0.008	0.007	0.007	0.008	0.008	0.008
9	0.127	0.128	0.109	0.124	0.141	0.116	0.029	0.025	0.022	0.028	0.024	0.022
10	0.252	0.237	0.272	0.332	0.309	0.347	0.030	0.025	0.023	0.031	0.026	0.024
11	0.267	0.280	0.279	0.303	0.308	0.308	0.025	0.022	0.020	0.025	0.022	0.020
12	0.122	0.118	0.147	0.122	0.142	0.167	0.034	0.028	0.026	0.036	0.029	0.026
13	0.269	0.263	0.274	0.324	0.309	0.329	0.030	0.025	0.023	0.035	0.028	0.025
14	0.312	0.298	0.315	0.307	0.315	0.318	0.034	0.028	0.025	0.033	0.027	0.025
15	0.216	0.207	0.198	0.237	0.227	0.230	0.020	0.019	0.017	0.020	0.019	0.017
16	0.362	0.313	0.295	0.472	0.360	0.385	0.057	0.038	0.036	0.059	0.038	0.036
17	0.050	0.063	0.064	0.067	0.082	0.086	0.018	0.017	0.016	0.023	0.021	0.019
18	0.301	0.274	0.277	0.342	0.328	0.335	0.036	0.029	0.026	0.034	0.028	0.025
19	0.077	0.105	0.109	0.165	0.195	0.184	0.027	0.024	0.022	0.041	0.031	0.028
20	0.064	0.065	0.071	0.100	0.102	0.103	0.018	0.017	0.016	0.020	0.019	0.018
21	0.192	0.230	0.204	0.253	0.272	0.245	0.036	0.029	0.026	0.040	0.031	0.027
22	0.078	0.107	0.095	0.089	0.117	0.095	0.028	0.024	0.022	0.032	0.027	0.024
23	0.283	0.279	0.281	0.339	0.342	0.343	0.031	0.026	0.024	0.034	0.028	0.025
24	0.192	0.189	0.190	0.193	0.211	0.212	0.032	0.027	0.024	0.029	0.025	0.023
25	0.177	0.179	0.173	0.239	0.232	0.220	0.037	0.030	0.028	0.043	0.032	0.029
26	0.166	0.160	0.160	0.212	0.207	0.203	0.020	0.018	0.017	0.022	0.020	0.018
27	0.207	0.190	0.191	0.225	0.239	0.243	0.037	0.030	0.026	0.034	0.029	0.026
28	0.110	0.110	0.108	0.126	0.125	0.122	0.014	0.013	0.012	0.013	0.013	0.012
29	0.222	0.229	0.218	0.258	0.261	0.252	0.025	0.022	0.020	0.023	0.021	0.019
30	0.219	0.216	0.211	0.256	0.254	0.253	0.017	0.016	0.015	0.018	0.017	0.016
31	0.090	0.089	0.082	0.094	0.095	0.089	0.014	0.014	0.013	0.014	0.013	0.013
32	0.282	0.248	0.207	0.213	0.232	0.201	0.053	0.036	0.031	0.043	0.033	0.029
33	0.108	0.110	0.112	0.122	0.127	0.131	0.014	0.013	0.013	0.013	0.013	0.012
34	0.228	0.213	0.200	0.280	0.276	0.229	0.054	0.036	0.034	0.058	0.038	0.034
35	0.224	0.219	0.218	0.246	0.239	0.247	0.026	0.023	0.021	0.025	0.022	0.020
36	0.174	0.175	0.177	0.214	0.221	0.217	0.021	0.019	0.018	0.022	0.020	0.019
37	0.308	0.259	0.266	0.329	0.276	0.302	0.042	0.032	0.029	0.042	0.032	0.029
38	0.263	0.253	0.274	0.286	0.272	0.296	0.027	0.024	0.021	0.026	0.023	0.021
39	0.095	0.101	0.100	0.128	0.136	0.141	0.017	0.016	0.015	0.020	0.018	0.017
40	0.234	0.202	0.219	0.438	0.259	0.292	0.061	0.038	0.036	0.071	0.040	0.038
41	0.209	0.213	0.204	0.228	0.234	0.229	0.020	0.019	0.017	0.020	0.019	0.017
42	0.247	0.181	0.204	0.604	0.230	0.269	0.107	0.044	0.046	0.126	0.045	0.049
43	0.125	0.132	0.130	0.174	0.171	0.164	0.029	0.025	0.023	0.033	0.027	0.025
44	0.083	0.110	0.124	0.151	0.175	0.202	0.033	0.028	0.027	0.045	0.033	0.032
45	0.250	0.238	0.244	0.220	0.241	0.247	0.029	0.025	0.022	0.028	0.024	0.022
46	0.137	0.136	0.144	0.139	0.141	0.149	0.017	0.016	0.015	0.014	0.014	0.013
47	0.165	0.158	0.173	0.210	0.199	0.219	0.024	0.022	0.020	0.027	0.023	0.022
48	0.092	0.091	0.088	0.099	0.102	0.095	0.014	0.013	0.013	0.014	0.014	0.013
49	0.332	0.337	0.333	0.268	0.312	0.306	0.048	0.036	0.034	0.046	0.035	0.033
50	0.101	0.099	0.100	0.136	0.137	0.138	0.014	0.014	0.013	0.017	0.016	0.015
51	0.334	0.322	0.328	0.388	0.384	0.388	0.035	0.029	0.027	0.035	0.029	0.027
52	0.236	0.243	0.248	0.251	0.291	0.285	0.037	0.031	0.028	0.034	0.028	0.026

Tabla 6.29. Estimaciones de la proporción de pobreza y de sus raíces de errores cuadráticos medios en 2006.

Conclusiones generales

En este último capítulo se exponen unas conclusiones generales referidas a los estudios que se ha realizado en los distintos capítulos sobre modelos lineales mixtos multivariantes de área. También se aprovecha para realizar una breve exposición de los posibles avances que se podrían conseguir a partir de algunos de los resultados que se han presentado. Así pues, en lo que sigue se desarrollan dos secciones; la primera sobre conclusiones generales y la segunda sobre líneas de investigación posibles.

7.1. Conclusiones generales

En numerosos estudios sobre las más diversas áreas del conocimiento, como pueden ser la economía, la sociología, la medicina, etc., los investigadores están interesados en más de una variable objetivo. En general, las variables objetivo están correladas y sobre ellas se dispone de cierta información auxiliar. No es menos habitual que el interés del estudio esté centrado en la estimación de indicadores asociados a distintas partes (dominios) de la población. En tales casos, la muestra disponible en cada uno de los dominios es pequeña y no permite obtener estimaciones directas fiables. Como consecuencia de lo anterior, se introducen y estudian modelos de predicción que abarquen un enfoque multivariante y que contemplen la distinción de las distintas áreas pequeñas que conforman la muestra total.

