R y lme4: Uniendo fuerzas para la Estimación en Áreas Pequeñas

Naomi Diz-Rosales[1], María José Lombardía[1], Domingo Morales[2].

[1]Universidade da Coruña, CITIC, Spain.
[2]Universidad Miguel Hernández de Elche, IUCIO, Spain.

X Xornada de Usuarios de R en Galicia.
18/10/2023.

1 AÑO

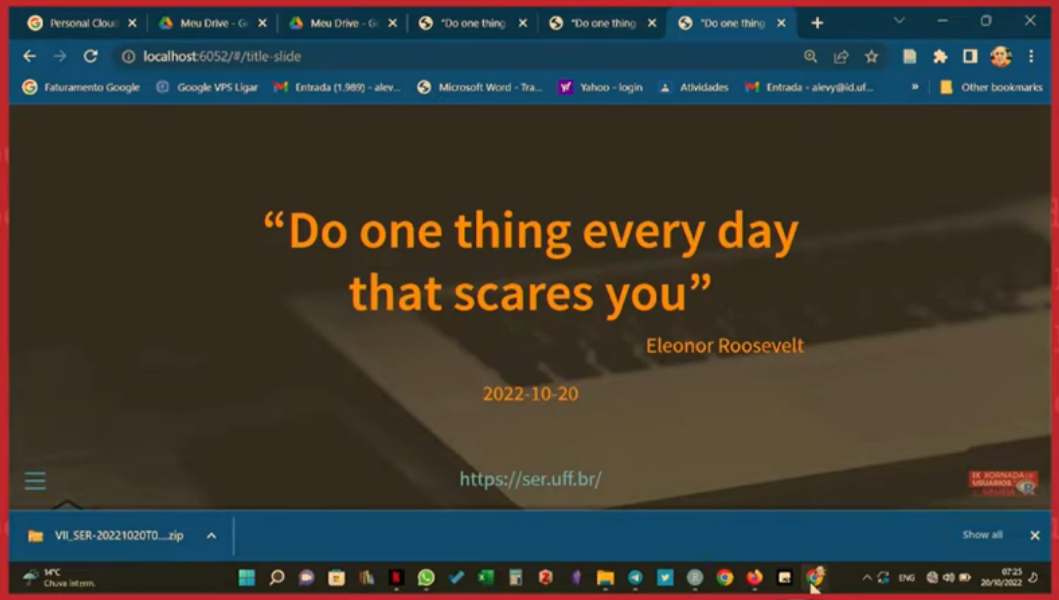
2 DÍAS

4 HORAS


Diabetes
Assesment model
Dispositivos vestibles
Sentence complexity
Rexións de referencia
Datos composicionales
Docencia
Regresión bivariada
Dependence
Cultura
Correlation
Emparellamento
Number of Groups
Selection methods
Invasión
Simulation
Distancia cultural
Toma de decisiones
Sampling
Cluster
Visualización gráfica
Linguagem R
Conservación biodiversidad
Multiple Regression Curves
Forecasting
Comparación de métodos de reparto
Estadística
Modelos predictivos de distribución de especies (SDM)
Smart city
Estimación non-paramétrica
Dynamic regression models
Gráficos de Control
Análisis de componentes principales no lineales del kernel
Linguagem R
Biosensores
Mockup
R language
Wireframe
Shiny
Mandala
Heterogeneidad espacial
Marco input-output
Processamento de texto
Soluciones puntuales
Clúster cultural
Desigualdad social
Multiple Survival Curves
R package chave
Programação
Innovación
Prototipo
Datos funcionales
Distance-correlation
Camino de Santiago
Evento SER
Condiciones de vida
COVID-19
Primeira lei de Zipf
Tomada de decisão
AEDA
Políticas públicas
Aup
Augas residuais
Juegos colacionales
Curvas paramétricas
Matching
Transformações
Independence
Paquete de R
Diagnóstico clínico
Climate change
Card-game
Time series
R-shiny
Soluciones de conjunto
Surplus production
Netrunner
RAD

Do one thing every day that scares you

Ariel Levy | Universidade Federal Fluminense (Brasil)



R y lme4: Uniendo fuerzas para la Estimación en Áreas Pequeñas

Naomi Diz-Rosales[1], María José Lombardía[1], Domingo Morales[2].

[1]Universidade da Coruña, CITIC, Spain.
[2]Universidad Miguel Hernández de Elche, IUCIO, Spain.

X Xornada de Usuarios de R en Galicia.
18/10/2023.

Introducción a la Estimación en Áreas Pequeñas

Estimación en Áreas Pequeñas (SAE)

Estimación en Áreas Pequeñas (SAE)

Rama multidisciplinar de la estadística que objetiva producir estimaciones precisas de variables de interés en dominios o áreas con pequeño o nulo tamaño muestral.

UNITED NATIONS

THE WORLD BANK

eurostat

United States® Census Bureau

Estimación en Áreas Pequeñas (SAE)

Demanda de datos más desglosados para el seguimiento de los Objetivos de Desarrollo Sostenible (ODS)

Uno de los principios básicos de la Agenda 2030 para el Desarrollo Sostenible es no dejar a nadie atrás. Por lo tanto, una parte de los procesos de revisión es el **suministro de datos de alta calidad, accesibles, fiables y desglosados** por ingresos, sexo, edad, raza, origen étnico, situación migratoria, discapacidad y ubicación geográfica, así como otras características pertinentes en los contextos nacionales. **El desglose de datos para los indicadores de los ODS garantiza el seguimiento de las desigualdades** (Kreutzmann y Chen, 2022).

Estimación en Áreas Pequeñas (SAE)

Rama multidisciplinar de la estadística que objetiva producir estimaciones precisas de variables de interés en dominios o áreas con pequeño o nulo tamaño muestral.

- Dominio: Grupo geográfico, sociodemográfico o de otra índole con pocas observaciones disponibles de un evento de interés.
- Aproximaciones metodológicas.
  - Basadas en diseño.
  - **Basadas en modelos.**
- Referencias clave: Fay y Herriot (1979); Battese et al. (1988); Prasad y Rao (1990); Jiang y Lahiri (2001); Rao y Molina (2015); y Morales et al. (2021).

Estimación en Áreas Pequeñas (SAE) y Modelos Mixtos

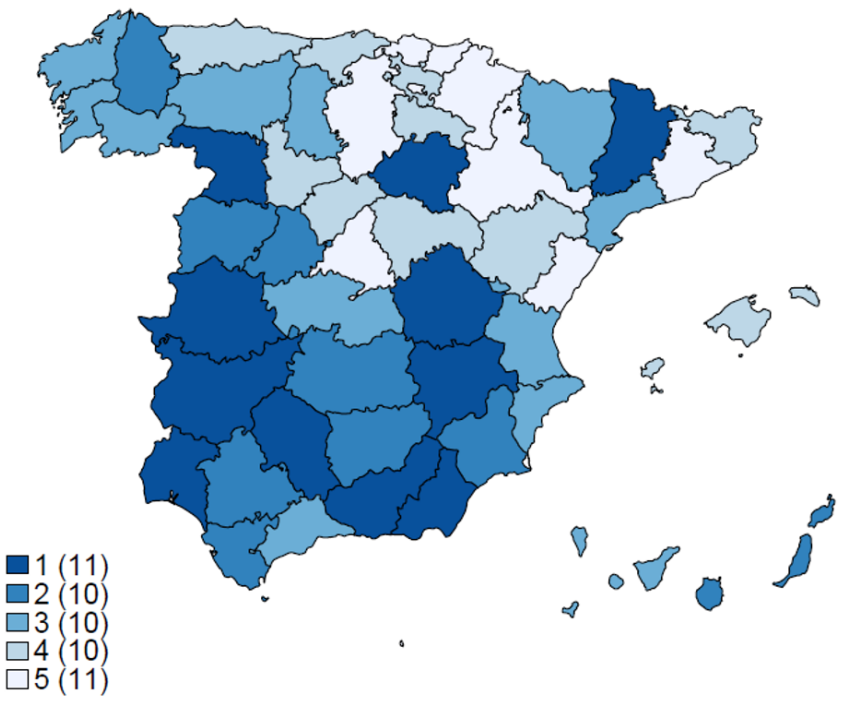
Rama multidisciplinar de la estadística que objetiva producir estimaciones precisas de variables de interés en dominios o áreas con pequeño o nulo tamaño muestral.

