Traballo Fin de Máster

Estimación do MSE en modelos de área con efecto temporal

Paula Lois Alfonsín

Máster en Técnicas Estadísticas

Curso 2014-2015

Proposta de Traballo Fin de Máster

Título en galego: Estimación do MSE en modelos de área con efecto temporal
Título en español: Estimación del MSE en modelos de área con efecto temporal
English title: MSE estimation for area level models with time effects
Modalidade: Modalidade B
Autor/a: Paula Lois Alfonsín, Universidade de Vigo
Director/a: María José Lombardía, UDC;
Titor/a: Esther López Vizcaíno, IGE;
Breve resumo do traballo: O obxectivo deste TFM é estudar o erro cadrático medio (MSE) dun modelo de área con efecto temporal, considerando o tempo independente e o tempo correlado. O resultado deste estudo aplicarase sobre os datos do mercado de traballo, a Enquisa de Poboación Activa (EPA).

Dona María José Lombardía, profesora da UDC, dona Esther López Vizcaíno, responsable do Servizo de Difusión e Información do IGE, informan que o Traballo Fin de Máster titulado

Estimación do MSE en modelos de área con efecto temporal

foi realizado baixo a súa dirección por dona Paula Lois Alfonsín para o Máster en Técnicas Estadísticas. Estimando que o traballo está terminado, dan a súa conformidade para a súa presentación e defensa ante un tribunal.

En Santiago de Compostela, a 8 de Xullo de 2015.

A directora:

Dona María José Lombardía

A titora:

Dona Esther López Vizcaíno

A autora:

Dona Paula Lois Alfonsín

"O pensamento estadístico
será un día tan necesario para o cidadán eficiente
como a capacidade de ler e escribir."
H. G. Wells

Índice xeral

Resumo

Resumo en galego

O obxectivo deste traballo é o estudo do erro cadrático medio (MSE) dun modelo mixto de área con efectos aleatorios temporais, considerando os efectos aleatorios de tempo independentes e os efectos de tempo correlados. En particular, trataremos de dar alternativas ás aproximacións analíticas do MSE, usando métodos de remostraxe.

Tras describir tal metodoloxía, aplicarémola aos datos reais proporcionados polo Instituto Galego de Estatística (IGE). Tales datos son os do mercado de traballo que ofrece a Enquisa de Poboación Activa (EPA) nas áreas pequenas determinadas pola Clasificación Nacional de Actividades Económicas (CNAE) a dous díxitos, no período de tempo comprendido entre terceiro trimestre do ano 2009 e o cuarto trimestre do ano 2013.

Presentaremos os resultados por separado segundo os efectos de tempo sexan independentes ou correlados, comentando en todo momento as súas conclusións.

English abstract

The aim of this work is to study the mean square error (MSE) of a model area with time effects, firstly with independent time effects and then with correlated time effects. Particularly, we will try to give alternatives to an analytical expressions of MSE, using bootstrap methods.

When we have just described the development methodologies, we will aply it to the actual data provided by Galician Statistics Institute (IGE). Such data are of the labour market offering by Labour Force Survey (EPA) in the small areas provided by Spanish Classification of Economic Activities (CNAE) in the double digits, in the time period from third quarter of 2009 to fourth quarter of 2013.

We will submit each result separately as time effects are independents or correlates, commenting at all time on their conclusions.

Introdución

Esta memoria contén o Traballo de Fin de Máster (TFM) do Máster de Técnicas Estadísticas na modalidade B (segundo a súa normativa) titulado *Estimación do MSE en modelos de área con efecto temporal*; todo o traballo que imos desenvolver está feito en colaboración co Instituto Galego de Estatística -en diante IGE-. Faremos a continuación dun TFM das mesmas características presentado no curso 2011/12 no cal, dispoñendo de datos para as variables de interese dende o primeiro trimestre do ano 2009 ata o cuarto trimestre do ano 2010, estuda os estimadores do total de ocupados por actividade económica dando como medida do erro o erro cadrático medio -en adiante MSE- de *Prasad e Cao* (1990). No presente traballo traballamos con datos noutro período de tempo (dende o terceiro trimestre do ano 2009 ata o cuarto trimestre do ano 2012) e tratamos de estudar outras alternativas de aproximación ao MSE, utilizando métodos de remostraxe.

Antes, de continuar con esta introdución, precisamos adicar unhas liñas ás áreas pequenas, pois o obxectivo deste traballo é estudar o MSE dun modelo de áreas pequenas con efecto temporal para logo aplicar este estudo sobre os datos do mercado de traballo da Enquisa de Poboación Activa (EPA). En todo o traballo, aparecerá numerosas veces o termo *área, área pequena ou dominio* e simplemente fai referencia a unha subdivisión da poboación para a variable de interese. Acostúmanse definir como áreas xeográficas (concellos,distritos,$\cdots$), grupos socioeconómicos e sociodemográficos (grupos por sexo,idade,$\cdots$) ou outras subpoboacións (actividades económicas particulares,$\cdots$).

Tamén debemos ter en mente que é a CNAE. A CNAE é a Clasificación Nacional de Actividade Económicas e asigna un código a cada actividade económica das que se poden realizar; así a CNAE actual, do 2009, ten 4 díxitos. Xeralmente este código úsase en moitos formularios e impresos, tanto oficiais como a nivel de empresa. Coa CNAE pódese saber que código corresponde á actividade que se está realizando. Se consideramos a CNAE a dous díxitos, contén unha lista de actividades económicas xunto cunha codificación que vai dende 01 ao 99 (está totalmente descrita no Apéndice A). Traballaremos coa poboación dividida en cada unhas das actividades económicas da CNAE que son áreas pequenas.

A estimación en áreas pequenas (SAE) é unha parte da ciencia estadística que despertou un grande interese nos últimos anos, dada a importancia que ten para os sectores públicos e privados obter información fiable acerca dos dominios en torno aos que estas técnicas centran a súa atención. Os antecedentes da estimación en áreas pequenas son os métodos demográficos, que dende fai décadas, úsanse para a estimación da poboación en carencias que sofren os censos e patróns.

As técnicas de estimación en áreas pequenas perseguen obter estimadores de parámetros como medias, totais e proporcións, das variables poboacionais de certas áreas, facendo uso para iso de datos mostrais recollidos atendendo a un deseño mostral no que tales áreas non recibiron unha consideración específica, senón que son entidades contidas nos estratos do deseño mostral. Tamén se fai un forte uso da información auxiliar dispoñible, tanto a referente aos dominios como as unidades mostrais.

A historia dos métodos de estimación para áreas pequenas remóntase a Inglaterra no século XI e a

Canadá no século XVII (*Brackstone*, 1987). Estes primeiros métodos usaban maioritariamente datos que proviñan de diversos rexistros administrativos e censos. Coa chegada das novas técnicas metodolóxicas no desenvolvemento de enquisas, diferentes organismos gobernamentais exploraron a posibilidade de usar datos que proveñen de enquisas na produción de estadísticas asociadas a áreas pequenas. Exemplos de usos da inferencia en áreas pequenas son o de *Fay e Herriot* en 1979 (entraremos máis a fondo neste tema na Sección 2.2 do Capítulo 2); en 1988 *Battese, Harter e Fuller* estimaron a superficie de cultivo de cereais e soia en doce condados; en 1999 *Malec, Davis e Cao* estudaron o predominio do sobrepeso en adultos estadounidenses por estados.

Nos institutos de estadística españois tamén temos unha larga lista de exemplos, por exemplo no ISTAC, EUSTAT ou no IGE, fixeron, respectivamente, estimacións en áreas pequenas cos datos da Enquisa de Poboación Activa en Canarias, da Enquisa Industrial no País Vasco e do ingreso medio mensual por comarca nos fogares galegos. No INE tamén hai un grupo de traballo en áreas pequenas, que establece procedementos para facer estimacións en áreas pequenas das variables máis significativas da EPA.

As principais fontes de datos das áreas pequenas foron, e son, os censos e os rexistros administrativos, pero a falta dun aproveitamento adecuado destas fontes, así como a necesidade de coñecer unha ampla variedade de aspectos económicos e sociais de xeito frecuente, orixinaron o desenvolvemento dunha numerosa metodoloxía neste campo. Polo tanto, non fai falta xustificar máis o crecente interese existente nas estimacións das áreas pequenas.

Asociada á expresión "área pequena" está a ausencia dunha mostra significativa. Polo tanto, o primeiro problema que nos xorde neste contexto é que o tamaño de mostra para as áreas pequenas de interese que manexaremos é moi pequeno ou cero. Isto queda solventado empregando información auxiliar que teñamos dispoñible da variable de interese, que describiremos no Capítulo 1.

Existen tres grandes clases de estimadores para a avaliación dos totais ou as medias das variables en áreas pequenas. A **estimación directa** é a que estima as áreas pequenas por métodos tradicionais baseados no deseño da mostra da enquisa, como o tamaño de mostra será moi pequena provoca estimadores cunha varianza moi grande. Na **estimación indirecta** emprégase entón información auxiliar para definir os estimadores para áreas pequenas, e dentro desta estimación indirecta atopamos os baseados en **modelos explícitos**: asumen que a variabilidade entre os dominios da variable resposta pode ser explicada de forma íntegra en termos da variabilidade da información auxiliar (**modelos de efectos fixos**) ou requiren supoñer que a variabilidade específica do dominio queda sen explicar tras considerar a información auxiliar e incorpora efectos aleatorios específicos por cada dominio, que explican as variacións entre as áreas non explicadas polos efectos fixos (**modelos mixtos**). Ademais, segundo a dispoñibilidade da información auxiliar os modelos de áreas pequenas clasifícanse en **modelos a nivel de área** cando se dispón de información auxiliar só a nivel de área, que será o noso caso, ou **modelos a nivel de unidade** cando se dispón de información auxiliar sobre as unidades individuais da poboación. Da estimación en áreas pequenas dos ocupados segundo a súa actividade económica tratará o Capítulo 2, despois de, no Capítulo 1 tratar a análise exploratoria dos datos.

Usaremos dous modelos a nivel de área que usan os estimadores directos como variable resposta, incorporando aos modelos a información do deseño de mostra. Primeiramente trataremos o **modelo lineal mixto con efectos de tempo independentes** e logo o **modelo lineal mixto con efectos de tempo correlados**. E tendo en conta o modelo, obteremos os estimadores do MSE, primeiramente usando expresións analíticas e logo usando **métodos de remostraxe**, que describiremos no Capítulo 3. No Capítulo 4 faremos un estudo de simulación para, no Capítulo 5 aplicar todos os resultados aos datos reais. Finalmente, veremos conclusións do presente traballo e abordaremos algunha liña de investigación que se podería usar no futuro apoiada na presente memoria.

Dentro da estimación en áreas pequenas, ten especial relevancia a estimación do erro do estimador, máis concretamente o MSE. *Prasad e Rao*(1990) obtiveron expresións analíticas para aproximar o MSE. Pero, estas aproximacións analíticas asumen hipóteses específicas do modelo pois obtéñense baixo o modelo particular a considerar, polo tanto estas hipóteses son fortes neste sentido. Isto soluciónase usando métodos de remostraxe, neste traballo aplicaremos a metodoloxía do bootstrap paramétrico. *Molina et al.* en 2009 deron unha aproximación heurística do MSE baixo un modelo Fay-Herriot con efectos de tempo: o bootstrap paramétrico.

Finalmente mencionar, que ao final da memoria, inclúese un glosario de siglas usadas para poder consultalas en todo momento. E tamén inclúese un apéndice no que se pode ver a descrición completa da CNAE 09 (Apéndice A) e unha explicación de paquetes do *software estadístico R* usados ao longo da memoria (Apéndice B).

Capítulo 1

Descrición e análise exploratoria dos datos

Queremos estudar o erro cadrático medio dun modelo de área con efecto temporal para estimar o número de ocupados da EPA nas áreas pequenas determinadas pola CNAE a dous díxitos e por sexo. Neste nivel de descrición o tamaño de mostra é pequeno polo que necesitamos de variables auxiliares que nos axuden a estimar a variable resposta. A nosa variable resposta será o número de ocupados da EPA. As variables auxiliares son o número de afiliados á Seguridade Social -en adiante SS- e o número de contratos novos rexistrados.

Dispomos dos datos trimestrais para estas variables de interese dende o terceiro trimestre do ano 2009 ata o cuarto trimestre do ano 2012. Tales datos proveñen do IGE.

A primeira sección deste capítulo adicámola a explicar a orixe dos datos que consideraremos no noso estudo. Na segunda sección, faremos a análise exploratoria das variables previamente descritas. Para iso, primeiramente visualizaremos os diagramas de dispersión: verémolos para as variables en xeral e despois represéntanse para as variables agrupadas por sector de actividade (primario, industria, construción e servizos); e tamén as agruparemos por sexo.

1.1. Descrición dos datos

A variable obxecto de estudo é o número de ocupados da EPA, que é función de diversas enquisas. As variables auxiliares empregan o número de afiliados á SS a final de mes e o número de contratos rexistrados.

Falaremos a continuación do procedemento que segue a EPA e a obtención de datos das variables auxiliares.

1.1.1. A EPA

A EPA é a fonte principal deste traballo, a partir dela obteranse os ocupados por rama económica. Vexamos mais detalladamente en que consiste.

A EPA é unha investigación por mostraxe, continua e dirixida ás vivendas familiares que o Instituto Nacional de Estadística (INE) realiza dende 1964.

A obxectivo da EPA é coñecer a actividade económica no relativo ao seu compoñente humano. Está orientada a dar datos das principais categorías poboacionais en relación co mercado de traballo (ocupados, parados, activos, inactivos) e a obter clasificacións de estas categorías segundo diversas características. Tamén posibilita confeccionar series temporais homoxéneas de resultados.

A enquisa está dirixida á poboación que reside en vivendas familiares, é dicir, as utilizadas todo o ano ou a maior parte del como vivenda habitual ou permanente. Exclúense, por tanto, os aloxamentos colectivos como hospitais, hoteis, orfanatos, cuarteis, conventos, etc.

Algúns conceptos que temos que ter en conta para analizar a cada un dos enquisados son:

- **Activos:** É o conxunto de persoas que, nun período de referencia dado, subministran man de obra para a produción de bens e servizos económicos ou que están dispoñibles e fan xestións para incorporarse á produción. Nesta enquisa, por tanto, a poboación economicamente activa comprende toda as persoas de 16 ou máis anos que durante a semana de referencia cumpren as condicións para a súa inclusión entre as persoas ocupadas ou paradas.

- **Ocupados:** Son as persoas de 16 e máis anos que durante a semana de referencia tiveron un traballo por conta allea ou exerceron unha actividade por conta propia.

- **Parados:** Considéranse paradas todas aquelas persoas de 16 ou máis anos, que durante a semana de referencia estiveron simultaneamente:

  - Sen traballo, é dicir, que non tiveron emprego por conta allea ou por conta propia durante a semana de referencia.

  - Na procura de traballo, é dicir, que tomaron medidas concretas para buscar un traballo por conta allea ou fixeron xestións para establecerse pola súa conta durante o mes precedente.

  - Dispoñibles para traballar.

- **Inactivos:** A poboación inactiva abrangue todas as persoas de 16 ou máis anos non clasificadas como ocupadas ou paradas durante a semana de referencia.

O tipo de mostraxe utilizado é bietápico con estratificación nas unidades de primeira etapa, que están constituídas polas seccións censais. As unidades de segunda etapa son as vivendas familiares principais e os aloxamentos fixos. Nestas unidades de segunda etapa non se realiza submostraxe algunha de modo que recóllese información de todas as persoas que teñan a súa residencia habitual nas vivendas.

A comunidade autónoma galega está dividida en estratos. Os estratos constrúense seguindo un criterio demográfico, no que se asigna a cada concello o seu estrato correspondente, dependendo do tamaño do mesmo. Avalíanse preto de 8000 vivendas, que supoñen unha mostra total de 20000 persoas. Realízase a seguinte clasificación:

- **Estrato 1:** concellos capital de provincia.

- **Estrato 2:** concellos autorrepresentados, importantes en relación coa capital (no caso de Galicia son Santiago de Compostela e Vigo)

- **Estrato 3:** outros concellos autorrepresentados, importantes en relación coa capital (no caso de Galicia é Ferrol), excepto os anteriores.

- **Estrato 4:** concellos de 50.001 a 100.000 habitantes, excepto os anteriores.

- **Estrato 5:** concellos de 20.001 a 50.000 habitantes.

- **Estrato 6:** concellos de 10.001 a 20.000 habitantes.

- **Estrato 7:** concellos de 5.001 a 10.000 habitantes.
- **Estrato 8:** concellos de 2.001 a 5.000 habitantes.
- **Estrato 9:** concellos con poboación menor ou igual a 2.000 habitantes.

A EPA usa estimadores de razón aos que se lle aplican técnicas de reponderación para que se fagan coincidir os resultados que proporciona a enquisa coas proxeccións da poboación.

Unha área, área pequena ou dominio é unha subdivisión da poboación para a variable de interese. Acostúmanse definir como áreas xeográficas (concellos,distritos,$\cdots$), grupos socioeconómicos e sociodemográficos (grupos por sexo,idade,$\cdots$) ou outras subpoboacións (actividades económicas particulares,$\cdots$).

Denotaremos Y_d o número de ocupados EPA, sendo d a área pequena que indica a actividade económica á que fai referencia $d = 1, \cdots, D$; no noso estudo teremos $D = 86$ áreas pequenas. Eliminamos previamente da CNAE 09 as ramas de actividade para as cales non se dispón de datos suficientes para a estimación porque a actividade económica non ten presencia na comunidade autónoma ou porque a EPA non recolleu tales datos en algúns dos trimestres a estudar.

Denotando por S_p a submostra que pertence á provincia p, N_h a poboación de 16 anos ou máis no estrato h segundo a proxeccións demográficas e n_h o número de individuos da mostra no estrato h. O estimador que usa a EPA para determinar o número de ocupados na provincia p é o que segue:

$$\widehat{Y}_p^{EPA} = \sum_{j \in S_p} w_j \cdot y_j \tag{1.1}$$

sendo $w_j = \frac{N_h}{n_h}$ os factores de elevación empregados pola EPA que representan a cantidade de individuos que supón cada observación recollida na enquisa.