En esta memoria se ha optado por definir el modelo lineal multivariante general (2.1) que permite introducir varios modelos específicos y que incluye efectos aleatorios para los dominios o áreas pequeñas. Los modelos desarrollados representan distintas estructuras de correlación entre los efectos aleatorios. En concreto, se estudian tres modelos: *Modelo diagonal*, *Modelo ARI* y *Modelo ARI heterocedástico*. Para cada uno de los modelos se ha planteado su formulación teórica exponiendo cada uno de los elementos que intervienen en los mismos. Además, mediante simulaciones implementadas en lenguaje de programación R, se han realizado estudios empíricos sobre los algoritmos de ajuste, los predictores EBLUP y los estimadores del error cuadrático medio. Los resultados de las simulaciones se presentan en tablas que contienen sesgos y errores cuadráticos medios empíricos y en gráficos de dispersión y diagramas de caja. Los resultados de las simulaciones ilustran el comportamiento satisfactorio de la metodología estadística introducida.

Después de estudiar los tres modelos planteados en los capítulos iniciales, se han planteado aplicaciones a datos socioeconómicos de muestras de la encuesta de condiciones de vida. En concreto, se han considerado dos enfoques diferentes. Para el primer enfoque (multivariante) se ha centrado la atención en dos variables objetivo: *proporción de pobreza* y *brecha de pobreza* en un periodo de tiempo determinado (año 2006). En el segundo enfoque (temporal) se ha estudiado la variable *proporción de pobreza* en dos periodos de tiempo determinados (año 2005 y año 2006)

Se ha seguido un proceso similar para los dos enfoques que se han considerado. En cada enfoque se han tenido en cuenta los tres modelos, y la metodología se puede resumir en lo que sigue:

1. Ajustar los datos de la muestra disponible utilizando los tres modelos estudiados.
2. Determinar estimaciones de parámetros y realizar contrastes de hipótesis.
3. Representar gráficamente las estimaciones directas y las estimaciones obtenidas del modelo en cuestión.
4. Representar gráficamente los errores cuadráticos medios del estimador directo y las estimaciones obtenidas para el error cuadrático medio del modelo en cuestión.

Después de completar los distintos ajustes y estudiar las estimaciones y los gráficos de dispersión obtenidos, se ha optado por seleccionar el *modelo ARI heterocedástico* para el enfoque multivariante y el *modelo ARI* para el enfoque temporal. Una vez elegido el modelo apropiado para cada enfoque, se han expuesto en distintas tablas más detalles sobre las estimaciones obtenidas, así como gráficas sobre las áreas geográficas que intervienen en las muestras de datos.

González-Manteiga et al. (2008) estudiaron una clase de modelos Fay-Herriot multivariantes con un efecto aleatorio común a todas las componentes del vector de variables objetivo. Ellos introdujeron además estimadores bootstrap de los errores de predicción. El trabajo de estos autores es el punto de partida y la inspiración de esta memoria. Se ha introducido una clase de modelos multivariantes que utiliza un efecto aleatorio distinto para cada componente del vector de variables objetivo. Se han deducido estimadores EBLUP y se han dado procedimientos para estimar los errores de predicción.

Esteban et al. (2012) utilizan modelos Fay-Herriot temporales y datos de la encuesta de condiciones de vida para estimar indicadores de pobreza en provincias españolas. Los modelos de esta memoria también se pueden aplicar en el contexto temporal. Por tal motivo, la siguiente sección presenta una comparación de los resultados de Esteban et al. (2012) con los correspondientes del capítulo 6.

7.2. Comparaciones

Esta sección compara los estimadores de indicadores de pobreza provinciales de Esteban et al. (2012) con los correspondientes del capítulo 6 y con los estimadores directos. Por tal motivo, se consideran separadamente las subpoblaciones de hombres y de mujeres. Ordenamos las provincias por tamaño muestral y

presentamos los resultados de las posiciones $5 \times k + 1$, $k = 1, \dots, 10$. También incluimos los resultados de Barcelona.

Esteban et al. (2012) estudia varias extensiones univariantes del modelo Fay-Herriot a datos temporales. Estos autores recomiendan usar su modelo 3 con efectos aleatorios que tienen una correlación temporal AR(1) dentro de cada dominio. Ellos usan datos del pasado de los años 2004 y 2005 para dar estimaciones de 2006. Las tablas 7.7 y 7.8 etiquetan sus EBLUPs y sus estimaciones de las raíces cuadradas de los errores cuadráticos medios (root-MSEs) con E3 y rE3 respectivamente. Estas tablas incluyen los resultados de las aplicaciones 1 y 2 del capítulo 6, que se etiquetan equivalentemente con A1, A2, rA1 y rA2. Para los estimadores directos se usa la notación Dir y rDir respectivamente. Finalmente n_d denota el tamaño muestral del dominio d en la encuesta de condiciones de vida de 2006.

province	n_d	Dir	A1	A2	E3	rDir	rA1	rA2	rE3
Soria	24	0.247	0.189	0.203	0.236	0.107	0.031	0.045	0.026
Guadalajara	92	0.077	0.077	0.109	0.099	0.027	0.020	0.022	0.021
Huelva	124	0.192	0.181	0.211	0.221	0.036	0.023	0.026	0.024
Gerona	145	0.050	0.051	0.063	0.068	0.018	0.016	0.016	0.021
Albacete	173	0.237	0.233	0.251	0.254	0.035	0.023	0.026	0.026
Córdoba	221	0.312	0.311	0.317	0.317	0.034	0.023	0.025	0.030
Guipúzcoa	280	0.064	0.063	0.071	0.078	0.018	0.015	0.016	0.021
Santander	428	0.095	0.095	0.098	0.087	0.017	0.015	0.015	0.022
Badajoz	477	0.366	0.364	0.330	0.305	0.025	0.020	0.020	0.030
Zaragoza	556	0.101	0.100	0.100	0.100	0.014	0.013	0.013	0.022
Madrid	911	0.110	0.110	0.105	0.094	0.014	0.012	0.013	0.022
Barcelona	1367	0.083	0.084	0.083	0.086	0.008	0.007	0.007	0.022

Table 7.7. Proporciones de pobreza y root-MSEs para hombres en 2006.

province	n_d	Dir	A1	A2	E3	rDir	rA1	rA2	rE3
Soria	18	0.604	0.529	0.270	0.297	0.126	0.032	0.048	0.034
Guadalajara	86	0.165	0.154	0.187	0.156	0.041	0.025	0.028	0.023
Huelva	124	0.253	0.251	0.242	0.248	0.040	0.025	0.027	0.025
Gerona	138	0.067	0.067	0.084	0.083	0.023	0.019	0.019	0.021
Albacete	193	0.285	0.281	0.289	0.280	0.037	0.024	0.027	0.026
Córdoba	233	0.307	0.306	0.324	0.323	0.033	0.023	0.025	0.029
Guipúzcoa	292	0.100	0.098	0.103	0.102	0.020	0.017	0.018	0.022
Santander	448	0.128	0.128	0.139	0.123	0.020	0.017	0.017	0.022
Badajoz	517	0.393	0.392	0.362	0.334	0.025	0.019	0.020	0.031
Zaragoza	577	0.136	0.135	0.139	0.121	0.017	0.015	0.015	0.022
Madrid	1008	0.126	0.126	0.124	0.111	0.013	0.012	0.012	0.022
Barcelona	1494	0.108	0.108	0.108	0.106	0.008	0.008	0.008	0.022

Table 7.8. Proporciones de pobreza y root-MSEs para mujeres en 2006.