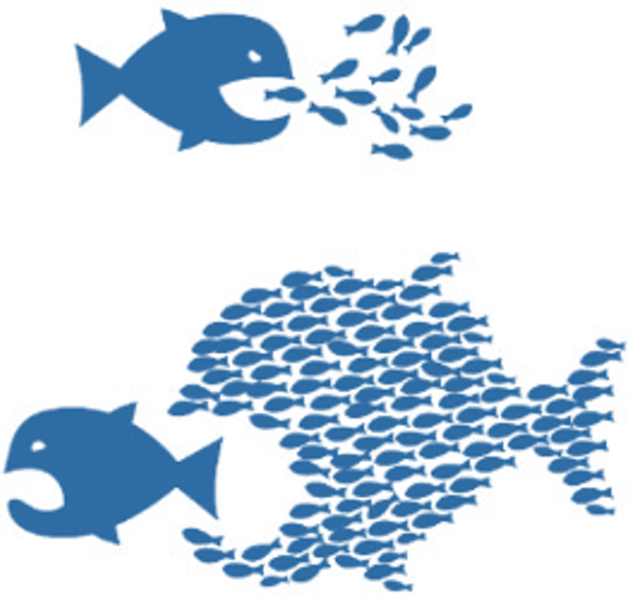
$$Y_{ij} = \beta_0 + \beta_1 x_{ij} + \sigma_u u_i + \varepsilon_{ij}$$

$$u_i \sim N(0, \sigma_u^2), \varepsilon_{ij} \sim N(0, \sigma^2)$$

Estimación en Áreas Pequeñas (SAE) y Modelos Mixtos

Rama multidisciplinar de la estadística que objetiva producir estimaciones precisas de variables de interés en dominios o áreas con pequeño o nulo tamaño muestral.





Nuestra línea de investigación

Nuestra línea de investigación se sustenta en **desarrollar** y **aplicar** modelos lineales generalizados mixtos o *Generalized linear mixed models* (GLMMs) con pendiente aleatoria a dos problemáticas de estimación en áreas pequeñas.

- Estimación de la proporción de pobreza por provincia y sexo en España.
- Estimación de la ocupación por COVID-19 en las unidades de cuidados intensivos (UCIs) por área sanitaria y día en España.

Introducción al paquete lme4

Paquete lme4

Creación y desarrollo: Douglas Bates, Martin Maechler, Ben Bolker, Steven Walker, Rune Haubo Bojesen Christensen, Henrik Singmann, Bin Dai, Fabian Scheipl, Gabor Grothendieck, Peter Green, John Fox, Alexander Bauer, Pavel N. Krivitsky. (Bates et al., 2023).

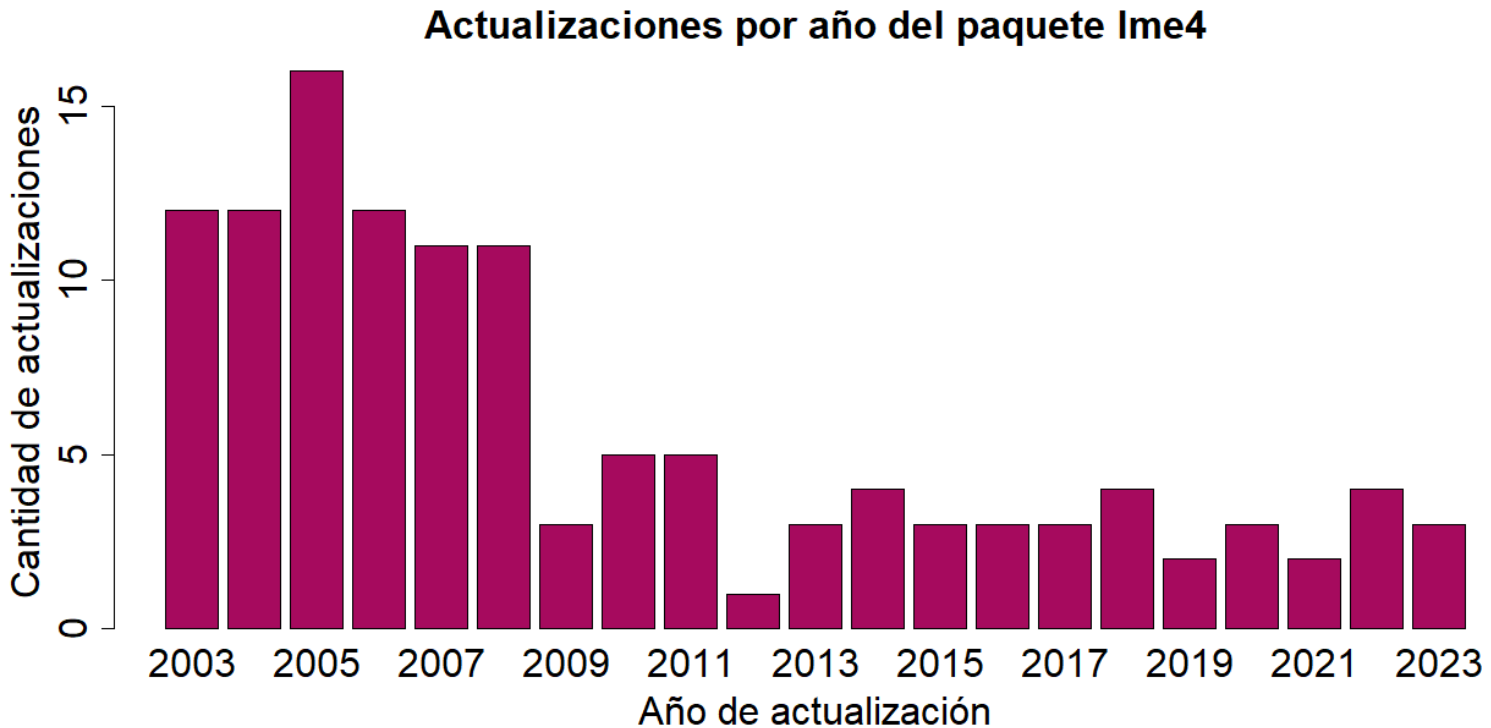
- **Objetivo**: Ajuste de modelos lineales mixtos (LMMs) y de modelos lineales generalizados mixtos (GLMMs).
- **Requisitos simples**: Actualmente solo es necesaria una versión de "R" >=3.5.0, y de "Matrix" >=1.2-1, así como disponer de "methods" y "stats".
- **Instalación sencilla**:
  - Directamente desde R:
    - `install.packages("lme4",dependencies=TRUE)`
  - Desde Github:
    - `library("devtools");`
      `install_github("lme4/lme4",dependencies=TRUE)`

Paquete lme4

- **Programación amigable**: lme4 emplea el mismo lenguaje de especificación de fórmulas que las funciones existentes en R para analizar modelos de regresión, como `lm()` y `glm()`.
- **Compatibilidad con paquetes clave**: Las estructuras generadas con lme4 se reconocen por otros paquetes de R, como ggplot2, cAIC4, boot, claves para diseño de gráficos, obtención de medidas de rendimiento, realizar simulaciones bootstrap u obtener intervalos de confianza.