A EPA non proporciona estimadores oficiais para os dominios (lembremos que no noso caso os dominios son as ramas de actividade económica recollidas na CNAE 09 a dous díxitos). Unha expresión equivalente para representar o estimador que calcula os ocupados por dominios é a que segue:

$$\widehat{Y}_p^{EPA} = \sum_{j \in S_d} w_j \cdot y_{dj} \quad d = 1, \cdots, 86 \tag{1.2}$$

As varianzas dos estimadores para os dominios aproxímanse coa librería *survey* de **R** (ver Apéndice B no que se describen todos paquetes usados nesta memoria), que usa fórmulas elementais da varianza da suma:

$$Var(\widehat{Y}_p^{EPA}) = \sum_j w_j^2 \cdot Var(\widehat{Y}_{dj}) + 2 \sum_{i<j} w_i \cdot w_j \cdot Cov(\widehat{Y_{di}}, \widehat{Y_{dj}}) \quad d = 1, \cdots, 86 \tag{1.3}$$

$$CV(\widehat{Y}_p^{EPA}) = \frac{\sqrt{Var(\widehat{Y}_p^{EPA})}}{\widehat{Y}_p^{EPA}} \cdot 100 \quad d = 1, \cdots, 86 \tag{1.4}$$

1.1.2. Variables auxiliares

Expoñemos a continuación as variables auxiliares que nos axudarán a estimar a variable resposta "número de ocupados".

Afiliacións á SS

Explicamos na sección anterior que a EPA proporciona información periódica sobre a ocupación. As afiliacións á SS é outra fonte estadística que nos da información periodicamente sobre a ocupación, proporciona información sobre o rexistro de afiliacións e cotizacións á SS.

A afiliación ao Sistema da SS é obrigatoria para todas as persoas incluídas no campo de aplicación da SS e única para toda a vida do traballador e para todo o sistema. Un traballador afíliase cando comeza a súa vida laboral pois dáse de alta nalgún dos réximes do Sistema da SS (alta inicial). Se cesa na súa actividade será dado de baixa pero seguirá afiliado en situación de baixa laboral. Se retoma a actividade producirase unha alta (alta sucesiva a efectos estatísticos), pero non terá que afiliarse novamente, pois a afiliación é única para toda a vida do traballador.

A súa periodicidade é mensual e utilizamos os datos a último día do mes. Tendo en conta que a periodicidade da EPA é trimestral, tivemos que agrupar trimestralmente os datos da variable afiliacións á SS.

Finalmente, dicir que hai que ter en conta que, por motivos lexislativos (referidos a axudas familiares ou cambios de leis), hai persoas ocupadas segundo a EPA que non están obrigadas a cotizar á SS e viceversa: hai persoas que están obrigadas de afiliarse á SS e non o están.

Contratos rexistrados

Esta variable auxiliar recolle o número de contratos novos rexistrados mensualmente por cada un dos dominios determinados pola CNAE a dous díxitos e por sexo.

Os empresarios teñen que formalizar por escrito á Oficina Pública de Emprego que corresponda o contido dos contratos de traballo celebrados e as prórrogas deles, nos dez días seguintes á súa concertación.

No último día do mes contabilízanse os contratos introducidos nese período de tempo con independencia de cando foi presentado o contrato, polo que, nun mes determinado, poden incluírse contratos rexistrados anteriormente e que, por causas varias, non se incluíron no seu día na base de datos.

Lembremos cales son os catro sectores económicos:

- Sector primario é o que ten os produtos directamente da natureza, materias primas, creacións, etc.
- Sector industria é o que transforma materias primas en produtos acabados ou semielaborados.
- Sector construción é o que inclúe as actividades do proceso de armado dunha estrutura.
- Sector servizos é o que só produce servizos, non produce bens.

Os nosos dominios (ou áreas pequenas) son as actividades económicas segundo a CNAE a dous díxitos, tales dominios pódense clasificar segundo o sector de actividade económica ao que fan referencia: sector primario, industria, construción e servizos. Así, o sector primario inclúe en tres actividades, o sector industria inclúe 34 actividades, o sector construción 3 e o sector servizos 54 actividades.

Pódese consultar o Apéndice A para ver o cadro do CNAE 09, observamos que presenta tres columnas. Na primeira dela atopamos un número que vai dende o 01 ata o 99 que representa o código da actividade. Na segunda columna descríbese a que actividade concreta fai referencia ese código. E na última columna vemos a que sector pertence ese Código e esa actividade. Por exemplo, o código 01

corresponde á actividade económica "agricultura,gandaría caza e servizos relacionados con elas" e pertence ao sector primario, do mesmo xeito o código 02 corresponde á actividade económica "silvicultura e explotación forestal" a cal pertence ao sector primario.

1.2. Análise exploratoria

Nesta sección centrarémonos en presentar unha análise exploratoria das variables. Presentaremos gráficas e cadros como procedemento previo ao vindeiro capítulo que trata dos modelos que traten de explicar o comportamento do número de ocupados da EPA.

Estudamos seguidamente os diagramas de dispersión para as variables na súa totalidade. Tamén veremos os diagramas de dispersión para as variables agrupadas polos catro sectores económicos dos que falamos anteriormente, e os diagramas de dispersión agrupados por sexo: home e muller. Ademais, estudaremos a normalidade das variables, para iso presentaremos gráficas cuantil-cuantil, histogramas e resultados dos tests de normalidade.

Na Figura 1.1 podemos ver os diagramas de dispersión dos "ocupados EPA" en función do "número de afiliados á SS" e do " número de contratos". Tras observar onde se sitúan os puntos, vemos que non existe unha clara relación lineal entre a variable resposta e as variables auxiliares.

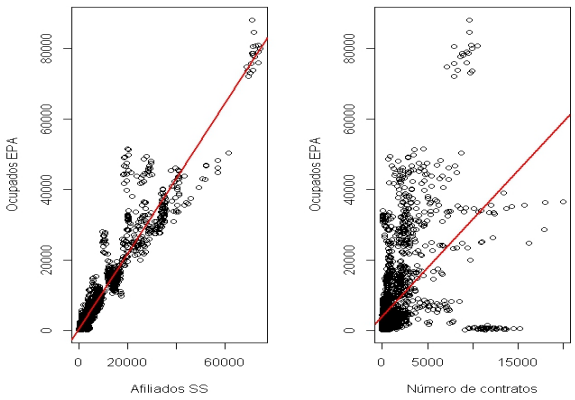

Figura 1.1: Diagrama de dispersión "ocupados EPA" en función do "nº afiliados SS" e do "nº de contratos"

Dada a importancia da hipótese básica da normalidade da variable resposta ("ocupados EPA"), faremos de seguido un estudo da mesma.

Represéntase na Figura 1.2 os gráficos cuantil-cuantil correspondentes á variable resposta na súa forma orixinal e na súa forma logarítmica e tamén visualizaremos o seus histogramas, ademais mostraremos os resultados dos tests de Shapiro-Wilks e Kolmogorov-Smirnov-Lilliefors na Táboa 1.2.

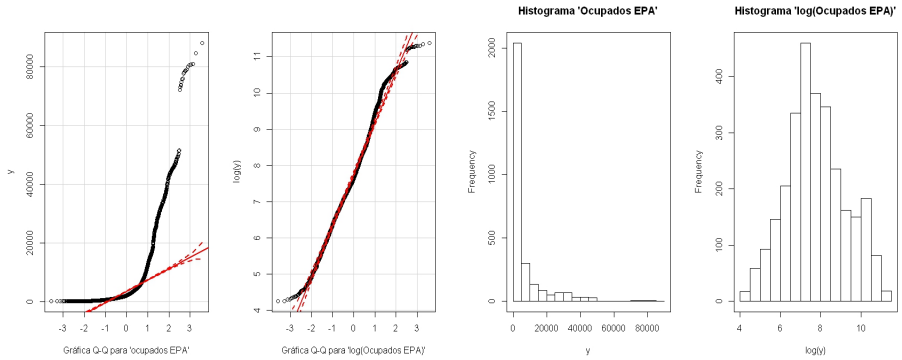

Figura 1.2: QQ-Plots e histograma dos "ocupados EPA" e do seu logaritmo

TEST	"**ocupados EPA**"	"log**(ocupados EPA)**"
Shapiro-Wilks	W=0.5903, p-valor $< 2{,}2e^{-16}$	W=0.9888, p-valor = 0,00
Kolmogorov-Smirnov	D=0.2805, p-valor $< 2{,}2e^{-16}$	D=0.0406, p-valor = 0,04

Cadro 1.1: Resultados dos tests de normalidade

Vemos que resultados nos ofrecen os dous tests de significación para determinar a normalidade dunha variable: o *test de Shapiro-Wilks* e o *test de Kolmogorov-Smirnov-Lilliefors*. En vista da información que recolle o Cadro 1.1 concluímos que parece axeitado transformar a variable resposta empregando a transformación logarítmica. Debemos ter en conta o tamaño de mostra é 2861 polo tanto nos fixaremos nos resultados que nos da o test de Kolmogorov-Smirnov-Lilliefors, pois, lembremos que o test de Shapiro-Wilks úsase para mostras pequenas, mentres que o test de Kolmogorov-Smirnov-Lilliefors utilízase para mostras grandes. Polo tanto, queda xustificado o uso da transformación logarítmica da variable resposta.

Na Figura 1.3, considérase o caso no que a variable resposta toma o seu valor logarítmico, pero non o fan as variables explicativas. Observamos que, no diagrama de dispersión dos ocupados EPA respecto do número de contratos non hai unha clara relación lineal monótona crecente. No caso do diagrama de dispersión dos ocupados EPA respecto do número de afiliados á SS pérdese a relación lineal aínda que continuamos observando unha relación monótona crecente, polo que xorde observar como sería tal diagrama se consideramos as variables explicativas tamén na forma logarítmica na Figura 1.4. En vista deste diagrama da Figura 1.4 vemos que as relacións son lineares.

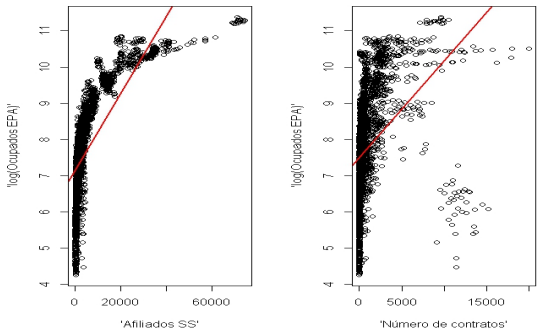

Figura 1.3: Diagrama de dispersión "log(ocupados EPA)" en función de "afiliados SS" e "nº de contratos"

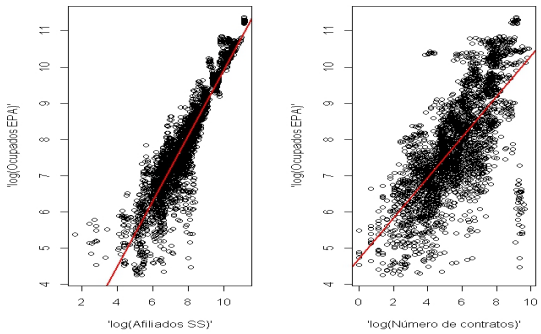

Figura 1.4: Diagrama de dispersión "log(ocupados EPA)" en función de "log(afiliados SS) e "log(nº de contratos)"

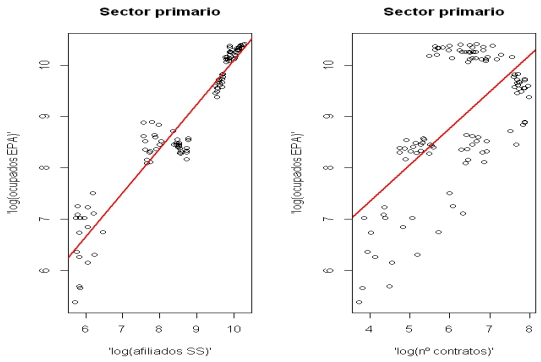

Figura 1.5: Diagramas de dispersión para as variables agrupadas polo sector primario

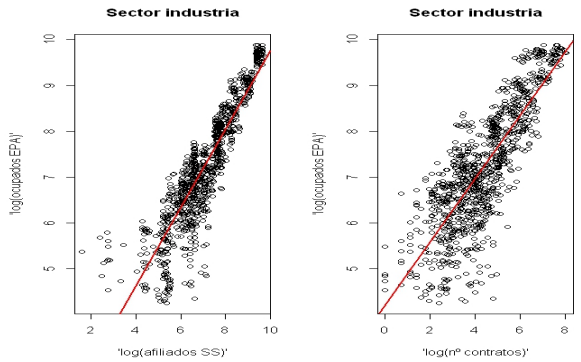

Figura 1.6: Diagramas de dispersión para as variables agrupadas polo sector industria

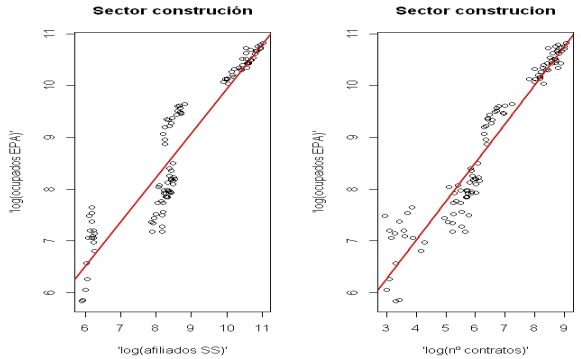

Figura 1.7: Diagramas de dispersión para as variables agrupadas polo sector construción

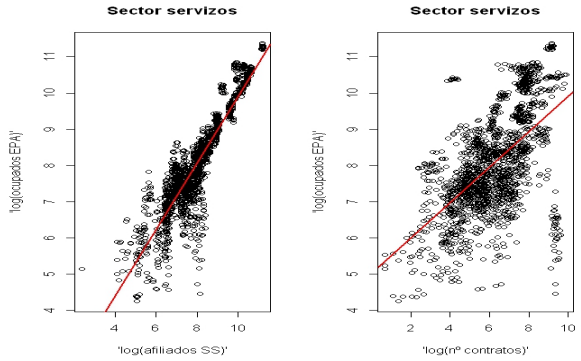

Figura 1.8: Diagramas de dispersión para as variables agrupadas polo sector servizos

Nas Figuras 1.5, 1.6, 1.7 e 1.8 podemos ver os diagramas de dispersión para as variables agrupadas polo sector de actividade: primario, industria, construción e servizos respectivamente. E, na Figura 1.9 temos os diagramas de dispersión para as variables agrupadas por sexos.

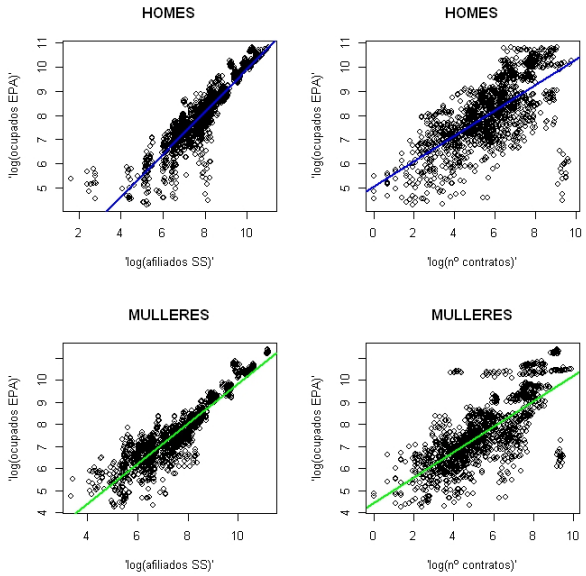

Figura 1.9: Diagramas de dispersión para as variables agrupadas por sexos

Realizamos na Figura 1.10 o diagrama de dispersión do tamaño de mostra da EPA e o coeficiente de variación do estimador directo nos diferentes dominios ordenados polo tamaño da mostra. Observamos que a medida que aumenta o tamaño de mostra o coeficiente de variación (cv) é menor. Nótese que para calcular o cv primeiramente calculamos as varianzas dos estimadores para os dominios (aproxímanse coa librería survey). Entón o coeficiente de variación calcúlase como o cociente entre a desviación típica (raíz cadrada da varianza) e a media.

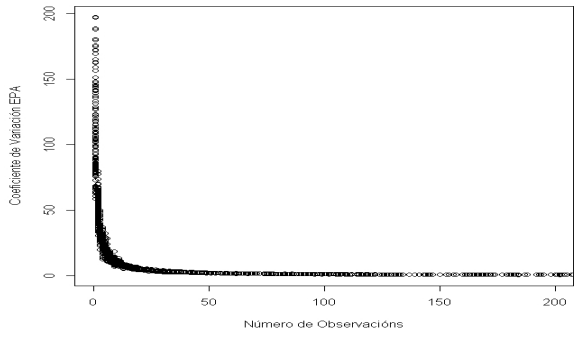

Figura 1.10: Diagrama de dispersión do coeficiente de variación dos ocupados EPA respecto ao tamaño de mostra

Preséntase no Cadro 1.2 os valores mínimo, máximo, media e mediana por sectores de actividade: primario, construción, industria e servizos nos cales observamos que cando hai poucas observacións o coeficiente de variación é maior.

Sector	Nº de dominios	Mínimo	Máximo	Media	Mediana
Primario	3	29.21 (266)	45.11 (2)	3.618	1.088 (68)
Industria	34	39.42 (147)	110.80 (1)	13.290	6.657 (8)
Construción	3	25.36 (360)	37.35 (2)	4.650	2.206 (42)
Servizos	48	16.60 (642)	208.60 (1)	11.850	5.988 (17)
Total	88	15.69 (642)	197.60 (1)	15.160	6.717 (15)

Cadro 1.2: Distribución do coeficiente de variación en % por sectores e do seu tamaño de mostra entre paréntese en cada caso

Neste Cadro 1.2 podemos ver a distribución do coeficiente de variación por sectores. Así, por exemplo, o sector primario, no cal hai 3 dominios ou áreas pequenas, o coeficiente de variación mínimo é 29.21 cun tamaño de mostra de 266, o máximo é 45.11 con tamaño de mostra 2, a media é 3.618, a mediana é 1.088 con tamaño de mostra de 68. A importancia dos resultados que este cadro nos ofrece é crucial para xustificar a importancia da estimación en áreas pequenas, pois observamos que no peor dos casos chégase a un coeficiente de 208,60 % que está asociado a unha actividade económica que ten só unha observación, e no mellor dos casos, con 642 observacións, o coeficiente de variación é do 16,6 %; polo tanto, queremos solventar este problema usando as estimacións en áreas pequenas que presentaremos nos vindeiros capítulos, e veremos que queda solventado tal problema ao longo desta memoria.