Se observa que los estimadores A1 están en general más cerca de los estimadores directos que los estimadores A2 y E3. Esto ocurre porque los estimadores directos de la proporción y de la brecha de pobreza están altamente correlados y el modelo AR(1) heterocedástico incluye las correlaciones muestrales de esos estimadores en la matriz 2×2 de covarianzas de los vectores $e_d = (e_{d1}, e_{d2})'$ de errores de muestreo. Por otro lado, los modelos que dan lugar a los EBLUPs A2 y E3 suponen que los errores muestrales de diferentes periodos de tiempo son independientes. En consecuencia, no incorporan las correspondientes correlaciones muestrales.

También se observa que los valores E3 se parecen más a los de A2 que a los de A1. Esto es un hecho natural dado que los valores de E3 y de A2 se obtienen a partir de modelos de correlación temporal. Las tablas 7.7 y 7.8 también presentan las root-MSEs. Se observa que los tres EBLUPs tienen root-MSE más pequeñas que las del estimador directo. También observamos que $rA1$ es en general más pequeño que $rA2$ y $rE3$. En ese sentido estaríamos tentados de recomendar A1 como la mejor opción. De todas formas, no damos la conclusión de que A1 es preferible a A2 o a E3 porque las root-MSEs no son comparables al estar basadas en modelos distintos. Comparando los residuos de los modelos para A1, A2 y E3, concluimos que los tres modelos tienen un ajuste similar a los datos. En consecuencia, recomendamos igualmente cualquiera de los tres EBLUPs.

7.3. Líneas futuras de investigación

De cara al futuro, y como continuación del presente trabajo, se plantean tres líneas de trabajo.

1. Estudiar nuevas estructuras de correlación que permitan el análisis de datos con dependencia temporal o espacial. Se pueden considerar estructuras de correlación temporal del tipo MA(1) o ARMA(1,1) y estructuras de correlación espacial del tipo SAR(1).
2. Desarrollar una teoría equivalente para modelos de unidad lineales mixtos y multivariantes. Esta línea de trabajo requerirá una programación eficiente de los algoritmos de ajuste y de los procedimientos de estimación de los errores cuadráticos medios de los predictores EBLUP.
3. Introducir modelos de área multinomiales mixtos, donde los efectos aleatorios presenten estructuras de correlación similares a las estudiadas en esta memoria.

A

Modelo Fay-Herriot univariante

En el presente trabajo se estudian modelos de area multivariantes basados en una teoría general que se presenta en el primer capítulo. En ocasiones se necesita estudiar el comportamiento de las variables estudiadas por separado, así como la determinación de estimaciones del parámetro asociado a la varianza del efecto aleatorio; por ello se incluye en este anexo la definición del modelo univariante. En concreto se va a estudiar la estimación que se obtiene por el método de los momentos y el que se obtiene por el método Henderson 3.

Con el propósito de poder realizar las comparaciones con el modelo multivariante de la forma más clara posible seguimos la misma notación utilizada en el modelo 3.1 y para ello basta con considerar el índice r fijo, y así queda lo siguiente

$$y_{dr} = \mu_{dr} + e_{dr}, \quad \text{donde} \quad \mu_{dr} = x_{dr}\beta_r + u_{dr}, \quad d = 1, \dots, D$$

El modelo univariante en forma matricial es el siguiente

$$\begin{pmatrix} y_{1r} \\ y_{2r} \\ \vdots \\ y_{Dr} \end{pmatrix} = \begin{pmatrix} x_{1r1} & x_{1r2} & \cdots & x_{1rp_r} \\ x_{2r1} & x_{2r2} & \cdots & x_{2rp_r} \\ \vdots & \vdots & \vdots & \vdots \\ x_{Dr1} & x_{Dr2} & \cdots & x_{Drp_r} \end{pmatrix} \begin{pmatrix} \beta_1 \\ \beta_2 \\ \vdots \\ \beta_{p_r} \end{pmatrix} + \begin{pmatrix} u_{1r} \\ u_{2r} \\ \vdots \\ u_{Dr} \end{pmatrix} + \begin{pmatrix} e_{1r} \\ e_{2r} \\ \vdots \\ e_{Dr} \end{pmatrix},$$

donde

$$u_{.r} \sim N_D(0, \sigma_{ur}^2 I_D) \quad \text{y} \quad e_{.r} \sim N_D(0, \sigma_0^2 W_r^{-1}).$$

y, además, los vectores $u_{.r}$ y $e_{.r}$ son independientes.

Finalmente, se puede expresar el modelo de una forma más resumida como se indica a continuación

$$y_{.r} = X_{.r}\beta_r + Z_{.r}u_{.r} + e_{.r}$$

La estimación de la varianza del efecto en el modelo Fay-Herriot univariante que se acaba de definir siguiendo el método Henderson 3 es

En el modelo que se está considerando se tiene que

$$W_r = \text{diag} (w_{dr}) \quad y \quad \sigma_0^2 = 1.$$

Método de los momentos

Una de las formas más inmediatas que facilitan una estimación del parámetro σ_u^2 consiste en considerar el modelo lineal sin la presencia del efecto aleatorio, es decir, $y_r = X_r \beta_r + e_r$ y el estimador de β_r que se deduce del mismo que es $\tilde{\beta} = (X_r^t X_r)^{-1} X_r^t y_r$. Ahora se considera el vector de los residuos del modelo $\tilde{e}_r = y_r - X_r \tilde{\beta}$ y se tiene que

$$\tilde{e}_r^t \tilde{e}_r = y_r^t P_{r1} y_r, \quad \text{donde} \quad P_{r1} = I_D - X_r (X_r^t X_r)^{-1} X_r^t$$

Ahora se plantea la relación

$$\tilde{e}_r^t \tilde{e}_r = E (\tilde{e}_r^t \tilde{e}_r); \quad \tilde{e}_r^t \tilde{e}_r = \sigma_u^2 \text{tr} (P_{r1} Z_r Z_r^t) + \text{tr} (P_{r1} V_{er})$$

A continuación, despejamos en la relación anterior y nos queda

$$\tilde{\sigma}_u^2 = \frac{\tilde{e}_r^t \tilde{e}_r - \text{tr} (P_{r1} V_{er})}{\text{tr} (P_{r1} Z_r Z_r^t)}$$

La expresión anterior puede dar como resultado un número negativo, en tal caso se le asigna el valor nulo; por ello, el estimador, finalmente, se define de la siguiente forma

$$\tilde{\sigma}_u^2 = \max \{0, \tilde{\sigma}_u^2\}$$

Método Henderson 3

Se puede obtener una estimación del parámetro σ_{ur}^2 utilizando el método Henderson 3 como se indica a continuación.