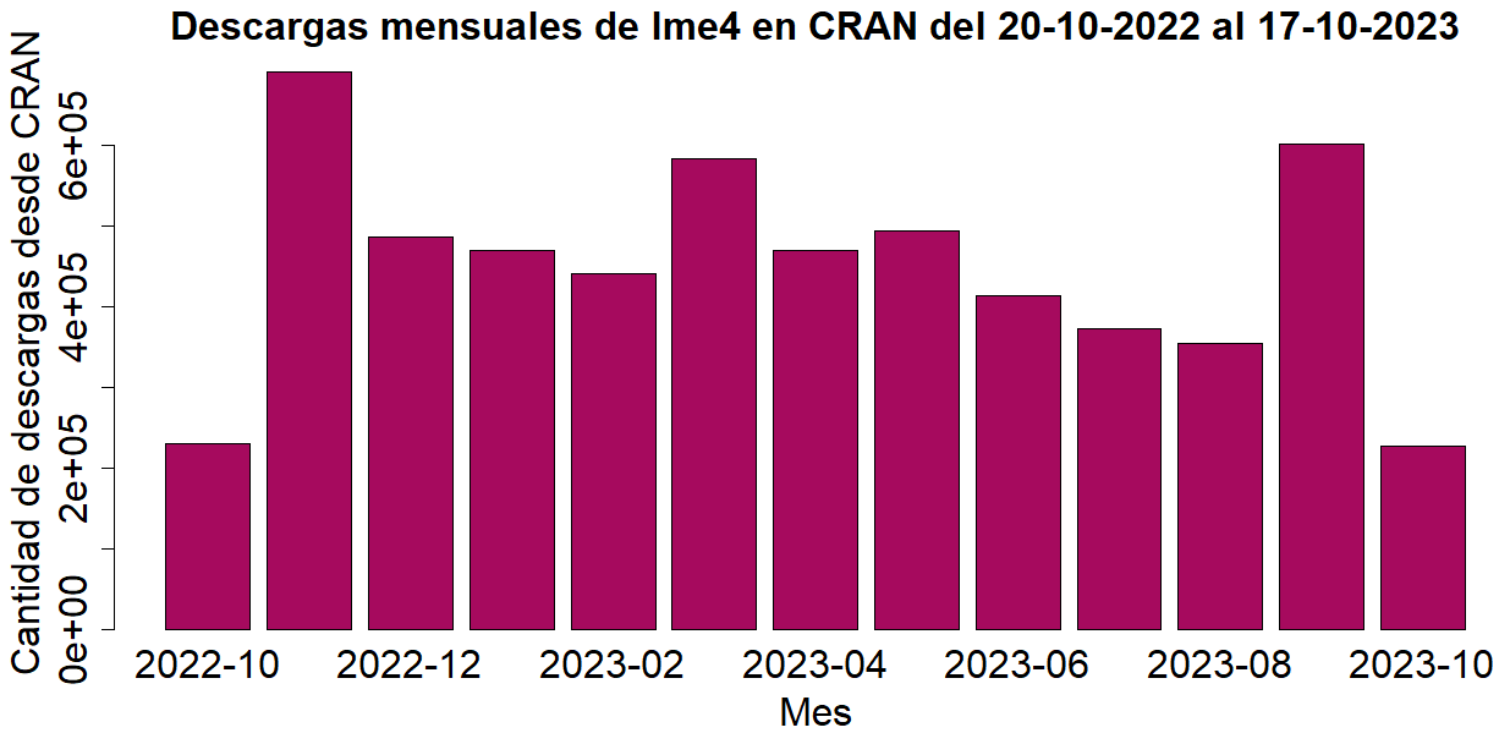
Paquete lme4

- **Mantenimiento y desarrollo continuo**: lme4 se actualiza entre 3 y 4 veces por año, resolviendo bugs detectados por la comunidad, optimizando el código e incorporando nuevas funcionalidades.



Paquete lme4

- **Comunidad activa y colaborativa**: lme4 goza de popularidad elevada, y una comunidad con una buena comunicación y soporte, en la que las personas creadoras y usuarias del paquete interactúan constantemente. `cranlogs::cran_downloads(package = "lme4", from = "2022-10-20", to = "2023-10-17")`



Uso del paquete lme4: library("lme4")

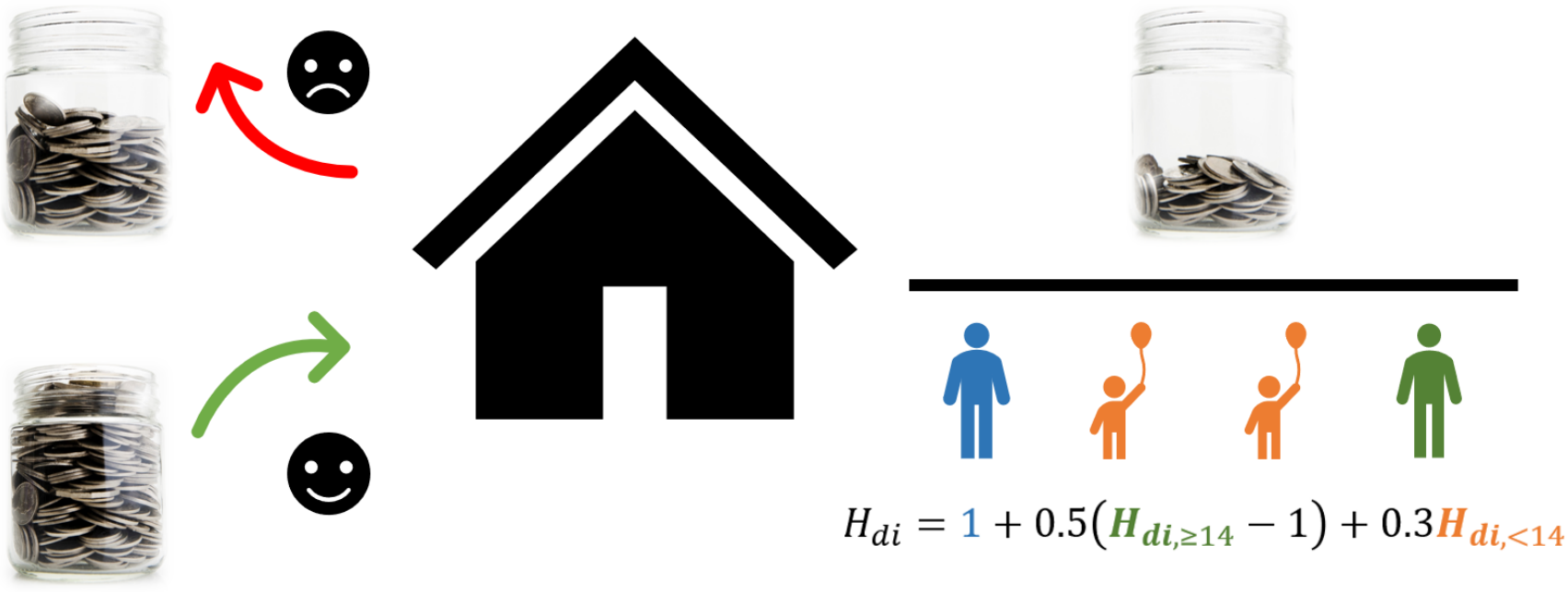
Nuestra línea de investigación

Nuestra línea de investigación se sustenta en **desarrollar** y **aplicar** modelos lineales generalizados mixtos o *Generalized linear mixed models* (GLMMs) con pendiente aleatoria a dos problemáticas de estimación en áreas pequeñas.

- **Estimación de la proporción de pobreza por provincia y sexo en España.**
- Estimación de la ocupación por COVID-19 en las unidades de cuidados intensivos (UCIs) por área sanitaria y día en España.