Queda así mostrada a incidencia que as áreas pequenas teñen no incremento dos coeficientes de variación. Vemos ademais que as actividades que presentan menos observacións e máis tamaño de mostra son as que pertencen aos sectores primario e construción. E conclúese que, cando hai poucas observacións, o coeficiente de variación é maior.

Capítulo 2

Metodoloxía

Neste segundo capítulo atópase toda a metodoloxía que aplicaremos aos datos reais, cuxos resultados se atopan no cuarto capítulo desta memoria.

Comezaremos adicando unha breve sección aos modelos lineais mixtos, para logo presentar o **Modelo Fay-Herriot**, modelo lineal mixto a nivel de áreas pequenas, modelo que empregaremos para estimar o número de ocupados da EPA nas áreas pequenas do noso estudo. Usaremos dous modelos, extensións do Modelo Fay-Herriot: o modelo a nivel de área con efectos de tempo independentes (que denotaremos abreviadamente de aquí en adiante **modelo 1**) e o modelo a nivel de área con efectos de tempo correlados (ao que nos referiremos no que segue como **modelo 2**).

Presentaranse estes dous modelos por separado, e, como dixemos, no Capítulo 4 presentaranse os resultados obtidos no casos dos datos reais para cada un deles por separado.

Tras presentar os modelos 1 e 2, e as estimacións dos seus parámetros, faremos a estimación do erro do estimador. En particular traballarase co MSE e, para o seu cálculo, procederemos de dous xeitos: usando expresións analíticas e usando métodos de remostraxe, estes métodos de remostraxe servirán para enlazar con vindeiro capítulo, que trata de simular unha poboación similar á nosa e ver como se comporta fronte aos métodos de remostraxe para calcular o MSE.

Ademais, finalmente será interesante facer unha comparación entre o estimador directo e o obtido con cada un dos modelos.

2.1. Modelo lineal mixto

Lembremos que unha área pequena leva asociado o problema da ausencia dunha mostra significativa. Usamos entón o termo área pequena para referirnos a áreas xeográficas pequenas, como municipios ou comarcas, ou a pequenas subpoboacións, como desempregados, xuventude, minusválidos, minorías étnicas...

Os modelos de áreas pequenas supoñen a existencia dun modelo subxacente que seguen todos os datos da poboación, pero que se estima cos datos da mostra (*Rao*, 2003). Para o obtención de estimadores a nivel de área, atopamos os modelos de efectos fixos e os modelos mixtos.

Os modelos de efectos fixos, asumen que a variabilidade entre as áreas pequenas da variable resposta pode ser totalmente explicada en termos da correspondente variabilidade da información auxiliar.

Nos modelos mixtos o predictor consta dun termo común de efectos fixos e outro diferenciado para os elementos de cada área pequena, chamémola d. Este termo diferenciado, está formado polos efectos aleatorios v_d, de xeito que todos os datos da mesma área pequena comparten o mesmo efecto aleatorio. Un caso particular de modelo lineal mixto para datos agregados é o que veremos na seguinte sección: o modelo Fay-Herriot.

Nos modelos de efectos fixos, non existen termos diferenciados por cada área pequena pois a parte sistemática $X\beta$ é común para todas as áreas pequenas. Pero, a especificade lógrase ao proxectar o coeficiente común β na información auxiliar específica X_d de cada área pequena.

Henderson desenvolveu en 1950 os estimadores BLUP (mellor preditor lineal insesgado) para estimar os efectos aleatorios dos modelos mixtos. A partires do ano 1962, empezaron a ser usados asumindo que as varianzas dos efectos aleatorios son coñecidas, na práctica non o son e débense estimar a partires dos datos.

Robinson propuxo en 1991 os estimadores EBLUP (mellor estimación lineal empírica insesgada). Están obtidos a partires dos BLUP reemprazando as compoñentes descoñecidas da varianza por estimadores asociados. Usaremos estes estimadores no noso estudo.

2.2. Modelo Fay-Herriot e Modelo Rao-Yu

Como xa se comentou na Introdución, os modelos de áreas pequenas pódense clasificar en dous tipos segundo temos dispoñible a información auxiliar: modelos a nivel de área e a nivel de unidade. Non podemos aplicar modelos a nivel de unidade, pois as variables auxiliais a este nivel non están dispoñibles coas variables auxiliares que estamos traballando nesta memoria: afiliados á SS e número de contratos rexistrados. Polo tanto, aplicaremos modelos a nivel de área.

Cando as variables auxiliares do noso estudo están dispoñibles a nivel de área o modelo que extensamente se usa é o chamado modelo Fay-Herriot, primeiro modelo lineal mixto proposto. Este modelo, como o seu nome indica, foi proposto por *Fay e Herriot* en 1979 para obter estimacións en áreas pequenas da media da renda per capita en áreas pequenas de Estados Unidos.

O modelo Fay-Herriot constrúese en dúas etapas. Sexa μ_d a característica de interese na área d e y_d o seu estimador directo. O modelo dinos que o estimador directo de $\{y_d\}$ se expresa como

$$y_d = \mu_d + e_d \qquad d = 1, \cdots, D \tag{2.1}$$

sendo D o número total de áreas ou dominios. Sexa $\{e_d\}$ o erro de mostraxe, o cal dado μ_d é independente coa varianza coñecida σ_d^2:

$$e_d \sim N(0, \sigma_d^2) \tag{2.2}$$

Na segunda etapa deste modelo, asúmese que μ_d varía linealmente co número p de variables auxiliares a nivel de área, é dicir,

$$\mu_d = x_d'\beta + u_d \quad d = 1, \cdots, D \tag{2.3}$$

onde x_d é o vector que contén os valores das p variables auxiliares para a área d, β é o vector dos coeficientes de regresión e u_d son os erros do modelo, os cales asumimos que son i.i.d. de unha $N(0, \sigma_u^2)$ con varianza σ_u^2 descoñecida e independentes de $\{e_d\}$. Finalmente, o modelo Fay-Herriot queda expresado como segue

$$y_d = x_d'\beta + u_d + e_d \quad d = 1, \cdots, D \tag{2.4}$$

Choudhry e Rao en 1989 estenderon o modelo básico de Fay-Herriot incluíndo a variación dos instantes de tempo e considerando unha estrutura autocorrelada para os erros de mostraxe:

$$y_{dt} = x_{dt}'\beta + u_d + e_{dt} \quad d = 1, \cdots, D, \quad t = 1 \cdots, T \tag{2.5}$$

sendo y_{dt} e x_{dt} a variable resposta e o vector coas variables auxiliares para a área d no instante de tempo t, con $\mu_{dt} = x'_{dt}\beta + u_d$ as características obxectivo para a mesma área e no mesmo instante. Para cada dominio d, asúmese que os erros $\{e_{dt}\}_{t=1}^{T}$ seguen un modelo autorregresivo de orden 1, AR(1).

En 1994, *Rao e Yu* propoñen o modelo:

$$y_{dt} = x'_{dt}\beta + u_{1d} + u_{2dt} + e_{dt} \quad d = 1, \cdots, D, \quad t = 1, \cdots, T \tag{2.6}$$

o cal é unha forma simple de compartir información entre os dominios ao longo do tempo mediante a introdución nel dos efectos aleatorios que teñen en conta a variabilidade entre dominios e ao longo do tempo, sendo $\{u_{1d}\}$ unha variable aleatoria que mide o efecto área e, para cada d, $\{u_{2dt}\}$ son os efectos que varían co tempo e seguen un modelo AR(1), pero son independentes a través das áreas ao igual que o son os erros de mostraxe $\{e_{dt}\}$. Este modelo Rao-Yu é o usado polo paquete *saery* que empregaremos para estimar os parámetros dos modelos que nas seguintes seccións explicaremos.

Para ver distintas extensións e máis detalles do modelo Fay-Herriot véxase, por exemplo, *Marhuenda, Y. Molina, I. and Morales, D. (2013).*

2.3. Modelo a nivel de área con efectos de tempo independentes

Vexamos primeiramente notación empregada. O índice d vai ser empregado para os dominios, $d = 1, \cdots, D$, sendo D o número total de dominios. Y_{dt} é o estimador directo do indicador de interese da área (dominio) para o dominio d, $d = 1, \cdots, D$ no instante t, $t = 1, \cdots, m_d$. X_{dt} vai ser o vector que contén os valores das p variables auxiliares para o dominio d, $d = 1, \cdots, D$ no instante t con $t = 1, \cdots, m_d$.

En particular nos nosos datos obxecto de estudo, Y_{dt} é o estimador directo dos ocupados da EPA para o dominio d, $d = 1, \cdots, 86$ no instante t, $t = 1, \cdots, m_d$ con $m_d = 18$. X_{dt} contén os valores das nosas $p = 2$ variables auxiliares: "número de afiliados á SS" e "número de contratos novos rexistrados".

u_{dt} son os efectos aleatorios que están incorrelados con media cero e varianza σ_u^2 descoñecida, asumimos sempre a normalidade de u_d, entón $u_{dt} \sim N(0, \sigma_u^2)$ sendo σ_u^2 descoñecida. Denotaremos aos erros por e_{dt}, con $e_{dt} \sim N(0, \sigma_{dt}^2)$ e σ_{dt}^2 coñecidos. Ademais, $\{e_{dt}\}$ e $\{u_{dt}\}$ son independentes. β é o p-vector dos parámetros de regresión.

Consideramos o seguinte modelo:

$$\text{Modelo 1} \begin{cases} Y_{dt} = X_{dt}\beta + u_{dt} + e_{dt} \\ \qquad d = 1, \cdots, D \\ \qquad t = 1, \cdots, m_d \end{cases}$$

Matricialmente podemos escribir o modelo M1 como $Y = X\beta + Zu + e$ sendo, para $1 \leq d \leq D$, $1 \leq t \leq m_d$ e $1 \leq i \leq p$, o que segue:

- $Y = \text{col}(Y_d)$, $Y_d = \text{col}(Y_{dt})$
- $X = \text{col}(X_d)$, $X_d = \text{col}(x_{dt})$ con $x_{dt} = \text{col}'(x_{dti})$
- $\beta = \text{col}(\beta_i)$
- $u_d = \text{col}(u_{dt})$

- $e = \text{col}(e_d)$ con $e_d = \text{col}(e_{dt})$
- $Z = I_M$ e $M = \sum_{d=1}^{D} m_d$

Nesta notación matricial, $u \sim N(0, V_u)$ e $e \sim N(0, V_e)$ son independentes con matrices de varianza-covarianza $V_u = \sigma_u^2 I_M$ con $I_M = \text{diag}(I_{m_d})$ e $V_e = \text{diag}(V_{ed})$ con $V_{ed} = \text{diag}(\sigma_{dt}^2)$.

2.3.1. Estimación do modelo

Durante esta sección, daremos as expresións que seguen as estimacións dos parámetros do modelo, máis detalles pódense atopar en *Rao, J. N. K. (2003).*

Como vimos facendo en todo o capítulo, denotamos por m_d ao número de instantes temporais do dominio d, con $d = 1, \cdots, D$.

Por unha banda, para a estimación da varianza dos efectos aleatorios σ_u^2, a cal é descoñecida, empregamos o método de máxima verosimilitude restrinxida -REML no que segue- e, como punto de partida deste algoritmo empregamos o estimador de *Henderson* de σ_u^2, que se define como segue:

$$\widehat{\sigma}_{uH}^2 = \frac{Y' P_2 Y - (M - P)}{tr\{P2\}} \tag{2.7}$$

sendo, para cada $d = 1, \cdots, D$,
$P_2 = \text{diag}(V_{ed}^{-1}) - \text{col}(V_{ed}^{-1} X_d) Q_2 \text{col}'(X_d V_{ed}^{-1})$,
$tr\{P2\} = \sum_{d=1}^{D} \sum_{t=1}^{m_d} \sigma_{dt}^{-2} - \sum_{d=1}^{D} tr\{X_d' V_{ed}^{-2} X_d Q_2\}$,
$Y' P_2 Y = \sum_{d=1}^{D} \sum_{t=1}^{m_d} \sigma_{dt}^{-2} Y_{dt}^2 - \left(\sum_{d=1}^{D} Y' V_{ed}^{-1} X_d\right) Q_2 \left(\sum_{d=1}^{D} Y' V_{ed}^{-1} X_d\right)'$ e
$Q_2 = \sum_{d=1}^{D} (X_{ed}' V_{ed}^{-1} X_d)'$.

Os estimadores REML estímanse co algoritmo de estimación de Fisher (*Rao*, 1965) do seguinte xeito:

$$(\sigma_u^2)^{x+1} = (\sigma_u^2)^x + F^{-1}\left((\sigma_u^2)^x\right) S\left((\sigma_u^2)^{x+1}\right) \tag{2.8}$$

sendo a puntuación REML $S\left(\sigma_u^2\right) = \frac{1}{2} tr(P) + \frac{1}{2} Y' P^2 Y$ e a cantidade de información de Fisher é $F\left(\sigma_u^2\right) = \frac{1}{2} tr(P^2)$.

Pola outra banda, os estimadores BLUE de β e u do modelo 1 son:

$$\widehat{\beta} = (X' V^{-1} X)^{-1} \tag{2.9}$$

$$\widehat{u} = V_u Z' V^{-1} (Y - X)\widehat{\beta} \tag{2.10}$$

sendo $Z = I_M$ con $M = \sum_{d=1}^{D} m_d$, m_d é o número de instantes de tempo no dominio d, $d = 1, \cdots, D$.

A varianza de Y, $Var(Y)$, calcúlase como segue:

$$Var(Y) = \sigma_u^2 \text{diag}(I_{md}) + V_e = \text{diag}(\sigma_u^2 I_{md} + V_e) \qquad d = 1, \cdots, D \tag{2.11}$$

Nótese que na análise exploratoria, decidiramos considerar a variable resposta na súa forma logarítmica, polo tanto, para calcular a varianza $Var(log(Y))$ procedemos polo método delta resultando $Var(log(Y)) = \frac{Var(Y)}{\widehat{Y}^2}$. En canto á varianza dos efectos aleatorios σ_u^2, descoñecida, para estimala

empregamos o método de máxima verosimilitude restrinxida (REML).

Na presente memoria estamos interesados en predicir $\mu_{dt} = X_{dt}\beta + u_{dt}$. Henderson (1975) obtivo o BLUP como unha combinación linear dos elementos fixos de β e os elementos aleatorios de u_{dt} no modelo lineal xeral. Seguindo os seus resultados, o BLUP de μ_{dt} é o que segue

$$\widehat{\mu}_{dt}^{BLUP} = \widehat{\mu}(\sigma_u^2, Y) = X\widehat{\beta}^{BLUP} + \widehat{u}_{dt}^{BLUP} \tag{2.12}$$

onde $\widehat{\beta}^{BLUP} = \widehat{\beta}(\sigma_u^2, Y) = (X^t V^{-1} X)^{-1} X^t V^{-1} Y$ e $\widehat{u}_{dt}^{BLUP} = \widehat{u}_{dt}(\sigma_u^2, Y) = \sigma_u^2 Z' V^{-1}(Y - X\widehat{\beta}^{BLUP})$. Pero, como indica a propia notación, $\widehat{\beta}^{BLUP}$ e $\widehat{u}_{dt}^{BLUP}$ dependen de σ_u^2 que é descoñecido na práctica. Reemprazando, o estimador $\widehat{\sigma}_u^2$ por σ_u^2 en (2.10) chegamos ao EBLUP

$$\widehat{\mu}_{dt}^{EBLUP} = \widehat{\mu}(\sigma_u^2, Y) = X\widehat{\beta}^{EBLUP} + \widehat{u}_{dt}^{EBLUP}$$

onde $\widehat{\beta}^{EBLUP} = \widehat{\beta}(\sigma_u^2, Y)$ e $\widehat{u}_{dt}^{EBLUP} = \widehat{u}_{dt}(\sigma_u^2, Y)$.

2.3.2. Estimación do MSE

Unha parte moi importante na estimación en áreas pequenas é a estimación do erro do estimador, máis concretamente trabállase co erro cadrático medio -MSE-.

Prasad e Rao (1990) obtiveron unha expresión analítica para o MSE do EBLUP e propuxeron unha corrección do estimador. *Butar e Lahiri* (2003) usaron un bootstrap paramétrico para calcular o estimador do MSE baixo un modelo linear mixto. Máis recentemente, *Hall e Maiti* (2006) deron o método bootstrap paramétrico para construír unha corrección dos estimadores do MSE.

O MSE de $\widehat{\mu}_{dt}^{EBLUP}$, que para simplificar notación chamaremos $\widehat{\mu}_{dt,E}$, defínese como

$$\text{MSE}(\widehat{\mu}_{dt,E}) = E[(\widehat{\mu}_{dt} - \mu)(\widehat{\mu}_{dt,E} - \mu)'].$$

Esta matriz pódese descompoñer como

$$MSE(\widehat{\mu}_{dt,E}) = MSE(\widehat{\mu}_{dt,B}) + E[(\widehat{\mu}_{dt,E} - \widehat{\mu}_{dt,B})(\widehat{\mu}_{dt,E} - \widehat{\mu}_{dt,B})'] +$$

$$E[(\widehat{\mu}_{dt,E} - \widehat{\mu}_{dt,B})(\widehat{\mu}_{dt,B} - \mu)'] + E[(\widehat{\mu}_{dt,B} - \mu)(\widehat{\mu}_{dt,E} - \widehat{\mu}_{dt,B})'].$$

O último termo da expresión representa o erro debido a predición de μ_{dt} cando σ_u^2 é coñecida, e o penúltimo termo expresa o aumento do erro de predición debido a estimación de σ_u^2.