$$E [SSR(u_r | \beta_r)] = E [SSR(\beta_r, u_r)] - E [SSR(\beta_r)] = \sigma_0^2 (\text{rango}(X_r Z_r) - \text{rango}(X_r)) + \text{tr} \{P_2\} \sigma_{ur}^2$$

donde

$$P_2 = \text{tr} \left\{ Z_r^t W_r \left(W_r^{-1} - X_r (X_r^t W_r X_r)^{-1} X_r^t \right) W_r Z_r \right\}$$

Ahora se obtiene una expresión para $SSR(u_r | \beta_r)$.

$$SSR(u_r | \beta_r) = SSE(\beta_r) - SSE(\beta_r, u_r) = y_r^t M_1 y_r - y_r^t M y_r,$$

donde

$$M_1 = W_r - W_r X_r (X_r^t W_r X_r)^{-1} X_r^t W_r$$

y

$$M = W_r - W_r X (X^t W_r X)^{-1} X^t W_r, \quad X = (X_r \ Z_r).$$

Se despeja el parámetro σ_{ur}^2 en la expresión y se obtiene el siguiente estimador

$$\hat{\sigma}_{urH}^2 = \frac{y_r^t (M_1 - M) y_r - \sigma_0^2 (\text{rango}(X_r \ Z_r) - \text{rango}(X_r))}{\text{tr}\{P_2\}}$$

donde

$$\begin{aligned} Q_2 &= (X^t V_{er}^{-1} X)^{-1} = \left(\sum_{d=1}^D \sigma_{dr}^{-2} x_{dr}^t x_{dr} \right)^{-1}, \\ P_2 &= V_{er}^{-1} - V_{er}^{-1} X Q_2 X^t V_{er}^{-1} = \text{diag}(\sigma_{edr}^{-2}) - \text{col}_{1 \leq d \leq D}(\sigma_{edr}^{-2} x_{dr}) Q_2 \text{col}'_{1 \leq d \leq D}(x_{dr}^t \sigma_{edr}^{-2}), \\ \text{tr}\{P_2\} &= \sum_{d=1}^D \sigma_{edr}^{-2} - \sum_{d=1}^D \sigma_{edr}^{-4} \text{tr}\{x_{dr}^t x_{dr} Q_2\}, \\ y_r^t P_2 y_r &= \text{col}'_{1 \leq d \leq D}(y_{dr}) \left[\text{diag}(\sigma_{edr}^{-2}) - \text{col}_{1 \leq d \leq D}(\sigma_{edr}^{-2} x_{dr}) Q_2 \text{col}'_{1 \leq d \leq D}(x_{dr}^t \sigma_{edr}^{-2}) \right] \text{col}_{1 \leq d \leq D}(y_{dr}) \\ &= \sum_{d=1}^D \sigma_{edr}^{-2} y_{dr}^2 - \left(\sum_{d=1}^D y_{dr} \sigma_{edr}^{-2} x_{dr} \right) Q_2 \left(\sum_{d=1}^D y_{dr} \sigma_{edr}^{-2} x_{dr} \right)^t \\ \sigma_{edr}^2 &= \sigma_0^2 W_{dr}^{-1} \end{aligned}$$

La expresión anterior puede dar como resultado un número negativo, en tal caso se le asigna el valor nulo; por ello, el estimador, finalmente, se define de la siguiente forma

$$\hat{\sigma}_{urH}^2 = \text{máx}\{0, \hat{\sigma}_{urH}^2\}$$



B

Modelo para el experimento de simulación

Con el objeto principal de realizar los experimentos de simulación apropiados para cada uno de los modelos planteados en los capítulos 3-5 del presente trabajo.

Como se trata de experimentos iniciales cuyo objetivo consiste en verificar los métodos de estimación utilizados se considerará un número reducido de variables y parámetros.

En el modelo se considerará $R = 2$, $p_1 = p_2 = 1$, $p = 2$ y para el número de áreas se considerarán los siguientes valores $D = 50, 100, 200, 400$. La forma matricial del modelo después de tener en cuenta lo anterior es el siguiente

$$\begin{pmatrix} y_{11} \\ y_{12} \\ y_{21} \\ y_{22} \\ \vdots \\ y_{D1} \\ y_{D2} \end{pmatrix} = \begin{pmatrix} x_{11} & 0 \\ 0 & x_{12} \\ x_{21} & 0 \\ 0 & x_{22} \\ \vdots & \\ x_{D1} & 0 \\ 0 & x_{D2} \end{pmatrix} \begin{pmatrix} \beta_1 \\ \beta_2 \end{pmatrix} + \begin{pmatrix} u_{11} \\ u_{12} \\ u_{21} \\ u_{22} \\ \vdots \\ u_{D1} \\ u_{D2} \end{pmatrix} + \begin{pmatrix} e_{11} \\ e_{12} \\ e_{21} \\ e_{22} \\ \vdots \\ e_{D1} \\ e_{D2} \end{pmatrix}$$

Las variables explicativas que se van a considerar se construyen de la forma siguiente

$$\begin{aligned} U_{dr} &= \frac{d-D}{D} + \frac{r}{R+1} \\ x_{d1} &= \mu_1 + \sigma_{x11}^{1/2} U_{d1} \\ x_{d2} &= \mu_2 + \sigma_{x22}^{1/2} \left(\rho_x U_{d1} + \sqrt{1 - \rho_x^2} U_{d2} \right) \end{aligned}$$

donde $\mu_1 = \mu_2 = 10$, $\theta_{x11} = 1$, $\sigma_{x22} = 2$ y $\rho_x = 0,2$

Además de las variables explicativas se necesitan los efectos y los errores aleatorios para poder generar la variable explicativa. Por ello, se realizarán las siguientes consideraciones sobre los mismos.

Para $d = 1, \dots, D$

$$u_d \sim N_2(0, V_{ud}), \quad \begin{pmatrix} e_{d1} \\ e_{d2} \end{pmatrix} \sim N_2\left(\begin{pmatrix} 0 \\ 0 \end{pmatrix}, V_{ed} = \begin{pmatrix} \sigma_{d11} & \sigma_{d12} \\ \sigma_{d12} & \sigma_{d22} \end{pmatrix}\right),$$

donde $\sigma_{d11} = 1$, $\sigma_{d22} = 2$, $\sigma_{d12} = \rho_e \sqrt{\sigma_{d11}\sigma_{d22}}$, $\rho_e = 1/2$ y la matriz V_{ud} depende de un vector de parámetros $\theta = (\theta_1, \dots, \theta_m)$ que en principio son desconocidos.