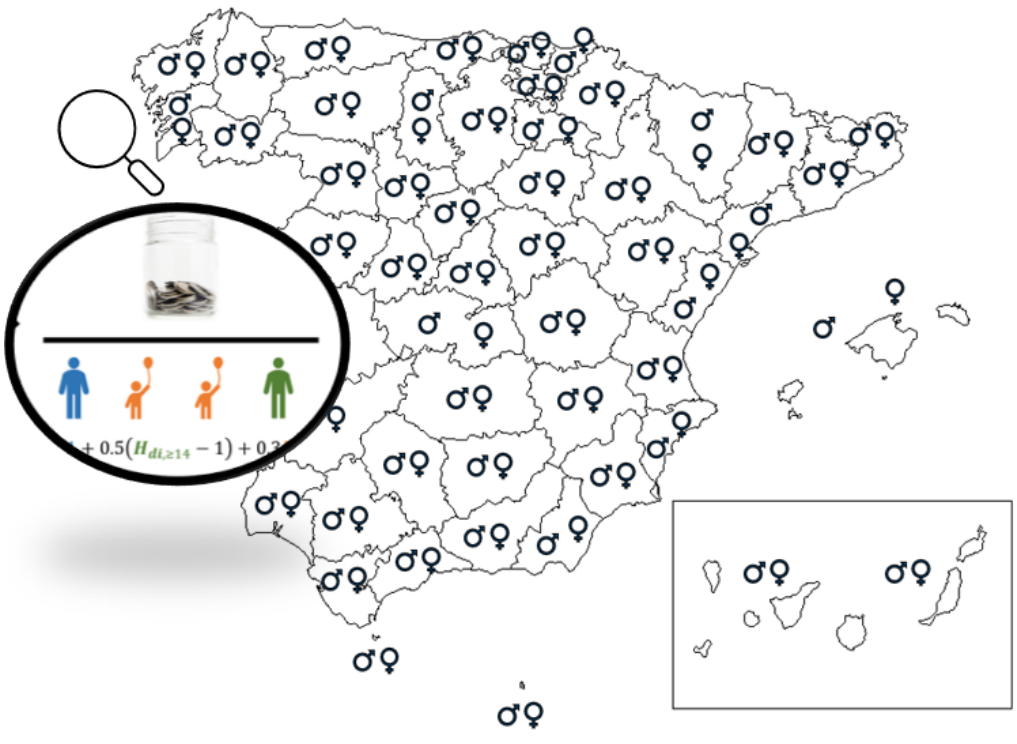
Proporción de pobreza

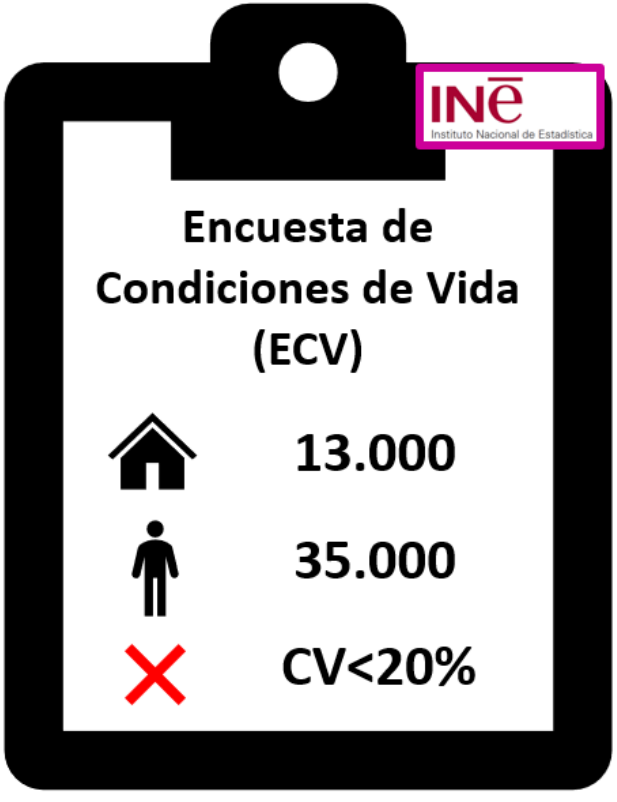
Proporción de las personas cuya **renta disponible equivalente<60% mediana** renta disponible equivalente nacional.



$$H_{di} = 1 + 0.5\left(H_{di,\geq 14} - 1\right) + 0.3H_{di,<14}$$

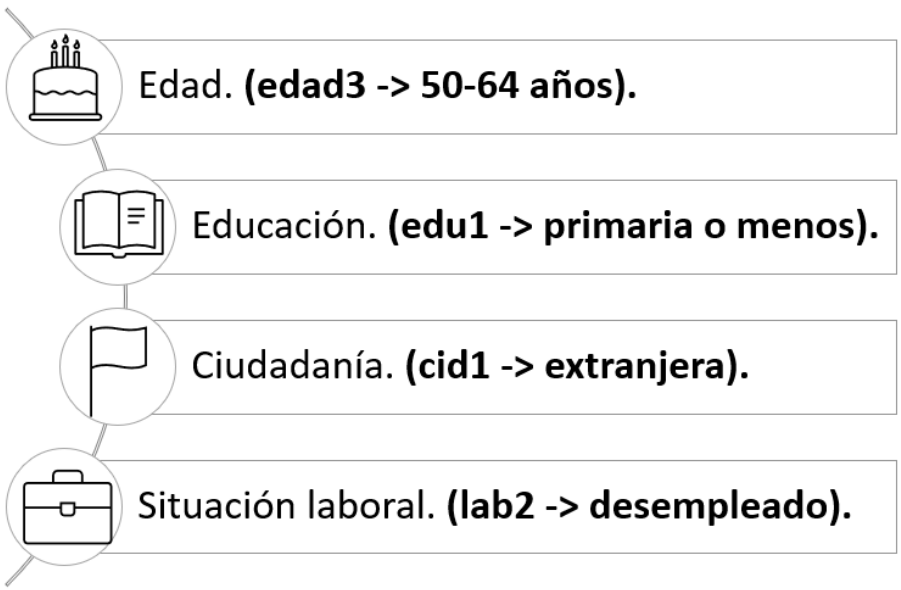
Estimación de la proporción de pobreza





Descripción de los datos

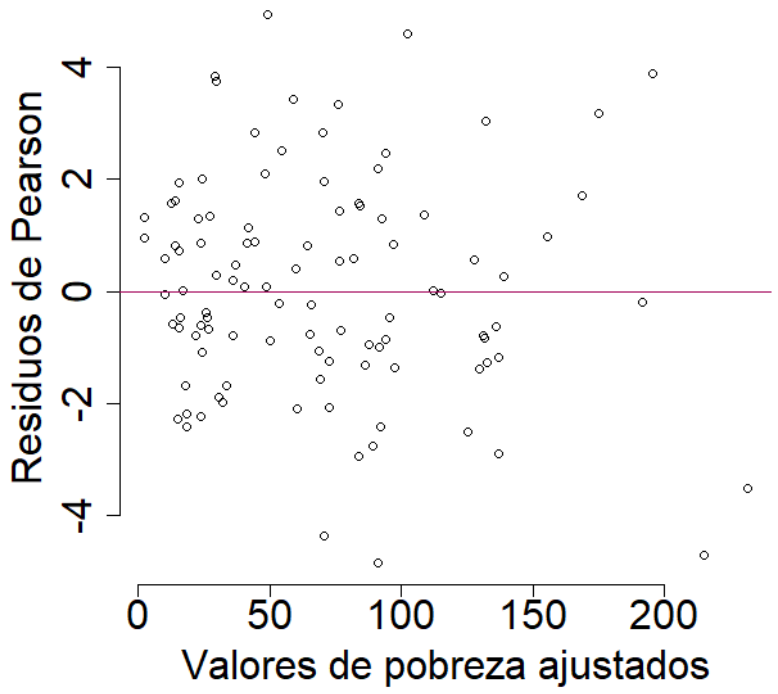
- **Variable objetivo: pobreza**. Procedente de los datos de la **Encuesta de Condiciones de Vida en España (ECV)** del 2008.
- **Variables auxiliares**: Procedente de los datos de la **Encuesta de Población Activa en España (EPA)** del 2008.