Seguindo a *Krackar e Harville* (1984) e seguindo a suposición de normalidade, é fácil de probar que para calquera estimador insesgado e traslación invariante de σ_u^2, os dous últimos termos son matrices de cero. Ademais, seguindo a *Henderson* (1975),

$$MSE(\widehat{\mu}_{dt,B}) = g_1(\sigma_u^2) + g_2(\sigma_u^2)$$

onde, para $Q = (X'V^{-1}X)^{-1}$ e $T = \sigma_u^2(I - \sigma_u^2 Z'V^{-1}Z)$,

$$g_1(\sigma_u^2) = ZTZ' \quad e \quad g_2(\sigma_u^2) = (X - ZTZ'\Sigma^{-1}X)Q(X' - X'\Sigma^{-1}ZTZ')$$

Temos a medida do erro o MSE de *Prasad e Rao* (1990), estimación EBLUP:

$$MSE(\widehat{Y}_{dt}^{EBLUP}) \approx g_1(\sigma_u^2) + g_2(\sigma_u^2) + g_3(\sigma_u^2) \tag{2.13}$$

$$\widehat{MSE}(\widehat{Y}_{dt}^{EBLUP}) = g_1(\widehat{\sigma}_u^2) + g_2(\widehat{\sigma}_u^2) + 2g_3(\widehat{\sigma}_u^2) \tag{2.14}$$

onde:

$g_1(\widehat{\sigma}_u^2)$, $g_2(\widehat{\sigma}_u^2)$ e $g_3(\widehat{\sigma}_u^2)$ son as expresións que foron estudadas por *Prasad e Rao* (1990):

$$\begin{aligned} g_1(\widehat{\sigma}_u^2) &= \frac{\sigma_u^2\sigma_{dt}^2}{\sigma_u^2+\sigma_{dt}^2} \\ g_2(\widehat{\sigma}_u^2) &= [a_d'X_d - \sigma_u^2 a_d' V_{ed}^{-1}X_d + \sigma_u^4 a_d' V_d^{-1}V_{ed}^{-1}X_d] \quad \cdot Q \\ &\cdot \quad [X_d'a_d - \sigma_u^2 X_d' V_{ed}^{-1}a_d + \sigma_u^4 X_d' V_{ed}^{-1}V_{ed}^{-1}a_d] \\ \text{con} \quad a_d &= col(\delta_{tk}), \quad 1 \le k \le m_d \\ g_3(\widehat{\sigma}_u^2) &= qF^{-1}(\sigma_u^2) \quad \text{sendo} \quad q = \frac{1}{\sigma_u^2+\sigma_{dt}^2} - \frac{2\sigma_u^2}{(\sigma_u^2+\sigma_{dt)^2})^2} \end{aligned}$$

onde F contén a cantidade da información de Fisher calculada co REML mediante a ecuación do algoritmo de puntuacións de Fisher.

Ademais, *Prasad e Rao*(1990) probaron que:

$$E[g_1(\widehat{\sigma}_u^2)] = g_1(\sigma_u^2) - g_3(\sigma_u^2) + [o(D^{-1})]$$

$$E[g_i(\widehat{\sigma}_u^2)] = g_i(\sigma_u^2) + [o(D^{-1})] \quad i = 2,3$$

Bootstrap paramétrico

As expresións analíticas, que se mostraron no punto anterior, son aproximacións que asumen hipóteses fortes e que son específicas do modelo. Os métodos de remostraxe baséanse en condicións máis débiles que as das aproximacións analíticas, pódense aplicar de maneira similar baixo calquera modelo estadístico e fan comparables os resultados obtidos polos distintos estimadores.

Para o estimador Fay-Herriot, utilízase un método Bootstrap paramétrico (*González-Manteiga et al.*, 2008) deseñado especialmente para o modelo. A continuación, vemos o procedemos bootstrap para aproximar $MSE(\widehat{\mu})$. Aínda que a suposición de normalidade é precisa para probar a consistencia das aproximacións, a aplicación do método non a require. Tamén daremos dúas alternativas aos estimadores bootstrap.

Tal bootstrap paramétrico procede como segue:

1. Calcular estimacións $\widehat{\sigma}_u^2 = \widehat{\sigma}_u^2(Y)$ e $\widehat{\beta}_E = \widehat{\beta}(\widehat{\sigma}_u^2, Y)$ de σ_u^2 e β respectivamente.
   Calcular $\Sigma_u = \text{diag}\{\sigma_u^2\}_{1\le d\le D}$, Σ_u é unha matriz diagonal.

2. Xerar D copias independentes dunha variable $W_1 \sim N(0,1)$. Construír o vector $u^* = (u_1^*, \cdots, u_D^*)'$ cos elementos $u_d^* = \widehat{\sigma}_u W_1, \quad d = 1, \cdots, D$.

3. Xerar D copias independentes dunha variable $W_2 \sim N(0,1)$,independente de W_1. Construír o vector $\epsilon^* = (\epsilon_1^*, \cdots, \epsilon_D^*)'$ cos elementos $\epsilon_d^* = \widehat{\sigma}_d W_2, \quad d = 1, \cdots, D$.

4. Construír o modelo bootstrap:
   $$Y^* = X\widehat{\beta}_E + u^* + \epsilon^*$$
   Baixo este modelo bootstrap, defínese o BLUP de $\mu^* = X\widehat{\beta}_E + u^*$ como:
   $$\widehat{\mu}_B^* = \widehat{\mu}(\widehat{\sigma}_u^2, Y^*) = X\widehat{\beta}_B^* + \widehat{u}_B^*$$
   sendo $\widehat{\beta}_B^* = \widehat{\beta}(\widehat{\sigma}_u^2, Y^*)$ e $\widehat{u}_B^* = \widehat{u}(\widehat{\sigma}_u^2, Y^*)$.
   Sexa $\widehat{\sigma}_u^{2*} = \widehat{\sigma}_u^2(Y^*)$ o estimador de σ_u^2 calculado no mundo bootstrap.

Usando esta información, conséguese o estimador bootstrap EBLUP:

$$\widehat{\mu}_E^* = \widehat{\mu}(\widehat{\sigma}_u^{2*}, Y^*) = X\widehat{\beta}_E^* + \widehat{u}_E^*$$

sendo $\widehat{\beta}_E^* = \widehat{\beta}(\widehat{\sigma}_u^{2*}, Y^*)$ e $\widehat{u}_E^* = \widehat{u}(\widehat{\sigma}_u^{2*}, Y^*)$

5. Xerar B vectores bootstrap $\widehat{\mu}^{*(b)}$, $b = 1, \cdots, B$ do modelo bootstrap definido no paso anterior. De cada vector $\widehat{\mu}^{*(b)}$ calcúlase a súa media real $\mu^{*(b)}$ e o seu estimador EBLUP $\widehat{\mu}_E^{*(b)}$. Finalmente calculamos o estimador do erro cadrático medio para cada dominio:

$$MSE^{*,1} = \widehat{\mathrm{MSE}}_B(\widehat{\mu}_{E,d}) = \frac{1}{B-1}\sum_{b=1}^{B}(\widehat{\mu}_{E,d}^{*(b)} - \mu_d^{*(b)})^2. \tag{2.15}$$

A técnica do método bootstrap é especialmente útil para a aproximación das cantidades descoñecidas. Por isto, é razoable aplicar este método soamente para o termo de $MSE(\widehat{\mu}_{dt})$ que non pode ser analiticamente calculado, é dicir para $E[(\widehat{\mu}_{dt,E} - \widehat{\mu}_{dt,B})(\widehat{\mu}_{dt,E} - \widehat{\mu}_{dt,B})']$. Seguindo esta idea, definimos o estimador bootstrap termo a termo como

$$MSE^{*,2} = g_1(\widehat{\sigma}_u^2) + g_2(\widehat{\sigma}_u^2) + E^*[(\widehat{\mu}_{dt,E}^* - \widehat{\mu}_{dt,B}^*)(\widehat{\mu}_{dt,E}^* - \widehat{\mu}_{dt,B}^*)']. \tag{2.16}$$

Usando as ideas de *Pfeffermann e Tiller* (2005), definimos o estimador bootstrap bias-corrector como

$$\begin{aligned} MSE^{*,3} = 2[g_1(\widehat{\sigma}_u^2) + g_2(\widehat{\sigma}_u^2)] - E^*[g_1(\widehat{\sigma}_u^{2,*}) + g_2(\widehat{\sigma}_u^{2,*})] + \\ E^*[(\widehat{\mu}_{dt,E}^* - \widehat{\mu}_{dt,B}^*)(\widehat{\mu}_{dt,E}^* - \widehat{\mu}_{dt,B}^*)']. \end{aligned} \tag{2.17}$$

2.4. Modelo a nivel de área con efectos de tempo correlados

De forma análoga a como procedemos co modelo 1 faremos co modelo a nivel de área con efectos de tempo correlados, modelo 2, comezando por describir a notación:

De novo, o índice d vai ser empregado para os dominios, $d = 1, \cdots, D$, sendo D o número total de dominios. Y_{dt} é o estimador directo do indicador de interese da área para o dominio d, $d = 1, \cdots, D$ no instante t, $t = 1, \cdots, m_d$. X_{dt} vai ser o vector que contén os valores das p variables auxiliares para o dominio d, $d = 1, \cdots, D$ no instante t, $t = 1, \cdots, m_d$.

u_{dt} son os efectos aleatorios temporais que seguen un proceso autorregresivo AR(1) con varianza σ_u^2 e autocorrelación ρ. Denotaremos aos erros por e_{dt}, con $e_{dt} \sim N(0, \sigma_{dt}^2)$ e σ_{dt}^2 coñecidos. Ademais, $\{e_{dt}\}$ e $\{u_{dt}\}$ son independentes. β é o p-vector dos parámetros de regresión.

Consideramos o seguinte modelo:

$$\text{Modelo 2} \begin{cases} Y_{dt} = X_{dt}\beta + u_{dt} + e_{dt} \\ \qquad d = 1, \cdots, D \\ \qquad t = 1, \cdots, m_d \end{cases}$$

Matricialmente podemos escribir o modelo 2 como $Y = X\beta + Zu + e$ sendo, para $1 \leq d \leq D$, $1 \leq t \leq m_d$ e $1 \leq i \leq p$, o que segue:

- $Y = \mathrm{col}(Y_d)$, $Y_d = \mathrm{col}(Y_{dt})$

- $X = \text{col}(X_d)$, $X_d = \text{col}(x_{dt})$ con $x_{dt} = \text{col}'(x_{dti})$
- $\beta = \text{col}(\beta_i)$
- $u = \text{col}(u_{dt})$
- $e = \text{col}(e_d)$ con $e_d = \text{col}(e_{dt})$
- $Z = I_M$ e $M = \sum_{d=1}^{D} m_d$

Nesta notación matricial, $u \sim N(0, V_u)$ e $e \sim N(0, V_e)$ son independentes con matrices de varianza-covarianza $V_u = \sigma_u^2 \Omega(\rho)$ con $\Omega(\rho) = \text{diag}\Omega_d(\rho)$ e $V_e = \text{diag}(V_{ed})$ con $V_{ed} = \text{diag}(\sigma_{dt}^2)$, sendo

$$\Omega_d = \Omega_d(\rho) = \frac{1}{1-\rho^2} \begin{pmatrix} 1 & \rho & ... & \rho^{m_d-2} & \rho^{m_d-1} \\ \rho & 1 & ... & ... & \rho^{m_d-2} \\ ... & ... & ... & ... & ... \\ \rho^{m_d-2} & ... & ... & 1 & \rho \\ \rho^{m_d-1} & \rho^{m_d-2} & ... & \rho & 1 \end{pmatrix}$$

2.4.1. Estimación do modelo

Os estimadores BLUE de β e u do modelo 2 son os seguintes:

$$\widehat{\beta} = (X'V^{-1}X)^{-1}(X'V^{-1}Y) \tag{2.18}$$

$$\widehat{u} = \sigma_u^2 \text{col}\left(\Omega_d(\rho)V_d^{-1}(Y_d - X_d\widehat{\beta})\right) \quad 1 \le d \le D \tag{2.19}$$

Ao ser os compoñentes da varianza descoñecidos, os seus estimadores REML calcúlanse empregando o algoritmo de puntuación de Fisher coa seguinte fórmula

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

sendo $\theta = (\sigma_u^2, \rho)$, e usando as sementes $\rho = 0$ e $\sigma_u^{2(0)} = \widehat{\sigma}_{uH}^2$ con $\widehat{\sigma}_{uH}^2$ o estimador Henderson de σ_u^2 baixo o modelo restrinxido a $\rho = 0$. As puntuacións Fisher e as compoñentes da matriz de información de Fisher son

$$S_a = -\frac{1}{2}\text{tr}(PV_a) + \frac{1}{2}Y'PV_aPY, \quad F_{ab} = \frac{1}{2}(PV_aPV_b), \quad a, b = 1, 2$$

onde, para cada $1 \le d \le D$:

$$V_1 = \frac{\delta V}{\delta \sigma_u^2} = \text{diag}(\Omega_d(\rho)), \quad V_2 = \frac{\delta V}{\delta \rho} = \sigma_u^2 \text{diag}(\Omega_d(\rho)), \quad Q = (\sum_{d=1}^{D} X_d' V_d^{-1} X_d)^{-1},$$

$$P = \text{diag}(V_d^{-1}) - \text{col}(V_d^{-1}X_d)Q\text{col}'(X_d'V_d^{-1}),$$

$$PV_a = \text{diag}(V_d^{-1}V_{ad}) - \text{col}(V_d^{-1}X_d)Q\text{col}'(X_d'V_d^{-1}V_{ad}),$$

$$tr(PV_a) = \sum_{d=1}^{D}(V_d^{-1}V_{ad}) - \sum_{d=1}^{D} tr(X_d'V_d^{-1}V_{ad}V_d^{-1}X_dQ),$$

$$tr(PV_aPV_b) = \sum_{d=1}^{D} tr(V_d^{-1}V_{ad}V_d^{-1}V_{bd}) - 2\sum_{d=1}^{D} tr(X_d'V_d^{-1}V_{ad}V_d^{-1}V_{db}V_d^{-1}X_dQ) +$$

$$tr\left\{(\sum_{d=1}^{D} X_d'V_d^{-1}V_{ad}V_d^{-1}X_d)Q(\sum_{d=1}^{D} X_d'V_d^{-1}V_{bd}V_d^{-1}X_d)Q\right\},$$

$$Y'PV_aPY = \sum_{d=1}^{D} Y_d'V_d^{-1}V_{ad}Y_d - (\sum_{d=1}^{D} Y_d'V_d^{-1}V_{ad}X_d)Q(\sum_{d=1}^{D} Y_d'V_d^{-1}X_d)' -$$

$$(\sum_{d=1}^{D} Y_d'V_d^{-1}X_d)Q(\sum_{d=1}^{D} Y_d'V_d^{-1}V_{ad}Y_d) +$$

$$(\sum_{d=1}^{D} Y_d'V_d^{-1}X_d)Q(\sum_{d=1}^{D} Y_d'V_d^{-1}V_{ad}Y_d)(\sum_{d=1}^{D} Y_d'V_d^{-1}X_d)'.$$

Por último, a derivada da matriz $\Omega_d(\rho)$ respecto de ρ, que denotamos $\delta\Omega_d(\rho)$ é

$$\delta\Omega_d(\rho) = \frac{1}{1-\rho^2}\begin{pmatrix} 0 & 1 & \cdots & \cdots & (m_d-1)\rho^{m_d-2} \\ 1 & 0 & \cdots & \cdots & (m_d-2)\rho^{m_d-3} \\ \cdots & \cdots & \cdots & \cdots & \cdots \\ (m_d-2)\rho^{m_d-3} & \cdots & \cdots & 0 & 1 \\ (m_d-1)\rho^{m_d-2} & \cdots & \cdots & 1 & 0 \end{pmatrix} + \frac{2\rho\Omega_d(\rho)}{(1-\rho^2)^2}$$

En canto ao estimador REML de β, simplemente se calcula aplicando a fórmula

$$\widehat{\beta}_{REML} = (X'\widehat{V}^{-1}X)^{-1}(X'\widehat{V}^{-1}X).$$

Ademais, as distribucións asintóticas dos estimadores REML e θ é $\widehat{\theta} \sim N(\theta, F^{-1}(\theta))$ e a de β é $\widehat{\beta} \sim N(\beta, (X'V^{-1}X)^{-1})$.

Igual que no modelo 1, estamos interesados en predicir $\mu_{dt} = X_{dt}\beta + u_{dt}$ co mellor predictor lineal non sesgado empírico (EBLUP): $\widehat{\mu} = X_{dt}\widehat{\beta} + \widehat{u}_{dt}$.

2.4.2. Estimación do MSE

Empregando de novo a metodoloxía de *Prasad e Rao* (1990), o erro cadrático medio de $\widehat{\mu}_{dt}^{\text{EBLUP}}$ aproxímase por:

$$MSE\left(\widehat{\mu}_{dt}^{\text{EBLUP}}\right) \approx g_1(\theta) + g_2(\theta) + g_3(\theta)$$

Polo tanto, o seu estimador é

$$\widehat{MSE}\left(\widehat{\mu}_{dt}^{\text{EBLUP}}\right) = g_1(\widehat{\theta}) + g_2(\widehat{\theta}) + g_3(\widehat{\theta})$$

con $\theta = (\sigma_u^2, \rho)$ e :

$$g_1(\theta) = \sigma_u^2 a_d'\Omega_d a_d - \sigma_u^4 a_d'\Omega_d V_d^{-1}\Omega_d a_d, \quad \text{sendo} \quad a_d = \text{col}(\delta_{tk}), 1 \le k \le m_d$$

$$g_2(\theta) = (a_d'X_d - \sigma_u^2 a_d'\Omega_d V_{ed}^{-1}X_d + \sigma_u^4 a_d'\Omega_d V_d^{-1}\Omega_d V_{ed}^{-1}X_d)Q(X_d'a_d - \sigma_u^2 X_d'V_{ed}^{-1}\Omega_d a_d +$$

$$\sigma_u^4 X_d' V_{ed}^{-1} \Omega_d V_d^{-1} \Omega_d a_d)$$

$$g_3(\theta) \approx tr\left[\begin{pmatrix} q_{11} & q_{12} \\ q_{21} & q_{22} \end{pmatrix}\begin{pmatrix} F_{11} & F_{12} \\ F_{21} & F_{22} \end{pmatrix}^{-1}\right]$$

sendo F_{ab} o elemento da matriz de información de Fisher calculado por REML e

$$q_{11} = a_d'\Omega_d V_d^{-1}\Omega_d a_d - 2\sigma_u^2 a_d'\Omega_d V_d^{-1}\Omega_d V_d^{-1}\Omega_d a_d + \sigma_u^4 a_d'\Omega_d V_d^{-1}\Omega_d V_d^{-1}\Omega_d V_d^{-1}\Omega_d a_d$$

$$q_{12} = \sigma_u^2 a_d'\Omega_d V_d^{-1}\delta\Omega_d a_d - \sigma_u^4 a_d'\Omega_d V_d^{-1}\delta\Omega_d V_d^{-1}\Omega_d a_d - \sigma_u^4 a_d'\Omega_d V_d^{-1}\Omega_d V_d^{-1}\delta\Omega_d a_d$$
$$+\sigma_u^6 a_d'\Omega_d V_d^{-1}\Omega_d V_d^{-1}\delta\Omega_d V_d^{-1}\Omega_d a_d$$

$$q_{22} = \sigma_u^4 a_d'\delta\Omega_d V_d^{-1} a_d - 2\sigma_u^6 a_d'\Omega_d V_d^{-1}\delta\Omega_d V_d^{-1}\delta\Omega_d a_d$$
$$+\sigma_u^8 a_d'\Omega_d V_d^{-1}\delta\Omega_d V_d^{-1}\delta\Omega_d V_d^{-1}\Omega_d a_d$$

Bootstrap paramétrico

Para o estimador Fay-Herriot, utilízase un método Bootstrap paramétrico (*González-Manteiga et al.*, 2008) deseñado especialmente para o modelo. A continuación, vemos o procedemos bootstrap para aproximar $MSE(\widehat{\mu})$. Aínda que a suposición de normalidade é precisa para probar a consistencia das aproximacións, a aplicación do método non a require. Tamén daremos dúas alternativas aos estimadores bootstrap.