Finalmente, se considera $\beta_1 = \beta_2 = 1$ y se tiene que la variable objeto de estudio es

$$y_{dr} = x_{dr}\beta_r + u_{dr} + e_{dr}.$$

La matriz de covarianzas del modelo es la siguiente

$$V = \text{var}(y) = Z'V_u Z + V_e = V_u + V_e = \text{diag}(V_d)_{1 \leq d \leq D}$$

donde $V_d = V_{ud} + V_{ed}$. La matriz V_{ud} dependerá de un vector de parámetros θ . La dimensión del vector θ será a lo sumo 3, ya que la matriz V_{ud} siempre será de orden 2.

C

Código R para realizar estimaciones

En todo el trabajo que se ha realizado se ha empleado el programa R para obtener los resultados de las simulaciones planteadas y de las aplicaciones a datos de una muestra real; a modo de ilustración se presenta en este apéndice el código más representativo que se ha elaborado para conseguir los fines que se acaban de citar.

Con el propósito de no prolongar en exceso se ha optado por el modelo de varianzas diagonal, ya que en los restantes las diferencias se limitan al cambio de modelo y nada aportaría a las estrategias de programación seguidas.

Función REMLarea.modelo1

La función **REMLarea.modelo1** se define mediante la instrucción

$$\text{REMLarea.modelo1} < -\text{function}(Xd, yd, D, R, \text{sigmau}, \text{sigmae}, \text{MAXITER}).$$

Se requiere cuando se precisa calcular estimadores utilizando el método de la máxima verosimilitud residual. Los argumentos de la función son los siguientes

Xd: lista que contiene las matrices $R \times R$ con los valores de las variables explicativas.

yd: lista que contiene los vectores $R \times 1$ con los valores de las variables objeto de estudio.

D: número de áreas.

R: número de variables que son objeto de estudio.

sigmau: vector con las varianzas de los efectos.

sigmae: vector con las varianzas de los errores.

MAXITER: número de iteraciones máxima para el algoritmo.

La función devuelve una lista de tres elementos. El primero de ellos es el vector **theta.f** que contiene las estimaciones REML de los parámetros, el segundo es la matriz **Fsig** que contiene la estimación de la matriz de la información de Fisher y el tercero es la matriz **Q** que aparece en el cálculo de los estimadores de β .

```
REMLarea.modelo1 <- function(Xd, yd, D, R, sigmau, sigmae, MAXITER){
  theta.f <- c(sigmau[1],sigmau[2])
  Vd <- matrix(0,nrow=R, ncol=R)
  for(ITER in 1:MAXITER){
    Vd.inv<-Vda<-Vdb<-VinvVda<-VinvVdb<-Vinvyd<-VinvXd<-list()
    XtVinvVdaVinvX<-XtVinvVdbVinvX<-VinvVdaVinvVda<-VinvVdaVinvVda<-VinvVdaVinvVda<-list()
    XtVinvVdaVinvVdaVinvX<-VinvVdbVinvVdb<-VinvVdaVinvVdb<-XtVinvVdbVinvVdbVinvX<-list()
    XtVinvVdaVinvVdbVinvX <- list()
    Q.inv <- matrix(0, nrow=R, ncol=R)
    tr.VinvVda<-tr.VinvVdb<-tr.VinvVdaVinvVda<-tr.VinvVdbVinvVdb<-tr.VinvVdaVinvVdb<-0
    ytVinvX<-ytVinvVdaVinvy<-SumXtVinvVdaVinvX<-ytVinvVdaVinvX<-ytVinvVdbVinvy<-0
    ytVinvVdbVinvX <- SumXtVinvVdbVinvX <- 0
    Vd <- diag(theta.f)+matrix(sigmae,nrow=2,ncol=2)
    Vd.inv[[1]] <- solve(Vd)
    for(d in 1:D){
      ### Cálculos matriciales para Sa, Sb, Faa, Fbb y Fab
      ### Derivadas de la matriz de covarianzas del modelo
      Vda[[d]]<-diag(c(1,0))
      Vdb[[d]]<-diag(c(0,1))
      Vd.inv[[d]] <- Vd.inv[[1]]
      Vinvyd[[d]] <- Vd.inv[[d]]%*%yd[[d]]
      VinvXd[[d]] <- Vd.inv[[d]]%*%Xd[[d]]
      ### Elaboración de la matriz Q
      Q.inv <- Q.inv + t(Xd[[d]])%*%VinvXd[[d]]
      ### Sa
      VinvVda[[d]] <- Vd.inv[[d]]%*%Vda[[d]]
      tr.VinvVda <- tr.VinvVda + sum(diag(VinvVda[[d]]))
      XtVinvVdaVinvX[[d]] <- t(VinvXd[[d]])%*%Vda[[d]]%*%VinvXd[[d]]
      ytVinvX <- ytVinvX + t(yd[[d]])%*%VinvXd[[d]]
      ytVinvVdaVinvy <- ytVinvVdaVinvy + t(Vinvyd[[d]])%*%Vda[[d]]%*%Vinvyd[[d]]
    }
  }
}
```

```

ytVinvVdaVinvX <- ytVinvVdaVinvX + t(Vinvyd[[d]])%*%Vda[[d]]%*%VinvXd[[d]]
SumXtVinvVdaVinvX <- SumXtVinvVdaVinvX + XtVinvVdaVinvX[[d]]
### Sb
VinvVdb[[d]] <- Vd.inv[[d]]%*%Vdb[[d]]
tr.VinvVdb <- tr.VinvVdb + sum(diag(VinvVdb[[d]]))
XtVinvVdbVinvX[[d]] <- t(VinvXd[[d]])%*%Vdb[[d]]%*%VinvXd[[d]]
ytVinvVdbVinvy <- ytVinvVdbVinvy + t(Vinvyd[[d]])%*%Vdb[[d]]%*%Vinvyd[[d]]
ytVinvVdbVinvX <-ytVinvVdbVinvX + t(Vinvyd[[d]])%*%Vdb[[d]]%*%VinvXd[[d]]
SumXtVinvVdbVinvX <- SumXtVinvVdbVinvX + XtVinvVdbVinvX[[d]]
### Faa
VinvVdaVinvVda[[d]]<-Vd.inv[[d]]%*%Vda[[d]]%*%Vd.inv[[d]]%*%Vda[[d]]
tr.VinvVdaVinvVda<-tr.VinvVdaVinvVda+sum(diag(VinvVdaVinvVda[[d]]))
XtVinvVdaVinvVdaVinvX[[d]]<-t(VinvXd[[d]])%*%Vda[[d]]%*%Vd.inv[[d]]%*%Vda[[d]]
%*%VinvXd[[d]]
###Fbb
VinvVdbVinvVdb[[d]]<-Vd.inv[[d]]%*%Vdb[[d]]%*%Vd.inv[[d]]%*%Vdb[[d]]
tr.VinvVdbVinvVdb<-tr.VinvVdbVinvVdb+sum(diag(VinvVdbVinvVdb[[d]]))
XtVinvVdbVinvVdbVinvX[[d]]<-t(VinvXd[[d]])%*%Vdb[[d]]%*%Vd.inv[[d]]%*%Vdb[[d]]
%*%VinvXd[[d]]
###Fab
VinvVdaVinvVdb[[d]]<-Vd.inv[[d]]%*%Vda[[d]]%*%Vd.inv[[d]]%*%Vdb[[d]]
tr.VinvVdaVinvVdb<-tr.VinvVdaVinvVdb+sum(diag(VinvVdaVinvVdb[[d]]))
XtVinvVdaVinvVdbVinvX[[d]]<-t(VinvXd[[d]])%*%Vda[[d]]%*%Vd.inv[[d]]%*%Vdb[[d]]
%*%VinvXd[[d]]
}
Q<-solve(Q.inv)
tr.XtVinvVdaVinvXQ<-tr.XtVinvVdbVinvXQ<-tr.XtVinvVdaVinvVdaVinvXQ
<-tr.XtVinvVdbVinvVdbVinvXQ<-0
tr.XtVinvVdaVinvVdbVinvXQ<-tr.XtVinvVdbVinvVdbVinvXQ<-XtVinvVdaVinvXQ
<-XtVinvVdbVinvXQ<-0
for(d in 1:D)
{
tr.XtVinvVdaVinvXQ <- tr.XtVinvVdaVinvXQ + sum(diag(XtVinvVdaVinvX[[d]]%*%Q))
tr.XtVinvVdbVinvXQ <- tr.XtVinvVdbVinvXQ + sum(diag(XtVinvVdbVinvX[[d]]%*%Q))
tr.XtVinvVdaVinvVdaVinvXQ<-tr.XtVinvVdaVinvVdaVinvXQ
+sum(diag(XtVinvVdaVinvVdaVinvX[[d]]%*%Q))
XtVinvVdaVinvXQ<-XtVinvVdaVinvXQ+XtVinvVdaVinvX[[d]]%*%Q
tr.XtVinvVdbVinvVdbVinvXQ<-tr.XtVinvVdbVinvVdbVinvXQ
+sum(diag(XtVinvVdbVinvVdbVinvX[[d]]%*%Q))
XtVinvVdbVinvXQ<-XtVinvVdbVinvXQ+XtVinvVdbVinvX[[d]]%*%Q
tr.XtVinvVdaVinvVdbVinvXQ<-tr.XtVinvVdaVinvVdbVinvXQ
+sum(diag(XtVinvVdaVinvVdbVinvX[[d]]%*%Q))
}