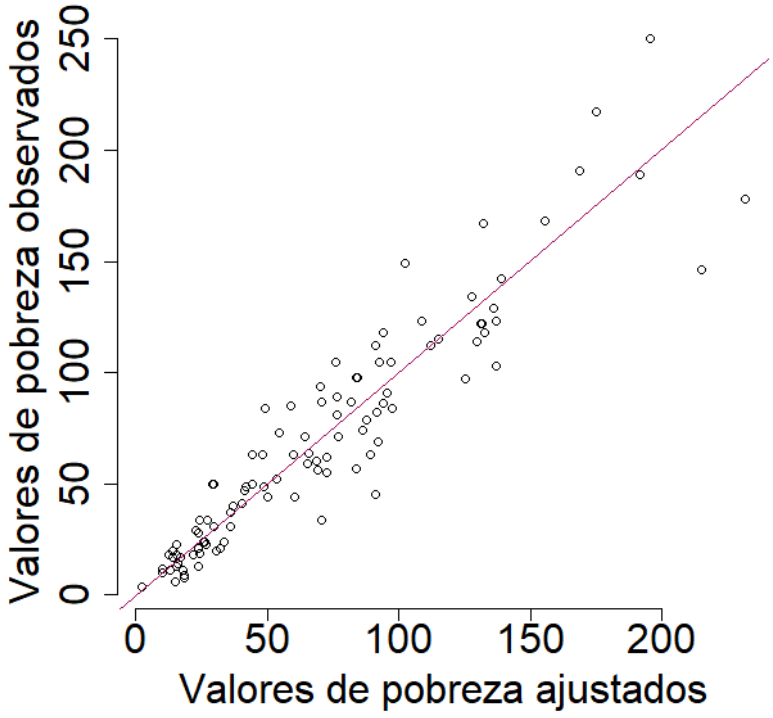


Estimación de la proporción de pobreza

Ajuste de un Modelo Lineal Generalizado (GLM).

```
M<-glm(pob~1+edad3+edu1+cid1+lab2,
offset=log(n),family="poisson",data=datos)
```





Estimación de la proporción de pobreza

Ajuste de un Modelo Lineal Generalizado Mixto (GLMM) con intercepto aleatorio.

MI<-glmer(pob~1+edad3+edu1+cid1+lab2 + **(1|factores agrupación)**,

offset=log(n),family="poisson",data=datos)

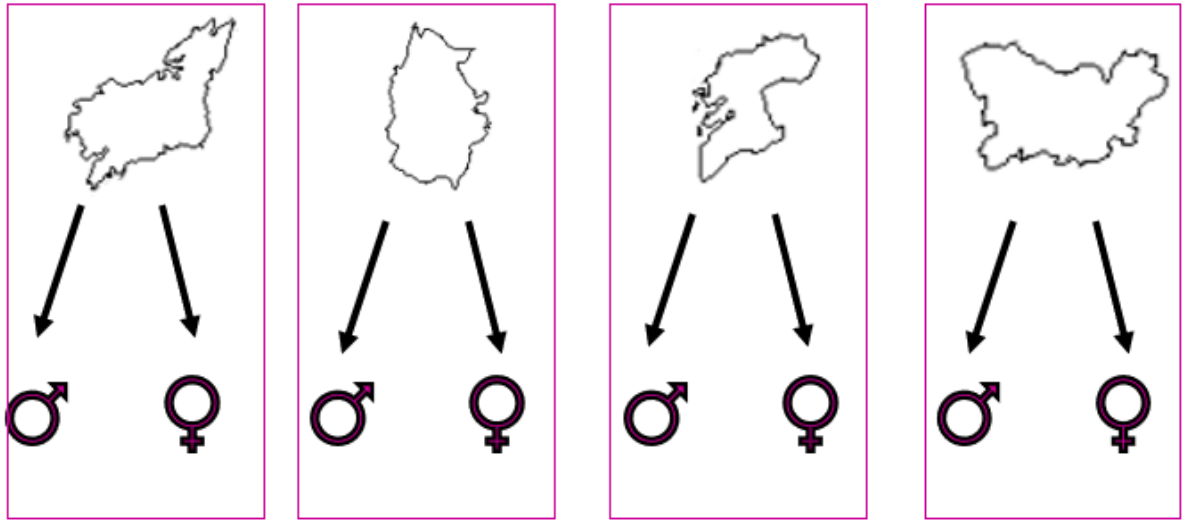
- ¿Cuál es el objetivo?: Estimar proporción de pobreza por **provincia y sexo**
- ¿Cuáles son los factores de agrupación?: **Provincia y sexo**
- ¿Los factores de agrupación están **anidados** o **cruzados**?

Estimación de la proporción de pobreza

Ajuste de un Modelo Lineal Generalizado Mixto (GLMM) con intercepto aleatorio.

Factores de agrupación **anidados**: El efecto de ser mujer u hombre en la proporción de pobreza estimada es diferente según la provincia. **Comparativa diferencia sexos dentro de la provincia**.

$$\mathbf{(1|provincia/sexo) = (1|provincia)+(1|provincia:sexo)}.$$



Estimación de la proporción de pobreza

Ajuste de un Modelo Lineal Generalizado Mixto (GLMM) con intercepto aleatorio.

Factores de agrupación **cruzados**: El efecto de ser mujer u hombre en la proporción de pobreza estimada es independiente de la provincia. **Comparativa diferencia sexos entre provincias**.

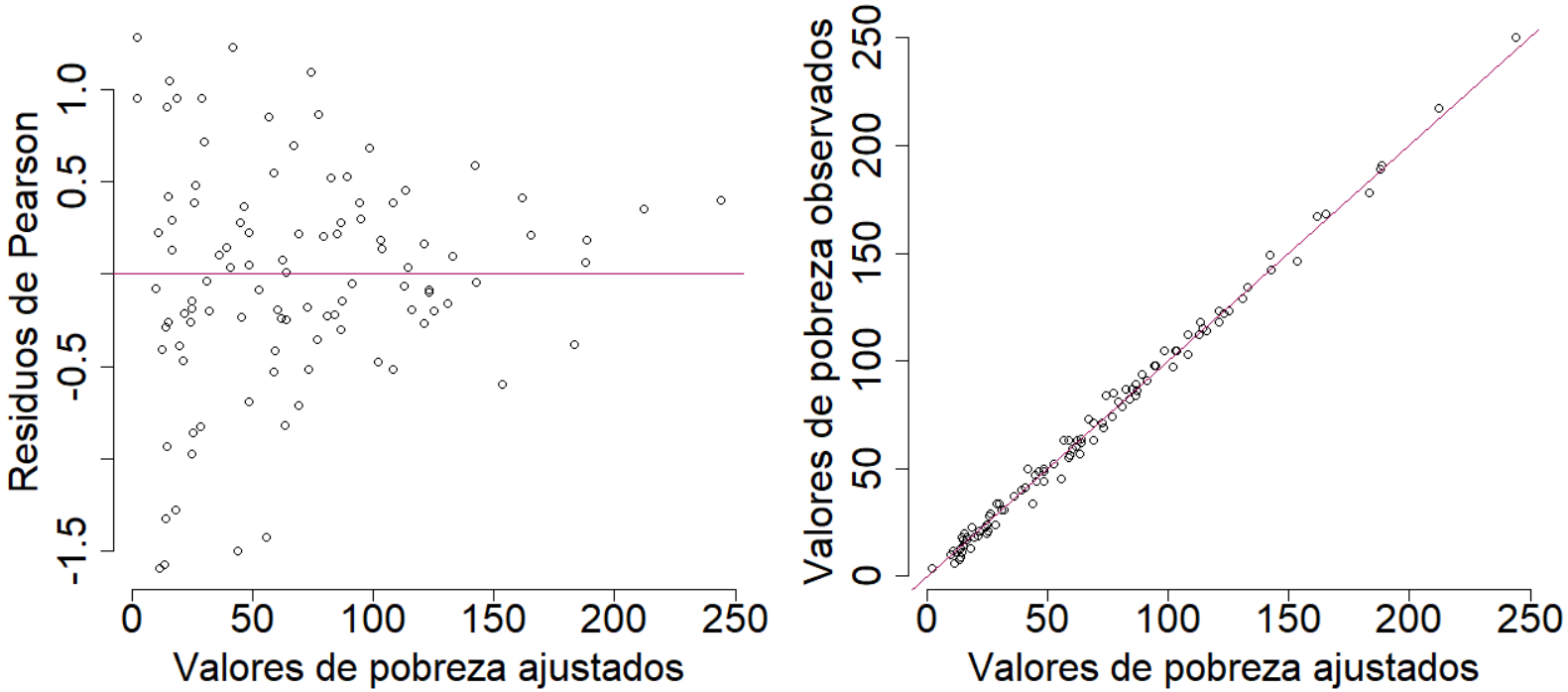
(1|provincia:sexo).