1. Calcular estimacións $\widehat{\sigma}_u^2 = \widehat{\sigma}_u^2(Y)$ e $\widehat{\beta}_E = \widehat{\beta}(\widehat{\sigma}_u^2, Y)$ de σ_u^2 e β respectivamente.
   Calcular $\Sigma_u = \sigma_u^2\Omega(\rho)$, unha estrutura de correlación $AR(1)$.

2. Xerar D copias independentes dunha variable $W_1 \sim N(0,1)$. Construír o vector $u^* = (u_1^*, \cdots, u_D^*)'$ cos elementos $u_d^* = \widehat{\Sigma_u}^{\frac{1}{2}} W_1$.

3. Xerar D copias independentes dunha variable $W_2 \sim N(0,1)$,independente de W_1. Construír o vector $\epsilon^* = (\epsilon_1^*, \cdots, \epsilon_D^*)'$ cos elementos $\epsilon_d^* = \widehat{\Sigma_e}^{\frac{1}{2}} W_e$, sendo $\Sigma_e = diag\{\sigma_{dt}^2\}$.

4. Construír o modelo bootstrap:
   $$Y^* = X\widehat{\beta}_E + u^* + \epsilon^*$$
   Baixo este modelo bootstrap, defínese o BLUP de $\mu^* = X\widehat{\beta}_E + u^*$ como:
   $$\widehat{\mu}_B^* = \widehat{\mu}(\widehat{\sigma}_u^2, Y^*) = X\widehat{\beta}_B^* + \widehat{u}_B^*$$
   sendo $\widehat{\beta}_B^* = \widehat{\beta}(\widehat{\sigma}_u^2, Y^*)$ e $\widehat{u}_B^* = \widehat{u}(\widehat{\sigma}_u^2, Y^*)$.
   Sexa $\widehat{\sigma}_u^{2*} = \widehat{\sigma}_u^2(Y^*)$ o estimador de σ_u^2 obtido no mundo bootstrap.
   Usando esta información, conséguese o estimador bootstrap EBLUP:
   $$\widehat{\mu}_E^* = \widehat{\mu}(\widehat{\sigma}_u^{2*}, Y^*) = X\widehat{\beta}_E^* + \widehat{u}_E^*$$
   sendo $\widehat{\beta}_E^* = \widehat{\beta}(\widehat{\sigma}_u^{2*}, Y^*)$ e $\widehat{u}_E^* = \widehat{u}(\widehat{\sigma}_u^{2*}, Y^*)$

5. Xerar B vectores bootstrap $\widehat{\mu}^{*(b)}$, $b = 1, \cdots, B$ do modelo bootstrap definido no paso anterior. De cada vector $Y^{*(b)}$ calcúlase a súa media real $\mu^{*(b)}$ e o seu estimador EBLUP $\widehat{\mu}_E^{*(b)}$. Finalmente calculamos o estimador do erro cadrático medio para cada dominio:
   $$MSE^{*,1} = \widehat{\mathrm{MSE}}_B(\widehat{\mu}_{E,d}) = \frac{1}{B-1}\sum_{b=1}^{B}(\widehat{\mu}_{E,d}^{*(b)} - \mu_d^{*(b)})^2. \tag{2.20}$$

A técnica do método bootstrap é especialmente útil para a aproximación das cantidades descoñecidas. Por isto, é razoable aplicar este método soamente para o termo de $MSE(\widehat{\mu}_{dt})$ que non pode ser analiticamente calculado, é dicir para $E[(\widehat{\mu}_{dt,E} - \widehat{\mu}_{dt,B})(\widehat{\mu}_{dt,E} - \widehat{\mu}_{dt,B})']$. Seguindo esta idea, definimos o estimador bootstrap termo a termo como

$$MSE^{*,2} = g_1(\widehat{\sigma}_u^2) + g_2(\widehat{\sigma}_u^2) + E^*[(\widehat{\mu}^*_{dt,E} - \widehat{\mu}^*_{dt,B})(\widehat{\mu}^*_{dt,E} - \widehat{\mu}^*_{dt,B})']. \tag{2.21}$$

Usando as ideas de *Pfeffermann e Tiller* (2005), definimos o estimador bootstrap bias-corrector como

$$\begin{aligned} MSE^{*,3} = 2[g_1(\widehat{\sigma}_u^2) + g_2(\widehat{\sigma}_u^2)] - E^*[g_1(\widehat{\sigma}_u^{2,*}) + g_2(\widehat{\sigma}_u^{2,*})] + \\ E^*[(\widehat{\mu}^*_{dt,E} - \widehat{\mu}^*_{dt,B})(\widehat{\mu}^*_{dt,E} - \widehat{\mu}^*_{dt,B})']. \end{aligned} \tag{2.22}$$

Capítulo 3

Estudo de simulación

No presente capítulo faremos un estudo de simulación. Para elo, simularemos catro poboacións: unha para o modelo con efectos de tempo independentes asumindo que os erros seguen unha distribución normal (modelo 1, escenario de simulación 1º), outra asumindo que os erros seguen unha distribución exponencial (modelo 1, escenario de simulación 2º), unha terceira poboación sobre o modelo de efectos de tempo correlados asumindo que os erros seguen unha distribución normal (modelo 2, escenario de simulación 1º) e por último os erros seguen unha distribución exponencial (modelo 2, escenario de simulación 2º).

Neste estudo de simulación estamos tratando de imitar o comportamento da poboación real.

Presentaremos polo tanto na próxima sección experimentos de simulación deseñados para analizar o comportamento dos estimadores do modelo e propoñer estimadores do MSE.

Nótese que para o presente capítulo, a bibliografía usada é *Marhuenda, Y. Molina, I. and Morales, D. (2013)*, onde se fai un estudo de simulación análogo ao que a continuación faremos e *W. González-Manteiga, M. J. Lombardía, I.Molina, D. Morales, L. Santamaría.(2008)*.

3.1. Experimento de simulación para o Modelo 1

Sexa D o número de áreas pequenas, T o número de instantes de tempo (lembremos que dada a natureza dos nosos datos, traballamos con trimestres). Tal e como introducimos no capítulo anterior, traballaremos instantes temporais con $t = 1, \cdots, T$ e número de dominios $d = 1, \cdots, D$.

Lembremos que no caso dos datos reais, o número de áreas pequenas é $D = 86$ que se corresponde coas actividades económicas determinadas pola CNAE a dous díxitos, e $T = 18$ pois ese é o número de trimestres do período de tempo no que traballamos, é dicir, os trimestres comprendidos dende o terceiro trimestre do ano 2009 ao cuarto trimestre do ano 2013. No experimento de simulación tomaremos $D \in \{50, 100, 150\}$ e $T \in \{4, 8, 12\}$, pois sabemos que se incrementamos máis o número de trimestres na simulación, o tempo de execución será demasiado elevado. Ademais, con estes períodos de tempo xa observamos o funcionamento da metodoloxía proposta cando vai aumentando o número de trimestres.

Tendo en conta o anterior, procedemos a simular o 1º escenario como segue:

- **Paso 1:** $x_{dt} = \dfrac{d + t/T}{D}$,
- **Paso 2:** $v_d \sim N(0, \sigma_v)$, sendo σ_v =1.2,

- **Paso 3:** $u_{dt} \sim N(0, \sigma_u)$, sendo $\sigma_u \in \{0.8{,}1\}$,
- **Paso 4:** $e_{dt} \sim N(0, \sigma_{dt})$, tendo σ^2_{dt} a seguinte expresión

$$\sigma^2_{dt} = \frac{(\alpha_1 - \alpha_0)[T(d-1)+t+1]}{DT-1} + \alpha_0 \tag{3.1}$$

  onde $\alpha_0 = \text{mín}_{d,t}\left\{\sigma^2_{dt}\right\} = 1{,}35$ e $\alpha_1 = \text{máx}_{d,t}\left\{\sigma^2_{dt}\right\} = 0.23$ son respectivamente os valores mínimo e máximo calculados dos datos reais. Tamén poderíamos coller $\alpha_0 = 0.8$ e $\alpha_1 = 1.2$ seguindo o estudo de simulación do artigo [8]: *Small area estimation with spatio-temporal Fay-Herriot models.*

Finalmente, o estimador directo é xerado do modelo Rao-Yu (pois é o usado no paquete *saery* que usamos para as estimacións), descrito en *Small area estimation by combining time series and cross sectional data*, páxinas 511-528, e que esencialmente é se diferencia do modelo Fay-Herriot en que ten un termo máis v_d de efectos aleatorios como

$$y_{dt} = \beta_0 + x_{dt}\beta_1 + v_d + u_{dt} + e_{dt} \tag{3.2}$$

e tomaremos $\beta_0 = 1$ e $\beta_1 = 0{,}75$.

Ademais podemos construír un escenario distinto, escenario 2º, asumindo no Paso 3 que os erros e_{dt} seguen unha distribución exponencial de parámetro $1/\sigma_{dt}$, para comprobar que é este o parámetro adecuado, simplemente temos en conta que a varianza dunha variable aleatoria con distribución exponencial de parámetro λ é $1/\lambda^2$, polo que

$$\sigma^2_{dt} = \frac{1}{\lambda^2} \Rightarrow \lambda = \frac{1}{\sigma_{dt}}.$$

3.2. Experimento de simulación para o Modelo 2

Neste segundo experimento de simulación tomaremos de novo $D \in \{50, 100, 15\}$ e $T \in \{4, 8, 12\}$ e, tamén, $\beta_0 = 1$ e $\beta_1 = 0.75$. Para cada $t = 1, \cdots, T$ e $d = 1, \cdots, D$, para construír o 1º escenario procedemos do seguinte xeito:

- **Paso 1:** $x_{dt} = \dfrac{d + t/T}{D}$,
- **Paso 2:** v_d segue un modelo autorregresivo $AR(1)$ con $\rho = 0.6$,
- **Paso 3:** u_d segue un modelo autorregresivo $AR(1)$ e tomamos ρ parecido aos datos reais, entón $\rho = 0.4$,
- **Paso 4:** $e_{dt} \sim N(0, \sigma_{dt})$, tendo σ^2_{dt} a seguinte expresión

$$\sigma^2_{dt} = \frac{(\alpha_1 - \alpha_0)[T(d-1)+t+1]}{DT-1} + \alpha_0 \tag{3.3}$$

  onde $\alpha_0 = \text{mín}\left\{\sigma^2_{dt}\right\} = 1.35$ e $\alpha_1 = \text{máx}_{d,t}\left\{\sigma^2_{dt}\right\} = 0.23$ son respectivamente os valores mínimo e máximo calculados dos datos reais. Igual que no anterior experimento de simulación, poderíamos coller $\alpha_0 = 0.8$ e $\alpha_1 = 1.2$ seguindo o estudo de simulación do artigo *Small area estimation with spatio-temporal Fay-Herriot models.*

Finalmente, o estimador directo xérase seguindo o modelo Rao-Yu como

$$y_{dt} = \beta_0 + x_{dt}\beta_1 + v_d + u_d + e_{dt} \tag{3.4}$$

De novo, podemos construír un 2º escenario, asumindo no Paso 3 que os erros e_{dt} seguen unha distribución exponencial de parámetro $1/\sigma_{dt}$.

3.3. Axuste do modelo e estimadores

A continuación faremos a simulación por Monte Carlo dos sesgos e dos erros cadráticos medios empíricos -en adiante BIAS e EMSE- dos estimadores dos parámetros.

Para cada unha das poboacións previas (o experimento de simulación 1 e máis o experimento de simulación 2 nos seus distintos escenarios) procedemos como segue:

- **Paso 1:** Repetir $i = 1, \cdots, I$, para $I = 1000$ veces e, para $d = 1, \cdots, D$:
- **Paso 1.1:** xerar unha mostra (y_d, x_d),
- **Paso 1.2:** calcular $\widehat{\mu}_d$, pois queremos predicir $\mu_{dt} = x_{dt}\beta + u_{dt} + e_{dt}$ co EBLUP $\widehat{\mu} = x_{dt}\widehat{\beta} + \widehat{u}_{dt}$,
- **Paso 1.3:** calcular $\widehat{\beta}_j^{(i)}$ con $j = 0, 1$, $\widehat{\sigma}_u^{2(i)}$ e $\widehat{\mu}_d$.
- **Paso 2:** A saída ,para $d \in \{1, D/2, D\}$ e $j \in \{0, 1\}$, é:

$$\text{BIAS}(\widehat{\beta}_j) = \frac{1}{I}\sum_{i=1}^{I}\left(\widehat{\beta}_j^{(i)} - \beta_j\right), \quad \text{BIAS}(\widehat{\sigma}_u^2) = \frac{1}{I}\sum_{i=1}^{I}\left(\widehat{\sigma}_u^{2(i)} - \sigma_u^2\right), \text{sendo} \quad j \in \{0, 1\};$$

$$\text{EMSE}(\widehat{\beta}_j) = \frac{1}{I}\sum_{i=1}^{I}\left(\widehat{\beta}_j^{(i)} - \beta_j\right)^2, \quad \text{EMSE}(\widehat{\sigma}_u^2) = \frac{1}{I}\sum_{i=1}^{I}\left(\widehat{\sigma}_u^{2(i)} - \sigma_u^2\right)^2, \text{sendo} \quad j \in \{0, 1\};$$

$$\text{BIAS}(\widehat{\sigma}_v^2) = \frac{1}{I}\sum_{i=1}^{I}\left(\widehat{\sigma}_v^{2(i)} - \sigma_v^2\right), \quad \text{EMSE}(\widehat{\sigma}_v^2) = \frac{1}{I}\sum_{i=1}^{I}\left(\widehat{\sigma}_v^{2(i)} - \sigma_v^2\right)^2;$$

$$\text{EMSE}_d = \frac{1}{I}\sum_{i=1}^{I}\left(\widehat{\mu}_d^{(i)} - \mu_d^{(i)}\right)\left(\widehat{\mu}_d^{(i)} - \mu_d^{(i)}\right)', \text{con} \quad d \in \{1, D/2, D\};$$

$$\text{RBIAS}(\widehat{\beta}_j) = \frac{\text{BIAS}(\widehat{\beta}_j)}{\beta_j}, \quad \text{RBIAS}(\widehat{\sigma}_u^2) = \frac{\text{BIAS}(\widehat{\sigma}_u^2)}{\sigma_u^2}, \text{sendo} \quad j \in \{0, 1\};$$

$$\text{REMSE}(\widehat{\beta}_j) = \frac{\sqrt{\text{EMSE}(\widehat{\beta}_j)}}{\beta_j}, \quad \text{REMSE}(\widehat{\sigma}_u^2) = \frac{\sqrt{\text{EMSE}(\widehat{\sigma}_u^2)}}{\sigma_u^2}, \text{sendo} \quad j \in \{0, 1\};$$

$$\text{RBIAS}(\widehat{\sigma}_v^2) = \frac{\text{BIAS}(\widehat{\sigma}_v^2)}{\sigma_v^2}, \quad \text{REMSE}(\widehat{\sigma}_v^2) = \frac{\sqrt{\text{EMSE}(\widehat{\sigma}_v^2)}}{\sigma_v^2}, \text{sendo} \quad j \in \{0, 1\};$$

$$\text{MEAN}_d = \frac{1}{I}\sum_{i=1}^{I}\left(\widehat{\mu}_d^{(i)}\right), \quad \text{REMSE}_d = \frac{\sqrt{\text{EMSE}_d}}{\text{MEAN}_d}, \text{con} \quad d \in \{1, D/2, D\}.$$

Ofrecemos nos seguintes cinco cadros os resultados para cada un dos modelos cos seus distintos escenarios de simulación.