```

```

}
tr.XtVinvVdaVinvXQXtVinvVdaVinvXQ<-sum(diag(XtVinvVdaVinvXQ%%XtVinvVdaVinvXQ))
tr.XtVinvVdbVinvXQXtVinvVdbVinvXQ<-sum(diag(XtVinvVdbVinvXQ%%XtVinvVdbVinvXQ))
tr.XtVinvVdaVinvXQXtVinvVdbVinvXQ<-sum(diag(XtVinvVdaVinvXQ%%XtVinvVdbVinvXQ))
tr.PVa<-tr.VinvVda-tr.XtVinvVdaVinvXQ
tr.PVb<-tr.VinvVdb-tr.XtVinvVdbVinvXQ
tr.PVaPVa<-tr.VinvVdaVinvVda-2*tr.XtVinvVdaVinvVdaVinvXQ
+tr.XtVinvVdaVinvXQXtVinvVdaVinvXQ
tr.PVbPVb<-tr.VinvVdbVinvVdb-2*tr.XtVinvVdbVinvVdbVinvXQ
+tr.XtVinvVdbVinvXQXtVinvVdbVinvXQ
tr.PVaPVb<-tr.VinvVdaVinvVdb-2*tr.XtVinvVdaVinvVdbVinvXQ
+tr.XtVinvVdaVinvXQXtVinvVdbVinvXQ
ytPVaPy<-ytVinvVdaVinvy-ytVinvVdaVinvX%%Q%%t(ytVinvX)-ytVinvX%%Q%%t(ytVinvVdaVinvX)
+ytVinvX%%Q%%SumXtVinvVdaVinvX%%Q%%t(ytVinvX)
ytPVbPy<-ytVinvVdbVinvy-ytVinvVdbVinvX%%Q%%t(ytVinvX)-ytVinvX%%Q%%t(ytVinvVdbVinvX)
+ytVinvX%%Q%%SumXtVinvVdbVinvX%%Q%%t(ytVinvX)
### Vector de las puntuaciones y matriz de la información de Fisher
Sa <- -0.5*tr.PVa + 0.5*ytPVaPy
Sb <- -0.5*tr.PVb + 0.5*ytPVbPy
Faa <- 0.5*tr.PVaPVa
Fbb <- 0.5*tr.PVbPVb
Fab <- 0.5*tr.PVaPVb
Ssig <- c(Sa,Sb)
Fsig <- matrix(c(Faa,Fab,Fab,Fbb), nrow=2, ncol=2)
### Algoritmo de Fisher-Scoring
Fsig.inv <- solve(Fsig)
dif <- Fsig.inv%%as.matrix(Ssig)
theta.f <- theta.f + as.vector(dif)
### Regla de parada del algoritmo (Si no se han producido diferencias significativas
### respecto de la iteración anterior se interrumpe el proceso)
if(abs(dif[1,1])<0.000001 && abs(dif[2,1])<0.000001)
break
}
return(list(as.vector(theta.f), Fsig, Q))
}

```

Función Betaarea.modelo1

Se requiere cuando se precisa calcular las estimaciones EBLUP del parámetro β del modelo. Los argumentos de la función son los siguientes

Xd: lista que contiene las matrices $R \times R$ con los valores de las variables explicativas.

yd: lista que contiene los vectores $R \times 1$ con los valores de las variables objeto de estudio.

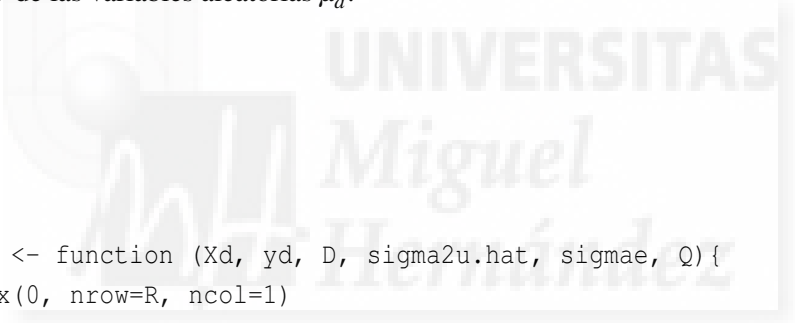
D: número de áreas.

sigma2u.hat: vector con las estimaciones REML de σ_u .

sigmae: vector con las varianzas de los errores.

Q: matriz $R \times R$ que aparece en el cálculo de los estimadores de β .