Estimación de la proporción de pobreza

Ajuste de un Modelo Lineal Generalizado Mixto (GLMM) con intercepto aleatorio.

Ml<-glmer(pob~1+edad3+edu1+cid1+lab2 + **(1|prov:sex)**,

offset=log(n),family="poisson",data=datos)

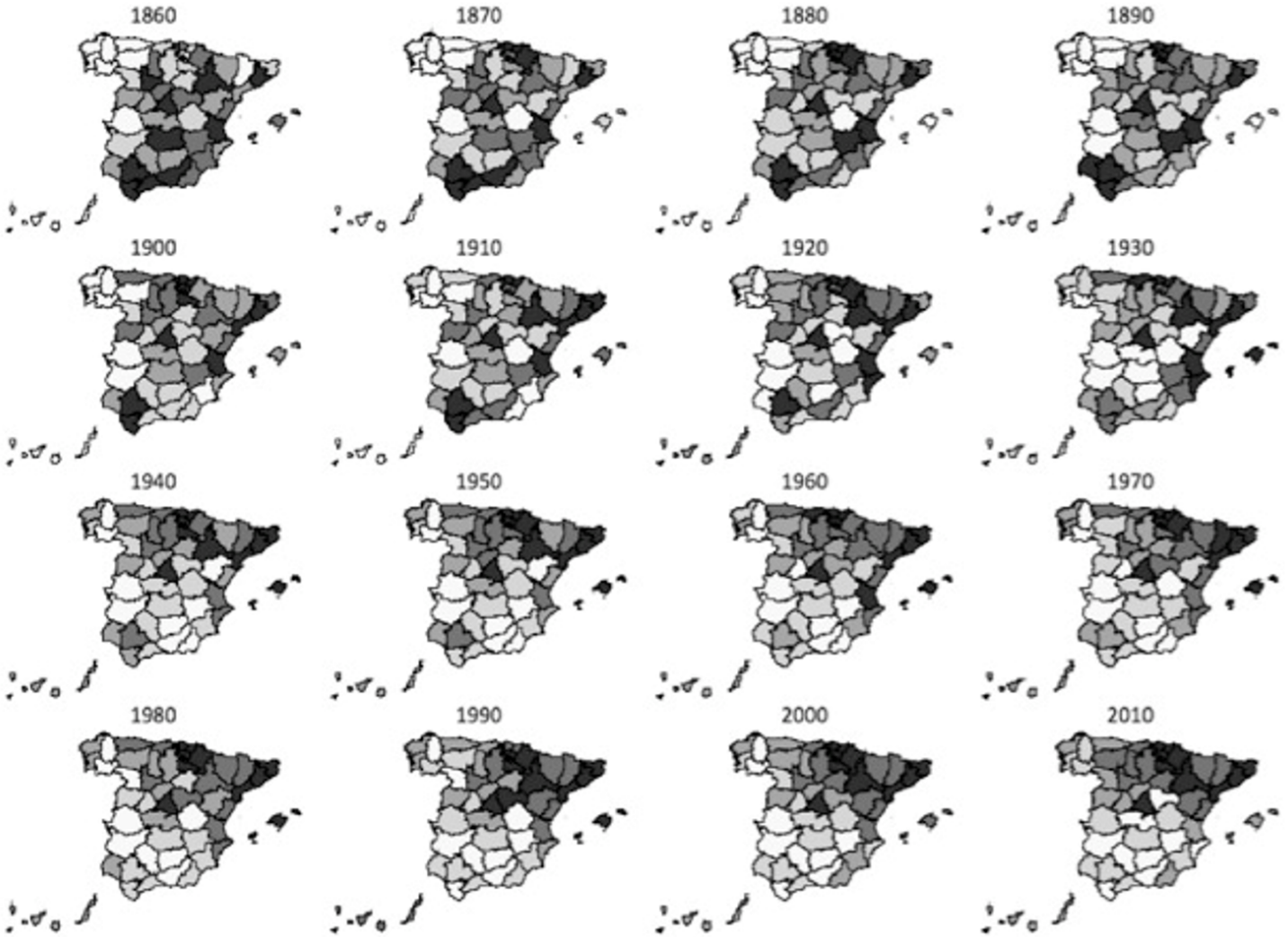


Estimación de la proporción de pobreza

¿Pero . . . es suficiente para modelar la heterogeneidad de las provincias?

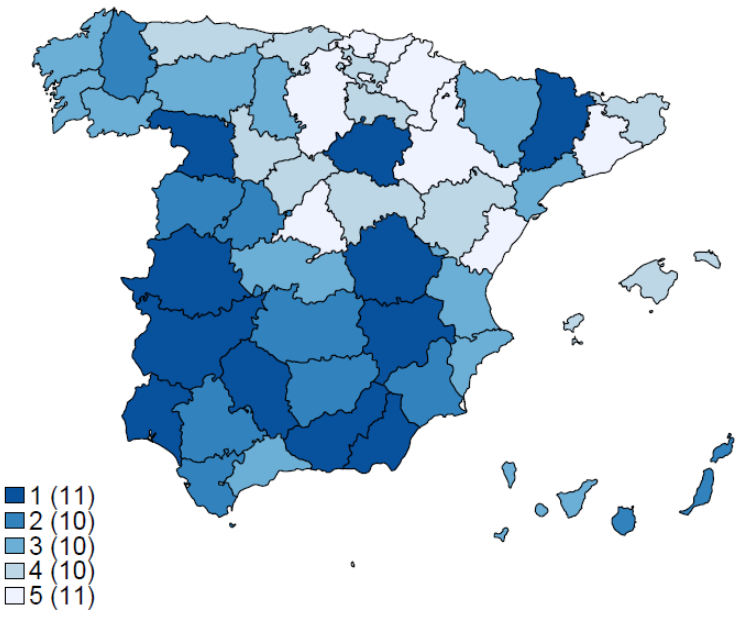
Estimación de la proporción de pobreza

Agrupación de las provincias por niveles de renta (Tirado et al., 2016)