Modelo 1 - escenario de simulación 1º						
RBIAS	β_0			β_1		
T ╲ D	50	100	150	50	100	150
4	-0.24	0.47	0.47	0.33	-0.0016	-0.001
8	-0.24	0.46	0.44	0.33	0.0003	0.003
12	0.46	0.45	0.38	0.009	-0.008	0.002
Modelo 1 - escenario de simulación 2º						
RBIAS	β_0			β_1		
T ╲ D	50	100	150	50	100	150
4	0.44	0.40	0.51	0.48	0.36	0.34
8	0.37	0.32	0.48	0.29	0.37	0.15
12	0.33	0.30	0.39	0.21	0.33	0.15
Modelo 2 - escenario de simulación 1º						
RBIAS	β_0			β_1		
T ╲ D	50	100	150	50	100	150
4	-0.29	0.57	0.47	0.33	-0.016	-0.03
8	-0.31	0.48	0.44	0.33	0.004	0.01
12	0.33	0.45	0.38	0.009	-0.009	0.001
Modelo 2 - escenario de simulación 2º						
RBIAS	β_0			β_1		
T ╲ D	50	100	150	50	100	150
4	0.45	0.44	0.51	0.48	0.39	0.38
8	0.41	0.39	0.48	0.29	0.38	0.21
12	0.39	0.32	0.39	0.21	0.33	0.19

Cadro 3.1: Para β_0 e β_1, cadro cos valores de *RBIAS* para distintos valores de D e T

Modelo 1 - escenario de simulación 1º						
REMSE	β_0			β_1		
T ╲ D	50	100	150	50	100	150
4	0.24	0.16	0.13	0.32	0.28	0.24
8	0.18	0.11	0.11	0.24	0.19	0.18
12	0.08	0.09	0.06	0.23	0.11	0.07
Modelo 1 - escenario de simulación 2º						
REMSE	β_0			β_1		
T ╲ D	50	100	150	50	100	150
4	0.44	0.40	0.37	0.59	0.55	0.47
8	0.42	0.39	0.33	0.45	0.43	0.44
12	0.40	0.37	0.30	0.42	0.39	0.31
Modelo 2 - escenario de simulación 1º						
REMSE	β_0			β_1		
T ╲ D	50	100	150	50	100	150
4	0.22	0.13	0.13	0.28	0.25	0.19
8	0.16	0.11	0.10	0.21	0.14	0.11
12	0.07	0.05	0.03	0.19	0.10	0.04
Modelo 2 - escenario de simulación 2º						
REMSE	β_0			β_1		
T ╲ D	50	100	150	50	100	150
4	0.38	0.37	0.35	0.41	0.43	0.40
8	0.32	0.36	0.33	0.38	0.37	0.33
12	0.30	0.34	0.31	0.36	0.33	0.31

Cadro 3.2: Para β_0 e β_1, tomando distintos valores de D e T, cadro cos valores de *REMSE*

Modelo 1 - escenario de simulación 1º						
$RBIAS$	(σ_u, σ_v) =(0.8, 1.2)			(σ_u, σ_v) =(1, 1.2)		
T ╲ D	50	100	150	50	100	150
4	0.25	0.21	0.23	0.13	0.15	-0.01
8	0.22	0.19	0.16	0.10	0.12	-0.02
12	-0.21	0.15	0.14	-0.09	0.04	-0.02

Modelo 1 - escenario de simulación 2º						
$RBIAS$	(σ_u, σ_v) =(0.8, 1.2)			(σ_u, σ_v) =(1, 1.2)		
T ╲ D	50	100	150	50	100	150
4	0.18	0.16	0.17	0.21	0.19	0.20
8	0.17	0.19	0.18	0.17	0.17	0.19
12	0.15	0.21	0.12	0.11	0.14	0.16

Modelo 2 - escenario de simulación 1º						
$RBIAS$	(σ_u, σ_v) =(0.8, 1.2)			(σ_u, σ_v) =(1, 1.2)		
T ╲ D	50	100	150	50	100	150
4	0.15	0.20	0.21	0.11	0.13	-0.0
8	0.21	0.17	0.13	0.10	0.11	-0.02
12	-0.19	0.12	0.13	-0.09	0.02	-0.01

Modelo 2 - escenario de simulación 2º						
$RBIAS$	(σ_u, σ_v) =(0.8, 1.2)			(σ_u, σ_v) =(1, 1.2)		
T ╲ D	50	100	150	50	100	150
4	0.15	0.14	0.15	0.21	0.19	0.21
8	0.11	0.16	0.13	0.19	0.19	0.13
12	0.13	0.28	0.11	0.17	0.12	0.12

Cadro 3.3: Para σ_u e σ_v cadro cos valores de $RBIAS$ para distintos valores de D e T

Modelo 1 - escenario de simulación 1º						
$REMSE$	(σ_u, σ_v) =(0.8, 1.2)			(σ_u, σ_v) =(1, 1.2)		
T ╲ D	50	100	150	50	100	150
4	0.09	0.09	0.04	0.02	0.009	0.005
8	0.05	0.02	0.001	0.02	0.009	0.004
12	0.04	0.001	0.001	0.01	0.007	0.001

Modelo 1 - escenario de simulación 2º						
$REMSE$	(σ_u, σ_v) =(0.8, 1.2)			(σ_u, σ_v) =(1, 1.2)		
T ╲ D	50	100	150	50	100	150
4	0.13	0.10	0.10	0.15	0.019	0.015
8	0.11	0.09	0.09	0.12	0.019	0.014
12	0.11	0.06	0.07	0.11	0.017	0.012

Modelo 2 - escenario de simulación 1º						
$REMSE$	(σ_u, σ_v) =(0.8, 1.2)			(σ_u, σ_v) =(1, 1.2)		
T ╲ D	50	100	150	50	100	150
4	0.03	0.01	0.008	0.002	0.009	0.005
8	0.02	0.009	0.001	0.001	0.009	0.004
12	0.01	0.008	0.001	0.001	0.007	0.0001

Modelo 2 - escenario de simulación 2º						
$REMSE$	(σ_u, σ_v) =(0.8, 1.2)			(σ_u, σ_v) =(1, 1.2)		
T ╲ D	50	100	150	50	100	150
4	0.04	0.04	0.01	0.09	0.08	0.05
8	0.03	0.02	0.01	0.07	0.08	0.04
12	0.02	0.02	0.01	0.05	0.03	0.01

Cadro 3.4: Para σ_u e σ_v cadro cos valores de $REMSE$ para distintos valores de D e T

Modelo 1 - escenario de simulación 1º									
$REMSE_d$	$d=1$			$d=D/2$			$d=D$		
T ╲ D	50	100	150	50	100	150	50	100	150
4	0.05	0.01	0.01	0.03	0.01	0.008	0.05	0.02	0.02
8	0.01	0.01	0.009	0.007	0.01	0.005	0.04	0.01	0.001
12	0.01	0.01	0.008	0.03	0.01	0.004	0.02	0.01	0.001

Modelo 1 - escenario de simulación 2º									
$REMSE_d$	$d=1$			$d=D/2$			$d=D$		
T ╲ D	50	100	150	50	100	150	50	100	150
4	0.04	0.03	0.01	0.13	0.09	0.02	0.05	0.03	0.01
8	0.03	0.02	0.009	0.007	0.01	0.005	0.04	0.01	0.001
12	0.03	0.02	0.009	0.03	0.01	0.004	0.002	0.01	0.001

Modelo 2 - escenario de simulación 1º									
$REMSE_d$	$d=1$			$d=D/2$			$d=D$		
T ╲ D	50	100	150	50	100	150	50	100	150
4	0.005	0.002	0.001	0.07	0.04	0.004	0.05	0.03	0.001
8	0.001	0.001	0.0009	0.03	0.01	0.003	0.04	0.001	0.001
12	0.001	0.001	0.0009	0.009	0.01	0.002	0.001	0.01	0.001

Modelo 2 - escenario de simulación 2º									
$REMSE_d$	$d=1$			$d=D/2$			$d=D$		
T ╲ D	50	100	150	50	100	150	50	100	150
4	0.05	0.03	0.02	0.13	0.09	0.07	0.05	0.04	0.01
8	0.03	0.03	0.01	0.007	0.01	0.06	0.04	0.02	0.001
12	0.01	0.01	0.009	0.03	0.01	0.009	0.009	0.02	0.001

Cadro 3.5: Para $d = 1, D/2, D$, cadro cos valores de $EMSE$ do EBLUP para distintos valores de D e T

O Cadro 3.1 presenta os valores do sesgo das estimacións de β_0 e β_1 para distintos valores de D e T, para cada un dos distintos escenarios de simulación.

No Cadro 3.2 podemos ver o EMSE dos estimadores de β_0 e β_1 para distintos valores de D e T nos catro escenarios distintos. Observamos neste cadro que presenta menor EMSE os modelos nos que os erros seguen unha distribución normal en vez de exponencial. Ademais cando aumentamos o número de T o EMSE diminúe, o mesmo ocorre con D.

No Cadro 3.4 atópanse os valores do sesgo das estimacións de σ_u e σ_v con distintos valores de D e T. E, no Cadro 3.3 atópase o EMSE das estimacións de σ_u e σ_v con distintos valores de D e T.

Finalmente, observamos no Cadro 3.5, que o EMSE das estimacións diminúe en calquera das poboacións segundo aumentamos o número de áreas pequenas ou o número de instantes de tempo (trimestres). Ademais, en ambos modelos elévase o EMSE se consideramos que os erros seguen unha distribución exponencial en vez de unha distribución normal.

3.4. MSE

Estamos agora interesados no comportamento dos estimadores do MSE, é dicir, queremos comprobar a estimación analítica (explicada do Tema 2):

$$\mathrm{mse}_{dt} = \widehat{\mathrm{MSE}}_{dt} = \mathrm{mse}_d(\widehat{\mu}_{dt}) = \hat{g}_1(\hat{\sigma}_u^2) + \hat{g}_2(\hat{\sigma}_u^2) + 2\hat{g}_3(\hat{\sigma}_u^2).$$

coa estimación bootstrap mse^*_{dt}. Para esta simulación procedemos como segue:

1. Para $i = 1, \cdots, I$, repetir $I = 500$ veces.
   - 1.1. Xerar unha mostra (y_{dt}, x_{dt}), $d = 1, \cdots, D$; $t = 1, \cdots, T$.
   - 1.2. Calcular $\widehat{\beta}$ e $\widehat{\Sigma}_u$. Sendo Σ_u unha matriz diagonal $\Sigma_u = \mathrm{diag}\{\sigma_u^2\}_{1 \le d \le D}$, ou unha estrutura de correlación $AR(1)$, $\Sigma_u = \sigma_u^2 \Omega(\rho)$, no caso de tempo independente ou correlado, respectivamente.
   - 1.3. Para $d = 1, \cdots, D$ y $t = 1, \cdots, T$ calcular $\widehat{\mu}_{dt} = x_{dt}\widehat{\beta} + \widehat{u}_{dt}$, usando $\widehat{\Sigma}_u$, $\widehat{\beta}$. Calcular

$$\mathrm{MSE}^{(i)}(\widehat{\mu}) = g_1(\Sigma_u) + g_2(\Sigma_u) + g_3(\Sigma_u).$$

$$\mathrm{mse}^{(i)} = g_1(\widehat{\Sigma}_u) + g_2(\widehat{\Sigma}_u) + 2g_3(\widehat{\Sigma}_u).$$

   - 1.4. Repetir $B = 500$ veces $(b = 1, \cdots, B)$
     - 1.4.1. Xerar $u^* = \widehat{\Sigma}_u^{1/2} W_1$ e $\varepsilon^* = \Sigma_e^{1/2} W_2$, tal que $w_{1d} \in N(0,1)$ e $w_{2d} \in N(0,1)$ para cada $d = 1, \cdots, D$ e $\Sigma_e = \mathrm{diag}\{\sigma_{dt}^2\}$.
     - 1.4.2. Xerar unha mostra bootstrap (y^*_{dt}, x_{dt}) do modelo

$$y^*_{dt} = x_{dt}\widehat{\beta} + u^*_{dt} + \varepsilon^*_{dt}.$$

     - 1.4.3. Calcular μ^*_{dt}.
     - 1.4.4. Calcular $\widehat{\beta}^*$ y $\widehat{\Sigma}^*_u$.
     - 1.4.5. Para $d = 1, \cdots, D$ y $t = 1, \cdots, T$, calcular $\widehat{\mu}^*_{dt}$.

1.5 Para $d = 1, \cdots, D$ y $t = 1, \cdots, T$, calcular

$$\text{mse}_{dt}^{*,1(i)} = \frac{1}{B}\sum_{b=1}^{B}(\widehat{\mu}_{dt}^{*(ib)} - \mu_{dt}^{*(ib)})$$

$$mse^{*,2(i)} = g_1(\widehat{\sigma}_u^2) + g_2(\widehat{\sigma}_u^2) + E^*[(\widehat{\mu}_{dt,E}^* - \widehat{\mu}_{dt,B}^*)(\widehat{\mu}_{dt,E}^* - \widehat{\mu}_{dt,B}^*)'].$$

$$mse^{*,3(i)} = 2[g_1(\widehat{\sigma}_u^2) + g_2(\widehat{\sigma}_u^2)] - E^*[g_1(\widehat{\sigma}_u^{2,*}) + g_2(\widehat{\sigma}_u^{2,*})] + E^*[(\widehat{\mu}_{dt,E}^* - \widehat{\mu}_{dt,B}^*)(\widehat{\mu}_{dt,E}^* - \widehat{\mu}_{dt,B}^*)'].$$

2. Para $d = 1, \cdots, D$ y $t = 1, \cdots, T$, ler os valores $EMSE_{dt}$ de simulación anterior

3. Saída:

$$B_{dt} = \frac{1}{I}\sum_{i=1}^{I}(\text{mse}_{dt}^{(i)} - \text{EMSE}_{dt}),\ B_{dt}^* = \frac{1}{I}\sum_{i=1}^{I}(\text{mse}_{dt}^{*(i)} - \text{EMSE}_{dt}).$$

$$RB_{dt} = \frac{\sqrt{B_{dt}}}{\text{EMSE}_{dt}},\ RB_{dt}^* = \frac{\sqrt{B_{dt}^*}}{\text{EMSE}_{dt}}.$$

$$E_{dt} = \frac{1}{I}\sum_{i=1}^{I}(\text{mse}_{dt}^{(i)} - \text{EMSE}_{dt})^2,\ E_{dt}^* = \frac{1}{I}\sum_{i=1}^{I}(\text{mse}_{dt}^{*(i)} - \text{EMSE}_{dt})^2,$$

$$RE_{dt} = \frac{\sqrt{E_{dt}}}{\text{EMSE}_{dt}},\ RE_{dt}^* = \frac{\sqrt{E_{dt}^*}}{\text{EMSE}_{dt}}.$$

Seguidamente faremos táboas e gráficos de caixa para comparar os resultados dos estimadores do MSE, nos dous escenarios das simulacións, descritos nas seccións 3.1 e 3.2, para distintos valores dos parámetros. E observamos, por unha banda, que usando as expresións analíticas obtemos peores estimacións do MSE que usando o método do bootstrap paramétrico; e, pola outra banda, que as estimacións do MSE empeoran no escenario 2 de cada un dos modelos.

Modelo 1- Escenario 1						
RB	σ_u =0.8, σ_v =1.2			σ_u =1, σ_v =1.2		
	$D = 50$	$D = 100$	$D = 150$	$D = 50$	$D = 100$	$D = 150$
mse	0.19	0.08	0.17	0.14	0.03	0.11
$mse^{*,1}$	-0.14	-0.08	-0.03	-0.10	-0.05	-0.02
$mse^{*,2}$	-0.09	-0.06	-0.04	-0.08	-0.06	-0.03
$mse^{*,3}$	-0.04	-0.03	-0.01	-0.03	-0.01	-0.01

Modelo 1-Escenario 2						
RB	σ_u =0.8, σ_v =1.2			σ_u =1, σ_v =1.2		
	$D = 50$	$D = 100$	$D = 150$	$D = 50$	$D = 100$	$D = 150$
mse	0.27	0.12	0.25	0.19	0.11	0.17
$mse^{*,1}$	0.11	0.08	0.09	0.05	-0.04	0.08
$mse^{*,2}$	0.13	0.10	0.12	0.09	0.05	0.03
$mse^{*,3}$	0.15	0.09	0.10	0.05	0.04	0.02

Modelo 2- Escenario 1						
RB	σ_u =0.8, σ_v =1.2			σ_u =1, σ_v =1.2		
	$D = 50$	$D = 100$	$D = 150$	$D = 50$	$D = 100$	$D = 150$
mse	0.15	0.09	0.13	0.13	0.02	0.09
$mse^{*,1}$	-0.11	-0.07	-0.02	-0.09	-0.04	-0.01
$mse^{*,2}$	-0.09	-0.05	-0.02	-0.04	-0.02	-0.01
$mse^{*,3}$	-0.04	-0.02	-0.01	-0.03	-0.03	-0.01

Modelo 2- Escenario 2						
RB	σ_u =0.8, σ_v =1.2			σ_u =1, σ_v =1.2		
	$D = 50$	$D = 100$	$D = 150$	$D = 50$	$D = 100$	$D = 150$
mse	0.17	0.10	0.15	0.14	0.09	0.09
$mse^{*,1}$	-0.19	-0.15	-0.13	-0.11	-0.09	-0.05
$mse^{*,2}$	-0.21	-0.18	-0.14	-0.16	-0.11	-0.08
$mse^{*,3}$	-0.23	-0.17	-0.12	-0.13	-0.06	-0.03

Escenario 1						
RE	σ_u =0.8, σ_v =1.2			σ_u =1, σ_v =1.2		
	$D=50$	$D=100$	$D=150$	$D=50$	$D=100$	$D=150$
mse	0.31	0.33	0.39	0.29	0.25	0.22
$mse^{*,1}$	0.16	0.14	0.11	0.14	0.12	0.09
$mse^{*,2}$	0.12	0.14	0.13	0.17	0.09	0.13
$mse^{*,3}$	0.10	0.09	0.12	0.15	0.07	0.09

Escenario 2						
RE	σ_u =0.8, σ_v =1.2			σ_u =1, σ_v =1.2		
	$D=50$	$D=100$	$D=150$	$D=50$	$D=100$	$D=150$
mse	0.39	0.41	0.54	0.37	0.43	0.49
$mse^{*,1}$	0.27	0.23	0.19	0.18	0.14	0.11
$mse^{*,2}$	0.22	0.17	0.20	0.17	0.12	0.11
$mse^{*,3}$	0.19	0.14	0.16	0.13	0.10	0.10

Escenario 1						
RE	σ_u =0.8, σ_v =1.2			σ_u =1, σ_v =1.2		
	$D=50$	$D=100$	$D=150$	$D=50$	$D=100$	$D=150$
mse	0.21	0.24	0.25	0.24	0.22	0.24
$mse^{*,1}$	0.12	0.18	0.13	0.14	0.18	0.09
$mse^{*,2}$	0.14	0.22	0.16	0.13	0.15	0.10
$mse^{*,3}$	0.11	0.17	0.12	0.11	0.18	0.09