La función devuelve una lista de dos elementos. El primero de ellos es el vector **beta.hat** que contiene las estimaciones REML de los parámetros β del modelo, el segundo es la lista **mud.hat** que contiene las estimaciones EBLUP de las variables aleatorias μ_d .



```
Betaarea.modelo1 <- function (Xd, yd, D, sigma2u.hat, sigmae, Q){
XtVinvy <- matrix(0, nrow=R, ncol=1)
beta.hat <- matrix(0, nrow=2, ncol=1)
u.hat <- mud.hat <- list()
Vd <- diag(sigma2u.hat)+matrix(sigmae,nrow=2,ncol=2)
Vd.inv[[1]] <- solve(Vd)
for (d in 1:D)
{
Vd.inv[[d]] <- Vd.inv[[1]]
XtVinvy <- XtVinvy+t(Xd[[d]])%*%Vd.inv[[d]]%*%yd[[d]]
}
beta.hat <- Q%*%XtVinvy
for (d in 1:D)
{
u.hat[[d]] <- diag(sigma2u.hat)%*%Vd.inv[[d]]%*%(yd[[d]]-Xd[[d]]%*%beta.hat)
mud.hat[[d]] <- Xd[[d]]%*%beta.hat+u.hat[[d]]
}
return(list(beta.hat, mud.hat))
}
```

Función MSEarea.modelo1

La función **MSEarea.modelo1** se define mediante la instrucción

$$MSEarea.modelo1 <- function(Xd, D, R, sigma2u.hat, sigmae, Q, Fsig.inv).$$

Se requiere cuando se precisa calcular la estimación del error cuadrático medio de $\hat{\mu}_d$. Los argumentos de la función son los siguientes

Xd: lista que contiene las matrices $R \times R$ con los valores de las variables explicativas.

D: número de áreas.

R: número de variables que son objeto de estudio.

sigma2u.hat: vector con las estimaciones REML del parámetro σ_u^2 .

sigmae: vector con las varianzas de los errores.

Q: matriz $R \times R$ que aparece en el cálculo de las estimaciones del parámetro β del modelo.

Fsig.inv: inversa de la matriz estimada de la matriz de la información de Fisher.

La función devuelve una lista de tres elementos que son las listas **g1.hat**, **g2.hat** y **g3.hat** que contienen, respectivamente, las estimaciones de G_1 , G_2 y G_3 que aparecen en la fórmula del error cuadrático medio de $\hat{\mu}_d$.

```
MSEarea.modelo1 <- function(Xd, D, R, sigma2u.hat, sigmae, Q, Fsig.inv){
  g1.hat <- g2.hat <- g3.hat <- mse.mud.hat <- list()
  Wd <- list()
  Vd <- Vud <- Vd.inv <- Ved.inv <- matrix(0, nrow=R, ncol=R)
  Td <- Ldi <- Ldj <- Sum <- matrix(0, nrow=R, ncol=R)
  Ved <- matrix(sigmae, nrow=R, ncol=R)
  Vud <- diag(sigma2u.hat)
  Vd <- Vud+Ved
  Vd.inv <- solve(Vd)
  Ved.inv <- solve(Ved)
  Wd[[1]] <- diag(c(1,0))
```

```

Wd[[2]] <- diag(c(0,1))
# En el experimento de simulación la matriz Td no varía.
# Las matrices de covarianzas son las mismas en cada una de las áreas.
# Por ello, sólo es necesario realizar una asignación
Td <- Vud-Vud%%Vd.inv%%Vud
# La matriz g1 estimada será siempre la misma por lo apuntado en el comentario anterior
# Cálculo de la matriz g2 estimada
for (d in 1:D){
g1.hat[[d]] <- Td
g2.hat[[d]] <- (Xd[[d]]-Td%%Ved.inv%%Xd[[d]])%%Q%%t(Xd[[d]]-Td%%Ved.inv%%Xd[[d]])
Sum <- matrix(0,R,R)
for (i in 1:R){
for (j in 1:R){
Ldi <- Wd[[i]]%%Vd.inv-Vud%%Vd.inv%%Wd[[i]]%%Vd.inv
Ldj <- Wd[[j]]%%Vd.inv-Vud%%Vd.inv%%Wd[[j]]%%Vd.inv
Sum <- Sum + Fsig.inv[i,j]*Ldi%%Vd%%t(Ldj)
}
}
g3.hat[[d]] <- Sum
mse.mud.hat[[d]] <- g1.hat[[d]]+g2.hat[[d]]+2*g3.hat[[d]]
}
return(list(mse.mud.hat,g1.hat,g2.hat,g3.hat))
}

```



Bibliografía

- Arnold, S. (1981). *The Theory of Linear Models and Multivariate Analysis*. John Wiley, New York.
- Battese, G. E., Harter, R. M. and Fuller, W. A. (1988). An error component model for prediction of county crop areas using survey and satellite data. *Journal of the American Statistical Association*, **83**, 28–36.
- Beckman, R.J., Nachtsheim, C.J., and Cook R.D. (1987). Diagnostics for Mixed-Model Analysis of Variance. *Technometrics*, **29**, 413–426.
- Belsley, D.A., Kuh, E., and Welsh, R.E. (1980). *Regression Diagnostics: Identifying Influential Data and Sources of Collinearity*. John Wiley, New York.
- Choudry, G.H., Rao, J.N.K. (1989). Small area estimation using models that combine time series and cross sectional data, in: Singh, A.C., Whitridge, P. (Eds), Proceedings of Statistics Canada Symposium on Analysis of Data in Time, pp. 67-74.
- Das, K., Jiang, J. and Rao, J. N. K. (2004). Mean squared error of empirical predictor. *The Annals of Statistics*, **32**, 818-840.
- Datta, G. S., Fay, R. E. and Ghosh, M. (1991). Hierarchical and Empirical Bayes Multivariate Analysis in Small Area Estimation. In: Proceedings of Bureau of the Census 1991 Annual Research Conference, U. S. Bureau of the Census, Washington, DC, 63-79.
- Datta, G. S., Ghosh, M., Nangia, N., and Natarajan, K. (1996). Estimation of median income of four-person families: a Bayesian approach. In: D. A. Berry, K. M. Chaloner and J. M. Geweke (Eds.), *Bayesian Analysis in Statistics and Econometrics*, Wiley, New York, 129-140.
- Datta, G.S., Lahiri, P., Maiti, T., Lu, K.L. (1999). Hierarchical Bayes estimation of unemployment rates for the U.S. states. *Journal of the American Statistical Association*, **94**, 1074-1082.