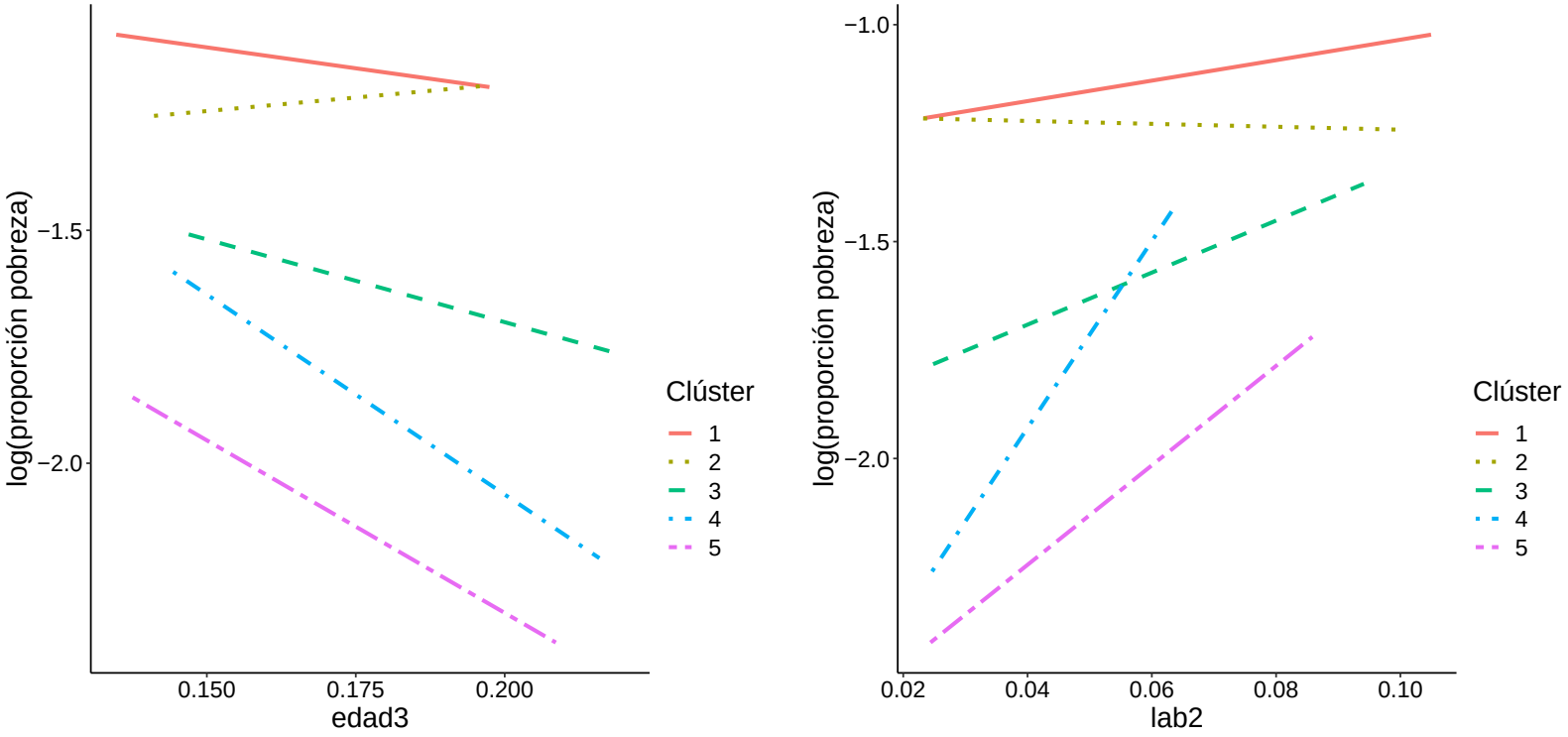
Estimación de la proporción de pobreza

Generación de la variable *grupo_ingreso*: Cinco clústers ($K = 5$), $\{k = 1, \ldots, K\}$, de acuerdo con el valor de la suma de la renta equivalente disponible para mujeres y hombres en cada provincia.



Estimación de la proporción de pobreza

Relación de la variable dependiente **Log-proporción-pobreza** con las variables auxiliares *edad*3 y *lab*2 por clúster k, $k = 1, \ldots, 5$, de la variable *grupo_ingreso*.

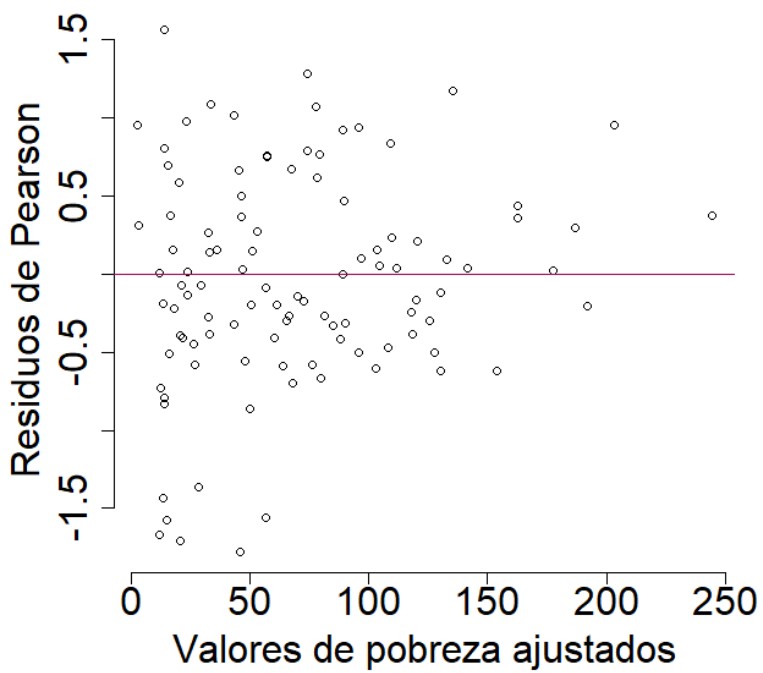


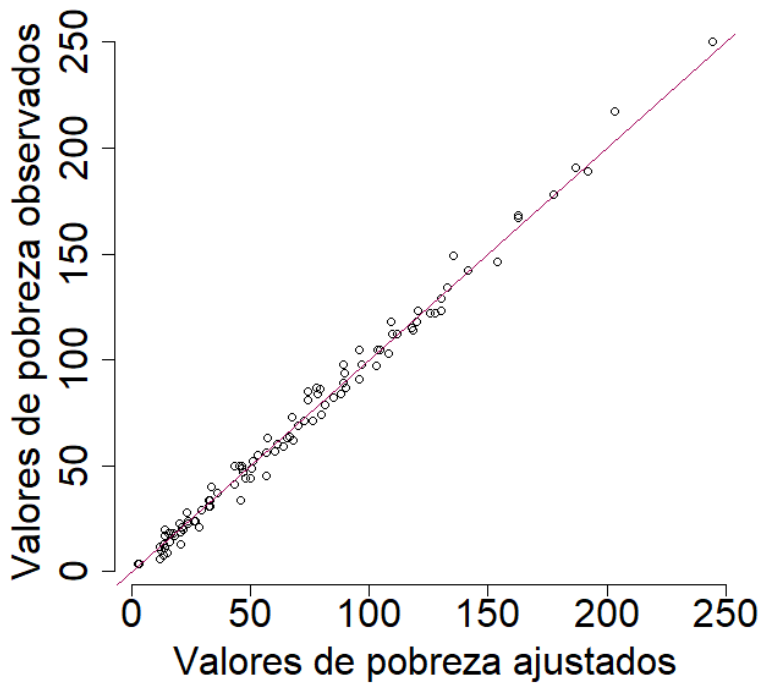
Estimación de la proporción de pobreza

Definimos, por primera vez en SAE, un modelo de área de Poisson con coeficientes de regresión aleatorios (**Modelo ARRCP**).