Escenario 2						
RE	σ_u =0.8, σ_v =1.2			σ_u =1, σ_v =1.2		
	$D=50$	$D=100$	$D=150$	$D=50$	$D=100$	$D=150$
mse	0.51	0.54	0.57	0.53	0.61	0.64
$mse^{*,1}$	0.42	0.39	0.33	0.31	0.28	0.23
$mse^{*,2}$	0.38	0. 35	0.30	0.27	0.25	0.20
$mse^{*,3}$	0.36	0.32	0.28	0.24	0.22	0.21

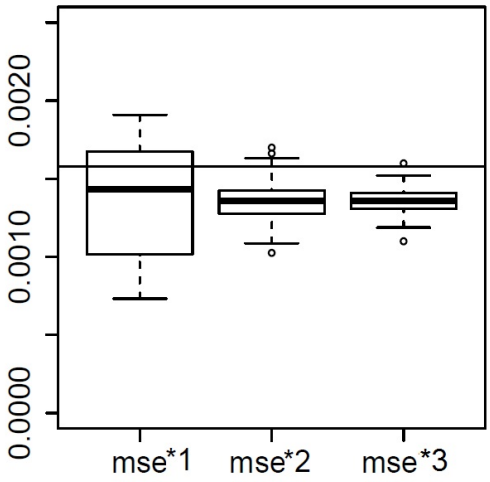

Figura 3.1: Modelo 1 - escenario 1, con T=12 e D=150

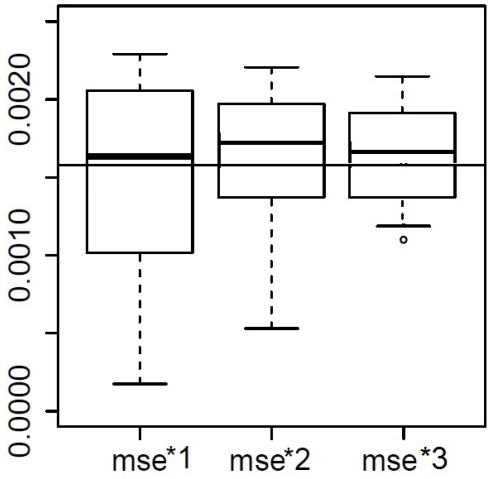

Figura 3.2: Modelo 1 - escenario 2, con T=12 e D=150

Nas Figuras 3.1, 3.2, 3.3 e 3.4 represéntanse todos os estimadores do MSE ($mse^{*,1}$, $mse^{*,2}$ e $mse^{*,3}$) ademais dunha liña horizontal que contén o valor real (o EMSE). Tendo en vista estas figuras, podemos sacar a conclusión de cal é o mellor estimador do MSE segundo o modelo e o escenario. Así, no modelo 1 e co escenario 1º, vemos que o mellor estimador é $mse^{*,1}$ pois é o que máis se aproxima ao valor real. O mesmo pasa no 2º escenario, aínda que ben é certo que as tres aproximacións do mse son peores que as do escenario 1º. En canto ao modelo 2, podemos facer conclusións análogas ao modelo 1, e ademais, dicir que se aproxima mellor o EMSE cando se usa este modelo.

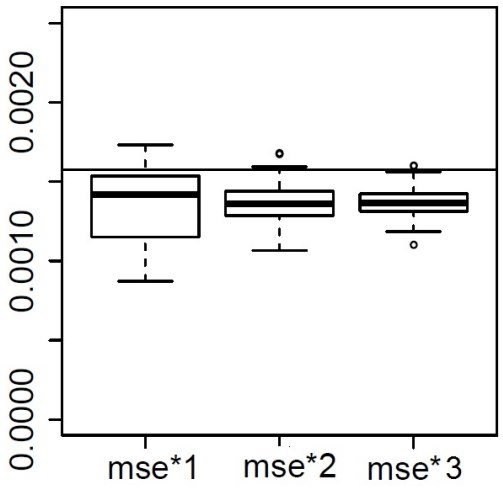

Figura 3.3: Modelo 2 - escenario 1, con T=12 e D=150

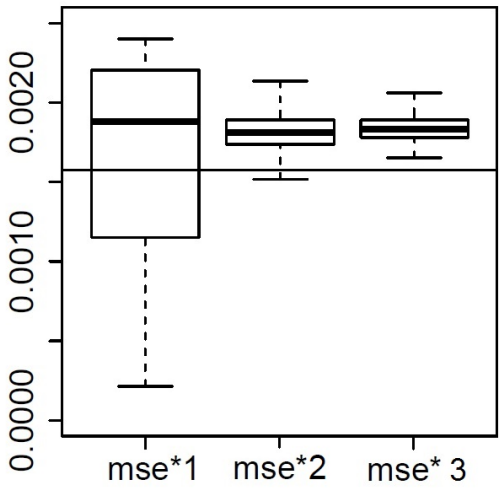

Figura 3.4: Modelo 2 - escenario 2, con T=12 e D=150

Capítulo 4

Aplicación aos datos reais

Neste capítulo aplicaremos aos datos reais toda a metodoloxía detallada nos capítulos previos.

Lembremos que o noso obxectivo é estudar o erro cadrático medio dos dous modelos de área con efecto temporal vistos anteriormente, para estimar o número de ocupados da EPA nas áreas pequenas determinadas pola CNAE a dous díxitos e por sexo. Neste nivel de descrición o tamaño de mostra é pequeno polo que precisamos de variables auxiliares: número de afilicados á SS e número de contratos novos rexistrados. A nosa variable resposta será o número de ocupados da EPA. Dispomos dos datos trimestrais para estas variables de interese dende o terceiro trimestre do ano 2009 ata o cuarto trimestre do ano 2013.

A análise exploratoria destes datos quedou feita no Capítulo 1, capítulo no que tamén se describiu a orixe dos datos. No Capítulo 2, tratamos as dúas extensións do modelo Fay-Herriot e vimos a estimación dos seus parámetros, así como a estimación do MSE, vimos as súas expresións analíticas e o procedemento bootstrap paramétrico con distintas extensións. Ademais, no Capítulo 3, fixemos un estudo de simulación e agora aplicaremos todo o anterior aos datos reais, obtendo os resultados para cada un dos modelos por separado, facendo comparacións e chegando a conclusións.

4.1. Modelo 1

Apliquemos o modelo a nivel de área con efectos de tempo independentes e as estimacións vistas no Capítulo 2 aos datos presentados no Capítulo 1.

Lembremos que, en vista da análise exploratoria realizada no Capítulo 1, decidimos considerar a variable resposta e as variables auxiliares "número de afiliados á SS" e "número de contratos" en forma logarítmica.

Lembremos tamén, que o número de dominios é $D = 86$, m_d é o número de trimestres comprendidos entre o 3^o trimestre do ano 2009 ata o 4^o trimestre do ano 2012, ou sexa, $m_d = 18$.

Os resultados que se obteñen para os efectos fixos e aleatorios xa estimado o modelo 1 figuran no seguinte Cadro 4.1 no cal se mostran os parámetros de regresión e a compoñente da varianza.

Parámetros	Valor	Erro estándar	p-valor
β_0	1.151	0.146	<0.01
β_1	$-9{,}09 \cdot 10^{-6}$	$1{,}89 \cdot 10^{-5}$	<0.01
β_2	0.882	0.028	<0.01
σ_u	0.207	0.0153	<0.01

Cadro 4.1: Estimación dos parámetros do modelo 1 co seu p-valor asociado

A estimación dos parámetros do modelo 1 os seus p-valores tamén está feita coa función "fit.saery" paquete *saery*. En vista do Cadro 4.1 todos os parámetros son significativos dentro do modelo para os contrastes de significatividade:
$H_0: \quad \beta_i = 0, \quad H_1: \quad \beta_i \neq 0$ e $H_0: \quad \sigma_u = \sigma_0, \quad H_1: \quad \sigma_u > \sigma_0$.

Móstrase na Figura 4.1 o gráfico do estimador directo -ecuación (1.1)- fronte ao obtido co modelo. Ponse de manifesto maior discrepancia entre o estimador directo e o obtido mediante o modelo 1 con valores pequenos no nº de ocupados.

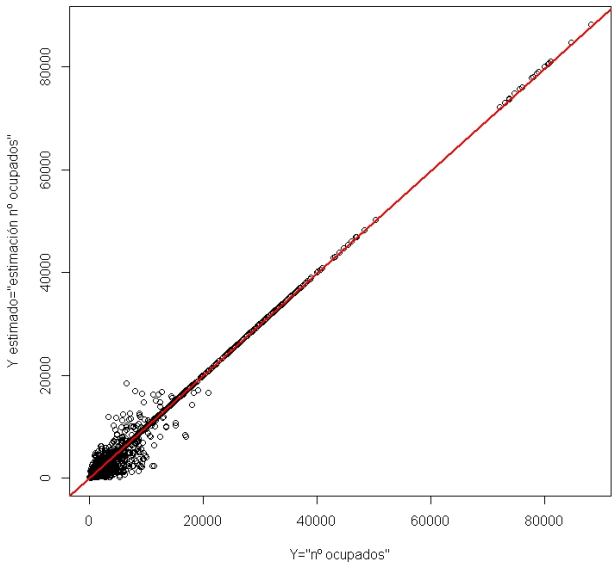

Figura 4.1: Diagrama de dispersión estimador directo fronte ao obtido do modelo 1

Na Figura 4.2 móstrase o estimador directo e o estimador obtido co modelo 1 (cos datos ordenados segundo o tamaño de mostra) para os 4 primeiros trimestres.

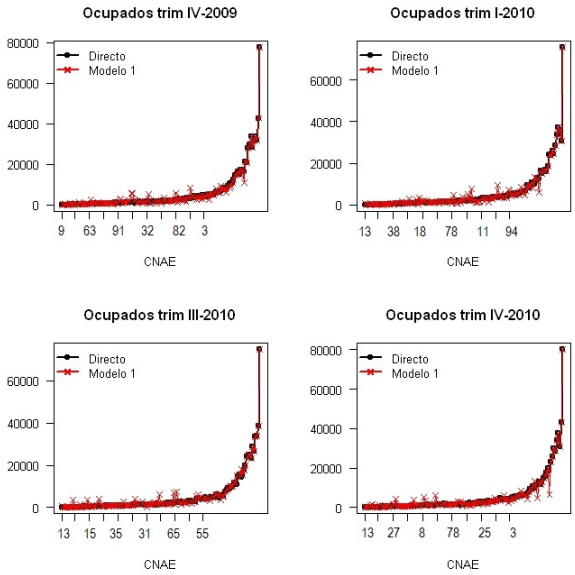

Figura 4.2: Estimador directo e o estimador obtido con modelo 1 para 4 trimestres. Ordenados polo tamaño de mostra

Na Figura 4.3 preséntase un zoom do mesmo para observar como son as estimacións do modelo para tamaños de mostra pequenos xa que cando o tamaño de mostra é grande os estimadores directos e baseados no modelo funcionan de xeito similar.

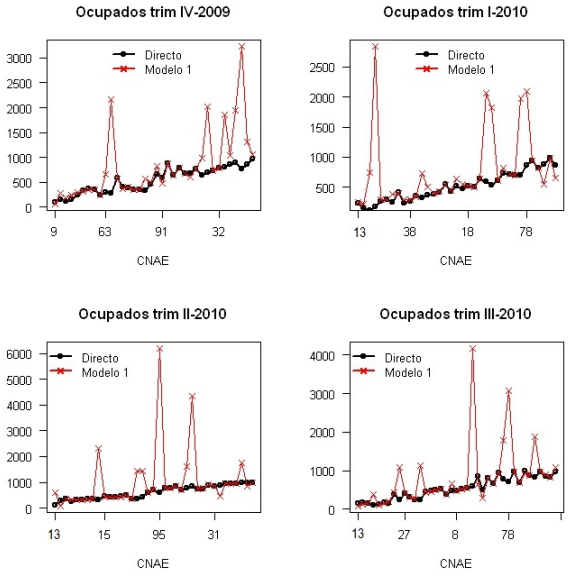

Figura 4.3: Zoom estimador directo e o estimador obtido con modelo 1 para os 4 primeiros trimestres

Na análise exploratoria, observamos uns coeficientes de variación moi altos para algunhas observacións do estimador directo. Convén comprobar que o modelo foi capaz de reducir o valor destes coeficientes. Para iso, na Figura 4.4 presentamos o coeficiente de variación do estimador directo fronte ao observado co modelo 1, o cal obtivemos coa expresión seguinte:

$$\widehat{CV} = \frac{\sqrt{\widehat{MSE}\left(\hat{\mu}_{dt}\right)}}{\hat{\mu}_{dt}} \tag{4.1}$$

Observando a Figura 4.4 podemos concluír que os coeficientes de variación diminuíron, e diminuíron en maior medida para os datos con poucas observacións, co cal avanzamos na estimación grazas ao modelo 1. O estimador do MSE representado é o que mellor funcionaba para o estudo de simulación, ou sexa, $mse^{*,1}$. Conclúese a partires das gráficas que para tamaños de mostra pequenos o estimador EBLUP do modelo 1 ten menos erro, polo tanto é máis adecuado.

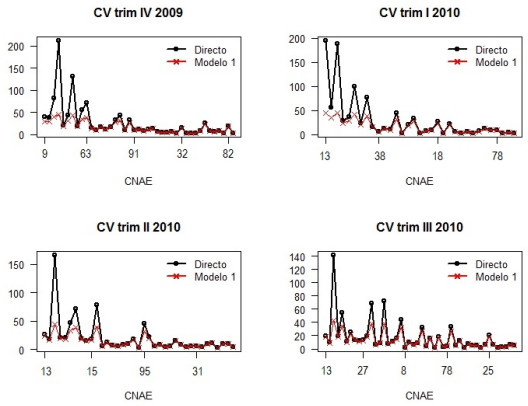

Figura 4.4: Coeficiente de variación fronte ao estimado obtido co modelo 1 para 4 trimestres

Na Figura 4.5 represéntase a evolución de mse, $mse^{*,1}$, $mse^{*,2}$ e $mse^{*,3}$ en % no 4º trimestre do ano 2009. Os datos están ordenados por tamaño de mostra, e obsérvase como diminúen as estimacións do MSE segundo o tamaño de mostra é máis pequeno.

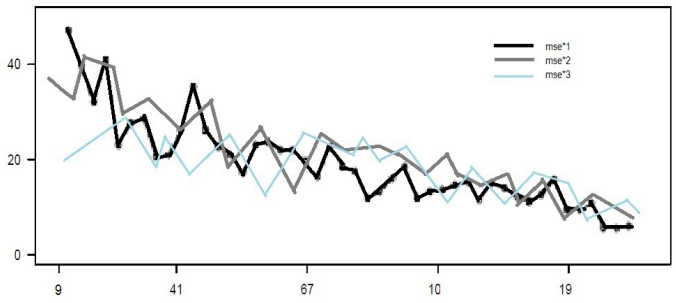

Figura 4.5: $mse^{*,1}$, $mse^{*,2}$,$mse^{*,3}$ en % obtidos co modelo 1 para o 4º trimestre do ano 2013

Na seguinte sección, o modelo 2 tratará de diminuír un pouco máis este erro, para isto considerará que os efectos aleatorios correlados e que seguen un proceso AR(1), o cal ten a vantaxe de ter en conta que o número de ocupados nun trimestre dependerá do número de ocupados no anterior trimestre. Presentaremos os resultados de forma análoga a como fixemos con este modelo 1.

4.2. Modelo 2

De xeito análogo a como procedemos co modelo 1, observamos a continuación no Cadro 4.2 os resultados que se obteñen para os efectos fixos e aleatorios xa estimado o modelo 2 e vemos que todos os parámetros son significativos. A estimación dos parámetros do modelo 2 os seus p-valores tamén está feita coa función "fit.saery" paquete *saery*. E, os contrastes de significatividade son:

$H_0: \quad \beta_i = 0, \quad H_1: \quad \beta_i \neq 0$ e $H_0: \quad \sigma_u = \sigma_0, \quad H_1: \quad \sigma_u > \sigma_0$.

Parámetros	Valor	Erro estándar	p-valor
β_0	0.9	0.148	$<$0.01
β_1	$-1{,}75 \cdot 10^{-6}$	$2{,}08 \cdot 10^{-5}$	$<$0.01
β_2	0.173	0.013	$<$0.01
σ_u	0.462	0.047	$<$0.01

Cadro 4.2: Estimación dos parámetros do modelo 2 co seu p-valor asociado

Móstrase na Figura 4.6 o gráfico do estimador directo fronte ao obtido co modelo 2 cos datos ordenados segundo o tamaño de mostra. Ao igual que pasaba na Figura 4.1 do modelo 1, ponse de manifesto maior discrepancia entre o estimador directo e o obtido mediante o modelo 2 con valores pequenos no número de ocupados.

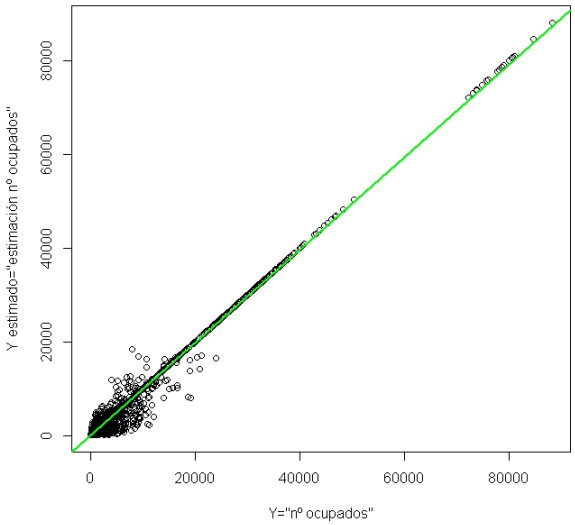

Figura 4.6: Diagrama de dispersión estimador directo fronte ao obtido do modelo 2

Na Figura 4.7 móstrase o estimador directo e o estimador obtido co modelo 2 para os 4 primeiros trimestres.