- Datta G.S., Lahiri P. (2000). A unified measure of uncertainty of estimated best linear unbiased predictors in small area estimation problems. *Statistica Sinica*, **10**, 613-627.
- Datta, G.S., Lahiri, P., Maiti, T. (2002). Empirical Bayes estimation of median income of four-person families by state using time series and cross-sectional data. *Journal of Statistical Planning and Inference*, **102**, 83-97.
- Datta, G., Kubokawa, T., Molina, I., Rao, J.N.K. (2011). Estimation of mean squared error of model-based small area estimators. *TEST*, **20**, 2, 367-388.
- Dick, P. (1995). Modelling net undercoverage in the 1991 Canadian census. *Survey Methodology*, **21**, 45-54.
- Dick, P. (1995). Modelling net undercoverage in the 1991 Canadian census, *Survey Methodology*, **21**, 45-54.
- Ericksen, E. P. and Kadane, J. B. (1985). Estimating the population in census year: 1980 and beyond (with discussion). *Journal of the American Statistical Association*, **80**, 98-131.
- Esteban, M.D., Morales, D., Pérez, A., Santamaría, L. (2011). Two Area-Level Time Models for Estimating Small Area Poverty Indicators. *Journal of the Indian Society of Agricultural Statistics*, **66**(1), 75-89.
- Esteban M.D., Morales D., Pérez A., Santamaría L. (2012). Small area estimation of poverty proportions under area-level time models. *Computational Statistics and Data Analysis*, **56**, 2840-2855.
- Fay, R. E. and Herriot, R. A. (1979). Estimates of income for Small Places: An Application of James-Stein Procedures to Census Data. *Journal of the American Statistical Association*, **74**, 366, 269-277.
- Fay, R. E. (1987). Application of Multivariate Regression of Small Domain Estimation. In: R. Platek, J. N. K. Rao, C. E. Särndal and M. P. Singh (Eds.), *Small Area Statistics*, Wiley, New York, 91-102.
- Ghosh, M. and Rao, J.N.K. (1994). Small area estimation: An appraisal. *Statistical Science*, **9**, 55-93.
- Ghosh, M., Nangia, N., Kim, D. (1996). Estimation of median income of four-person families: a Bayesian time series approach. *Journal of the American Statistical Association*, **91**, 1423-1431.
- González-Manteiga, W., Lombardía, M.J., Molina, I., Morales, D. y Santamaría, L. (2008). Analytic and Bootstrap Approximations of Prediction Errors under a Multivariate Fay-Herriot Model. *Computational Statistics and Data Analysis*, **52**, 5242-5252.
- González-Manteiga, W., Lombardía, M.J., Molina, I., Morales, D. y Santamaría, L. (2010). Small area estimation under Fay-Herriot models with nonparametric estimation of heteroscedasticity. *Statistical Modelling*, **10**, 2, 215-239.
- Herrador, M., Esteban, M.D., Hobza, T., Morales D. (2011). A Fay-Herriot model with different random effect variances *Communications in Statistics (Theory and Methods)*, **40**, 5, 785-797.

- Jiang, J. and Lahiri, P. (2006). Mixed model prediction and small area estimation. *Test*, **15**, 1–96.
- Jiang, J. (2007). *Linear and Generalized Linear Mixed Models and their Applications*. Springer-Verlag, New York.
- Jiang, J., Nguyen, T., Rao, J.S. (2011). Best Predictive Small Area Estimation. *Journal of the American Statistical Association*, **106**, 494, 732-745.
- Kackar, R. and Harville, D. A. (1981). Unbiasedness of two-stage estimation and prediction procedures for mixed linear models. *Communications in Statistics-Theory and Methods*, Ser. A, **10**, 1249-1261.
- Kackar, R. and Harville, D. A. (1984). Approximations for standard errors of estimators of fixed and random effects in mixed linear models. *Journal of the American Statistical Association*, **79**, 853–862.
- Kubokawa, T. (2011). On measuring uncertainty of small area estimators *Journal of the Japan Statistical Society*, **41**, 2, 93-119.
- Marhuenda, Y., Molina, I. and Morales, D. (2013). Small area estimation with spatio-temporal Fay-Herriot models. *Computational Statistics and Data Analysis*, **58**, 1, 308-325.
- Morales, D., Pagliarella, M.C., Salvatore R. (2015). Small area estimation of poverty indicators under partitioned area-level time models. *SORT-Statistics and Operations Research Transactions*, **39**, 1, 19-34, 2015.
- McCulloch, C.E. and Searle, S.R. (2001). *Generalized, Linear and Mixed Models*. John Wiley, New York.
- Molina, I. and Morales D. (2009) Small area estimation of poverty indicators. *BEIO*, **25**, 3, 218-225.
- National Research Council (2000). *Small-area estimates of school-age children in poverty: Evaluation of current methodology*. C.F. Citro and G. Kalton (Eds.), Committee on National Statistics, Washington DC: National Academy Press.
- Neter, J., Kutner, M.H., Nachtsheim, C.J., and Wasserman, W. (1990). *Applied Linear Statistical Models*. IRWIN, Chicago.
- Pfeffermann, D., Burck, L. (1990). Robust small area estimation combining time series and cross-sectional data. *Survey Methodology*, **16**, 217-237.
- Pfeffermann, D. and Buck, L. (2002). Small Area Estimation - New Developments and Directions. *International Statistical Review*, **70**, 125-143.
- Pfeffermann, D. and Tiller, R. (2005). Bootstrap approximation to prediction MSE for state-space models with estimated parameters. *Journal of Time Series Analysis*, **26**, 893–916.
- Pfeffermann, D. (2013). New important developments in small area estimation. *Statistical Science*, **28**, 40-68.

- Prasad, N. G. N. and Rao, J. N. K. (1990). The estimation of the mean squared error of small-area estimators. *Journal of the American Statistical Association*, **85**, 163–171.
- Rao, C.R. (1973). *Linear Statistical Inference and its Applications*. John Wiley, New York.
- Rao, J.N.K., Yu, M. (1994). Small area estimation by combining time series and cross sectional data. *Canadian Journal of Statistics*, **22**, 511-528.
- Rao, J.N.K. (1999). Some recent advances in model-based small area estimation. *Survey Methodology*, **25**, 175-186.
- Rao, J.N.K. (2003). *Small area estimation*. John Wiley, New York.
- Rencher, A.C. (2000). *Linear models in Statistics*. John Wiley, New York.
- Särndal, C.E., Swensson, B., and Wretman, J.H. (1992). *Model Assisted Survey Sampling*. Springer-Verlag, New York
- Searle, S.R. (1971). *Linear Models*. John Wiley, New York.
- Searle, S.R. (1997). Built-in restrictions on best linear unbiased predictors (BLUP) of random effects in mixed models. *The American Statistician*, **51**, 19—21.
- Searle, S.R., Casella, G. and McCulloch, C.E. (1992). *Variance components*. John Wiley, New York.
- Seber, G.A.F. (1977). *Linear Regression Analysis*. John Wiley, New York.
- Singh, B., Shukla, G., Kundu, D. (2005). Spatio-temporal models in small area estimation. *Survey Methodology*, **31**, 183-195.
- Slud, E.V., Maiti, T. (2011). Small-area estimation based on survey data from left censored Fay-Herriot models *Journal of Statistical Planning and Inference*, **141**, 3520-3535.
- Ybarra, L.M.R., Lohr, S.L.(2008). Small area estimation when auxiliary information is measured with error. *Biometrika*, **95**, 4, 919-931.
- You, Y., Rao, J.N.K. (2000). Hierarchical Bayes estimation of small area means using multi-level models. *Survey Methodology*, **26**, 173-181.