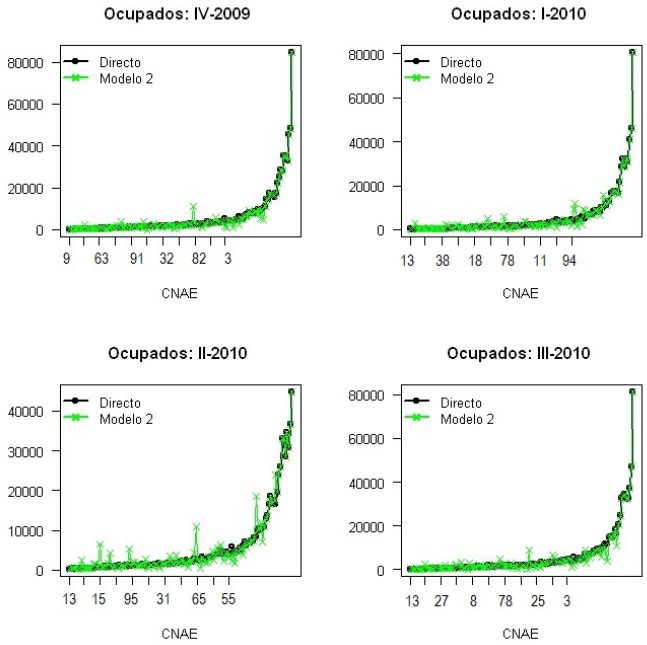

Figura 4.7: Estimador directo e o estimador obtido co modelo 2 para 4 trimestres

Na Figura 4.8 represéntase a evolución de $mse^{*,1}$, $mse^{*,2}$ e $mse^{*,3}$ no 4º trimestre do ano 2013.

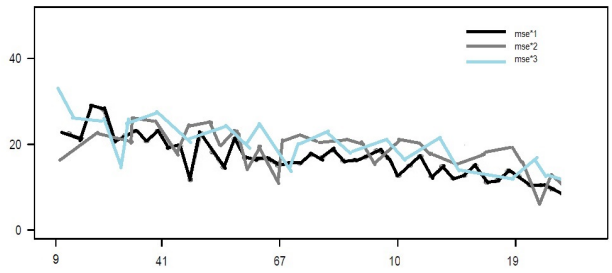

Figura 4.8: $mse^{*,1}$,$mse^{*,2}$,$mse^{*,3}$ en % obtidos co modelo 2 para o 4º trimestre do 2009

Finalmente, para comparar os resultados obtidos de ambos modelos, ver se a evolución dos modelos é a correcta e que avanzamos correctamente mellorando os resultados, observaremos os coeficientes de variación de catro trimestres por ambos modelos. Para iso, móstrase na Figura 4.9 a evolución do coeficiente de variación directo (en negro) e o dos obtidos cos modelos 1 (en vermello) e 2 (en verde). Pódese ver a mellora entre os coeficientes de variación dos modelos estudados.

Dicir que o estimador do MSE representado nos dous casos é o que mellor funcionaba para o estudo de simulación, ou sexa, $mse^{*,1}$. Así, podemos concluír afirmando, que observando a Figura 4.9, confirmamos que o modelo 2 mellora os coeficientes de variación.

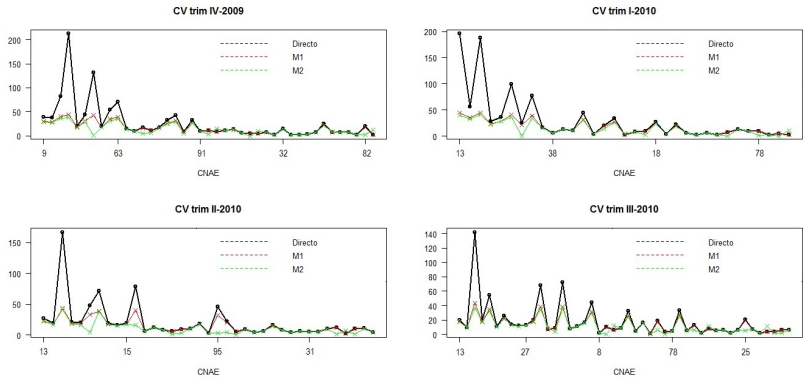

Figura 4.9: Comparación da evolución coeficiente de variación para 4 trimestres nos dous modelos

Recapitulando, neste último capítulo, aplicamos aos datos reais toda a metodoloxía descrita nos capítulos previos. Así, consideramos dous modelos de área con efecto temporal, os que chamamos modelo 1 e o modelo 2. Para cada un dos modelos, presentamos a estimación dos parámetros co seu p-valor asociado, representamos o estimador directo frente al EBLUP estudado nesta memoria e representamos os coeficientes de variación. En resumo, o MSE do estimador EBLUP en ambos modelo aproxímase por bootstrap, tomando como estimador $mse^{*,1}$. Ademais, vimos que cando consideramos os efectos aleatorios de tempo correlados en vez de independentes os coeficientes de variación melloran, o que implica unha evolución correcta dos dous modelos.

Conclusións e futuras liñas

Ao longo deste traballo vimos que se ofrece mellores estimacións do MSE usando a método bootstrap paramétrico en vez de usando as expresións analíticas. Tamén sacamos a conclusión de que ao incluír no modelo os efectos de tempo correlados mellórase respecto á inclusión dos efectos de tempo independentes, isto tamén se debe á natureza dos datos pois se efectivamente os datos son correlados temporalmente isto cáptao mellor o modelo 2.

En canto a posibles futuras investigacións deste traballo, poderíase usar outros métodos de remostraxe para estimar o MSE como, por exemplo, o método jackknife. Vexamos en que consiste para xustificar a futura ampliación.

Como vimos, temos a expresión medida do erro o MSE de *Prasad e Rao* (1990) detallada no Capítulo 2, sección 3º:

$$MSE(\hat{\mu}_{dt}^{EBLUP}) \approx g_1(\sigma_u^2) + g_2(\sigma_u^2) + g_3(\sigma_u^2) = M_1 + M_2 \tag{4.2}$$

sendo $M_1 = g_1(\sigma_u^2)$ e $M_2 = g_2(\sigma_u^2) + g_3(\sigma_u^2)$.

Jiang, Lahiri e Wan (2002) propuxeron unha estimación jackknife do MSE que consiste na avaliación explícita de $g_2(.)$ e $g_3(.)$ da fórmula anterior, pero require a derivación dos termos de $g_1(.)$. Chegados a este punto, aplicaron a idea jackknife de *Tukey* para conseguir o estimador jackknife de M_2.

Resumindo, no método jackknife disponse de tantas submostras como clusters teña a mostra, xa que se obteñen por sucesivas eliminacións de clusters na mostra orixinal. Para casa submostra, defínense novos pesos e calcúlase o estimador. Logo obtense o MSE.

Glosario de siglas usadas na memoria

BIAS	Sesgo empírico
BLUP	Mellor preditor lineal insesgado
CNAE	Clasificación Nacional de Actividades Económicas
CV	Coeficiente de Variación
EBLUP	Mellor estimación linear insesgada empírica
EMSE	Erro cadrático medio empírico
EPA	Enquisa da Poboación Activa
EUSTAT	Instituto Vasco de Estadística
IGE	Instituto Galego de Estatística
INE	Instituto Nacional de Estadística
ISTAC	Instituto Canario de Estadística
MSE	Erro cadrático medio
REML	Método de máxima verosimilitude restrinxida
SAE	Estimación en Áreas Pequenas
SS	Seguridade Social
TFM	Traballo de Fin de Grao

Apéndice A

Clasificación Nacional de Actividades Económicas (CNAE 09)

Código	**Descrición**	**Sector**
01	Agricultura, gandaría, caza e servizos relacionados con elas	Primario
02	Silvicultura e explotación forestal	Primario
03	Pesca e acuicultura	Primario
05	Extracción de antracita, hulla e lignito	Industria
06	Extracción de cru de petróleo e gas natural	Industria
07	Extracción de minerais metálicos	Industria
08	Outras industrias extractivas	Industria
09	Actividades de apoio ás industrias extractivas	Industria
10	Industria da alimentación	Industria
11	Fabricación de bebidas	Industria
12	Industria do tabaco	Industria
13	Industria téxtil	Industria
14	Confección de roupa de vestir	Industria
15	Industria do coiro e do calzado	Industria
16	Industria da madeira e da cortiza, agás mobles	Industria
17	Industria do papel	Industria
18	Artes gráficas e reprodución de soportes gravados	Industria
19	Coqueterías e refinación de petróleo	Industria
20	Industria química	Industria
21	Fabricación de produtos farmacéuticos	Industria
22	Fabricación de produtos de caucho e plásticos	Industria
23	Fabricación doutros produtos minerais non metálicos	Industria
24	Metalurxia	Industria

25	Fabricación de produtos metálicos, agás maquinaria e equipamento	Industria
26	Fabricación de produtos informáticos, electrónicos e ópticos	Industria
27	Fabricación de material e equipamento eléctrico	Industria
28	Fabricación de maquinaria e equipamento n.c.n.	Industria
29	Fabricación de vehículos de motor, remolques e semirremolques	Industria
30	Fabricación doutro material de transporte	Industria
31	Fabricación de mobles	Industria
32	Outras industrias manufactureiras	Industria
33	Reparación e instalación de maquinaria e equipamento	Industria
35	Fornecemento de enerxía eléctrica, gas, vapor e aire acondicionado	Industria
36	Captación, depuración e distribución de auga	Industria
37	Recolla e tratamento de augas residuais	Industria
38	Recolla, tratamento e eliminación de residuos	Industria
39	Actividades de incontaminación e outros servizos de xestión de residuos	Industria
41	Construción de edificios	Construción
42	Enxeñaría civil	Construción
43	Actividades de construción especializada	Construción
45	Venta e reparación de vehículos de motor e motocicletas	Servizos
46	Comercio por xunto e intermediarios do comercio, salvo de vehículos de motor	Servizos
47	Comercio a retallo, salvo de vehículos de motor e motocicletas	Servizos
49	Transporte terrestre e por tubaxe	Servizos
50	Transporte marítimo e por vías navegables interiores	Servizos
51	Transporte aéreo	Servizos

52	Almacenamento e actividades anexas ao transporte	Servizos
53	Actividades postais e de correos	Servizos
55	Servizos de aloxamento	Servizos
56	Servizos de comidas e bebidas	Servizos
58	Edición	Servizos
59	Actividades cinematográficas, de vídeo e de programas de televisión	Servizos
60	Actividades de programación e emisión de radio e televisión	Servizos
61	Telecomunicacións	Servizos
62	Programación, consultaría e outras actividades relacionadas coa informática	Servizos
63	Servizos de información	Servizos
64	Servizos financeiros, agás seguros e fondos de pensións	Servizos
65	Seguros, reaseguros e fondos de pensións, agás seguranza social obrigatoria	Servizos
66	Actividades auxiliares aos servizos financeiros e aos seguros	Servizos
68	Actividades inmobiliarias	Servizos
69	Actividades xurídicas e de contabilidade	Servizos
70	Actividades das sedes centrais	Servizos
71	Servizos técnicos de arquitectura e enxeñaría	Servizos
72	Investigación e desenvolvemento	Servizos
73	Publicidade e estudos de mercado	Servizos
74	Outras actividades profesionais, científicas e técnicas	Servizos
75	Actividades veterinarias	Servizos
77	Actividades de aluguer	Servizos
78	Actividades relacionadas co emprego	Servizos
79	Actividades de axencias de viaxes, operadores turísticos	Servizos

80	Actividades de seguranza e investigación	Servizos
81	Servizos a edificios e actividades de xardinaría	Servizos
82	Actividades administrativas de oficina e outras actividades auxiliares ás empresas	Servizos
84	Administración pública e defensa	Servizos
85	Educación	Servizos
86	Actividades sanitarias	Servizos
87	Asistencia en establecementos residenciais	Servizos
88	Actividades de servizos sociais sen aloxamento	Servizos
90	Actividades de creación, artísticas e espectáculos	Servizos
91	Actividades de bibliotecas, arquivos, museos e outras actividades culturais	Servizos
92	Actividades de xogos de azar e apostas	Servizos
93	Actividades deportivas, recreativas e de entretemento	Servizos
94	Actividades asociativas	Servizos
95	Reparación de ordenadores, efectos persoais e artigos de uso doméstico	Servizos
96	Outros servizos persoais	Servizos
97	Actividades dos fogares como empregadores de persoal doméstico	Servizos
98	Actividades dos fogares como produtores de bens e servizos para uso propio	Servizos
99	Actividades de organizacións e organismos extraterritoriais	Servizos

Apéndice B

Descrición dos paquetes usados en R

Paquete SURVEY: No paquete survey proporciónanse ferramentas para asistir ao usuario en todas as etapas do desenvolvemento dunha enquisa, dende o deseño ata a análise final. Con este paquete podemos traballar con resumos estadísticos, probas de dúas mostras, modelos lineais xenerais, modelos loglineais, mostras de estudo desigualmente ponderadas, postestratificación, calibración, deseños de submostraxe bifásicos, compoñentes principais, análise factorial, ···

Na nosa memoria, as funcións deste paquete que usamos son "svydesign" e " svytotal":

- A función "svydesign" serve para especificar os deseños de mostras complexas. Combina un marco de datos e todo o deseño da enquisa necesaria para analizala. Estes obxectos úsanse polo deseño da enquisa e as funcións de resumo.
- A función " svytotal" estima a poboación total.

No enderezo http://r-survey.r-forge.r-project.org/survey/ pódese atopar toda a información relativa a este paquete.

Paquete SAERY: Este paquete contén un conxunto moi completo de funcións para calcular as estimacións EBLUP e os seus erros cadráticos medios. Todas as estimacións deste paquete baséanse no modelo lineal mixto proposto por *Rao e Yu* en 1994 (modelo que presentamos no Capítulo 2, sección 2). O método REML é o que se usa para axustar este modelo.

Son dúas as principais funcións que usamos deste paquete. A función "eblup.saery" calcula o EBLUP e o MSE para un modelo. Previamente revísase o modelo coa función " fit.saery", a cal da os axustes dos parámetros do modelo.

Bibliografía

[1] Battese, G. E., Harter, R. M., and Fuller, W. A. (1988). *An error component model for prediction of county crop areas using survey and satelite data.* Journal of the American Statistical Association.

[2] Brackstone, G. J. (1987). *Small area data: policy issues and technical challenges.*

[3] Celdrán Bouzas, D., López-Vizcaíno, E., Lombardía M.J. y González-Manteiga, W. (2012). *Estimación trimestral en áreas pequeñas de los ocupados en la Encuesta de Población Activa según su actividad económica.* Trabajo Fin de máster, Máster en Técnicas estadísticas interuniversitario UDC-USC-UVigo.

[4] Chandra, H. and Chambers, R. (2009). *Multipurpose small area estimation. Journal of Oficial Statistics.*

[5] Chaudhuri, A. (1994). *Small domain statistics: A review.* Statistica Neerlandica.

[6] Choudhry, G.H., and Rao, J.N.K. (1989). *Small area estimation using models that combine time series and cross-sectional data.* Proceedings of the Statistics Canada Symposium on Analysis of Data in Time, eds. A.C. Singh and P Whifiridge, 67-74.

[7] Das, K., Jiang, J., and Rao, J. N. K. (2004). *Mean squared error of empirical predictor.* The Annals of Statistics.

[8] Domínguez R. , Lombardía M.J., González Manteiga W. y Prada-Sánchez J.M. (2009) *Estimación de áreas pequeñas: el ingreso medio mensual por comarcas en los hogares gallegos.* Instituto Galego de Estatística.

[9] Esteban, M.D., Morales, D., Perez, A., Santamaria, L. (2012). Small area estimation of poverty proportions under area-level time models. Computational Statistics and Data Analysis, 56 (10), 2840-2855.

[10] Fay R. E. y R. A. Herriot. (1979) *Estimates of income for small places: an application of james-stein procedures to census.* Journal of the American Statistical Association. 85, 398-409.

[11] Hall, P. and Maiti, T. (2006). *On parametric bootstrap methods for small-area prediction.* Journal of the Royal Statistical Society.

[12] Henderson C. R. (1975) *Best Linear Unbiased Estimation and Prediction under a Selection Model* International Biometric Society.

[13] Herrador, M. , Morales, D., Estaban, M.D., Sánchez, A., Santamaría, L., Marhuenda, Y., Pérez, A. y Molina I. (2009). *Estimadores de áreas pequeñas basados en modelos para la Encuesta de Población Activa.* Estadística Española, 51 , $133 - 172$.

[14] Jiang, J. and Lahiri, P. (2006). *Mixed model prediction and small area estimation.* Test, 15, $1 - 96$.

[15] Jiang, Lahiri and Wan (2002). *A unified jackknife theory for empirical best prediction with M-estimation.* The Annals of Statistics.

[16] 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.

[17] López-Vizcaíno, E., Lombardía, M.J. and Morales, D. (2013). *Multinomial-based small area estimation of labour force indicators.* Statistical Modelling, 13, 153 – 178.

[18] Malec D., Davis WW., Cao (1999) *Model-based small area estimates of overweight prevalence using sample selection adjustment.*

[19] Marhuenda, Y. Molina, I. and Morales, D. (2013). *Small area estimation with spatio-temporal Fay-Herriot models.* Computational Statistics and Data Analysis, 58, 308 – 325.

[20] Pfeffermann, D. and Tiller,R. (2002) *Bootstrap approximation to prediction mse for state-space model with estimaed parameters.* Hebrew University and University of Southampton, and Bureau of Labor Statistics, Washington, DC.

[21] Prasad and Rao (1990). *The estimation of the mean squared error of small-area estimators.* Journal of the American Statistical Association, 85, 163 – 171.

[22] Rao, J. N. K. (2003) *Small area estimation.* New Jersey, Wiley.

[23] Rao, J. N. K. *Jackknife and Bootstrap Methods for Small Area Estimation.* Section on Survey Research Methods. 2925-2929.

[24] 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.

[25] Robinson, G.K. (1991). *That BLUP is a Good Thing: The Estimation of Random Effects.* Statistical Science.

[26] Saei, A. and Chambers, R. (2003).*Small area estimation under linear an generalized linear mixed models with time and area efects.* Technical report, Southampton, UK.

[27] Serie Documentos Metodológicos, Observatorio Social, N°1 (2013). *Procedimiento de Cálculo de la Tasa de Pobreza a nivel Comunal, mediante la aplicación Metodología SAE.*

[28] W. González-Manteiga, M. J. Lombardía, I.Molina, D. Morales, L. Santamaría. (2008) *Analytic and bootstrap approximation of prediction errors under a multivariate Fay-Herriot model*, Computational Statistics and Data Analysis, 52, 5242 – 5252.