MS<-glmer(pob~1+edad3+edu1+cid1+lab2+(1|prov:sex)+

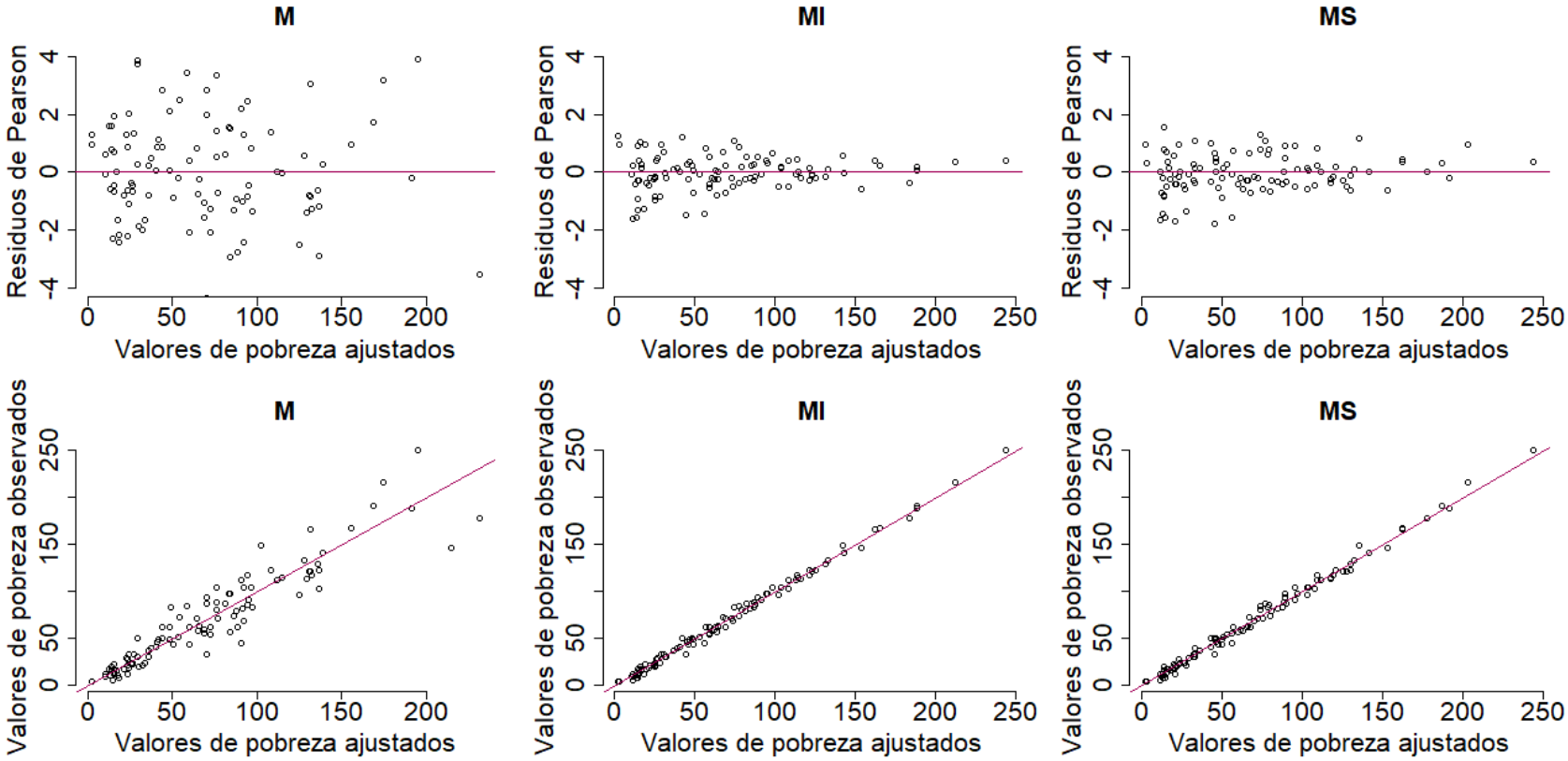
(0+edad3+lab2|grupo_ingreso),offset=log(n),family="poisson",data=datos)





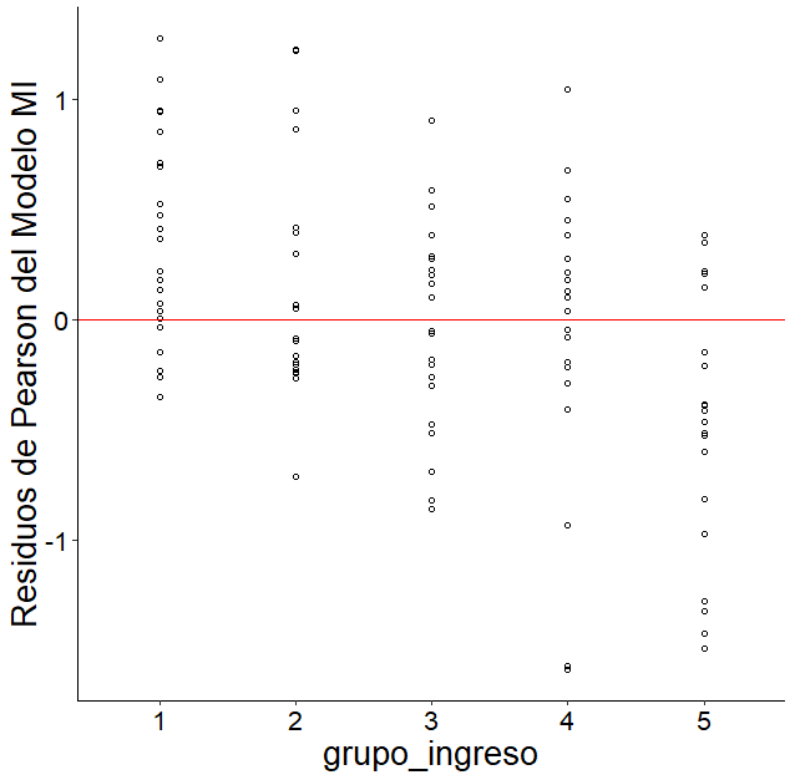
Estimación de la proporción de pobreza. Selección del modelo.

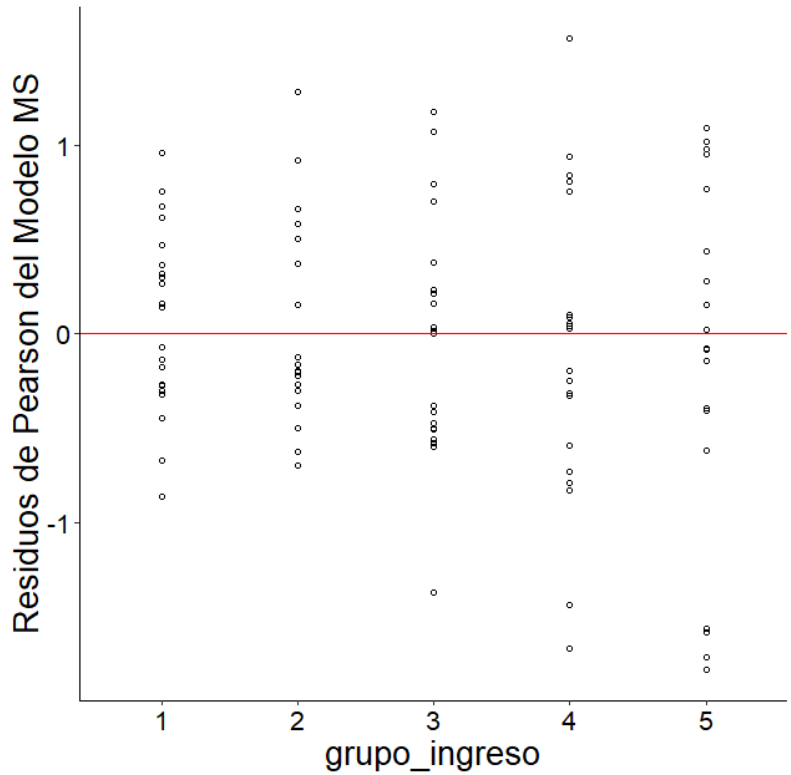
Comparativa bondad de ajuste de: **M** (GLM), **MI** (GLMM con intercepto aleatorio), **MS** (GLMM con intercepto y pendientes aleatorias).



Estimación de la proporción de pobreza. Selección del modelo.

Comparativa de la calidad de ajuste del modelo **MI** (GLMM con intercepto aleatorio) y **MS** (GLMM con intercepto y pendientes aleatorias) para cada uno de los clústers de la variable **grupo_ingreso**.





Estimación de la proporción de pobreza. Selección del modelo.

Obtención del AIC (Criterio de información de Akaike), métrica que evalúa la calidad del modelo con el objetivo de encontrar un equilibrio entre precisión y simplicidad.

- `AIC(M)=1005.927`

Obtención del cAIC (Criterio de información de Akaike condicionado), redefinición del AIC para incorporar la incerteza de la varianza de los efectos aleatorios. (Saefken et al., 2021).

```
install.packages("cAIC4",dependencies=TRUE); library("cAIC4")
```

- `set.seed(281095);`
  `cAIC(MI,method="conditionalBootstrap",B=500)=764.3284`
- `set.seed(281095);`
  `cAIC(MS,method="conditionalBootstrap",B=500)=743.1995`

Estimación de la proporción de pobreza. Intervalos de confianza.

Obtención de intervalos de confianza (IC).

- Método "Profile".

```
confint(MS,level=0.95,method="profile")
```

- Método "Wald".

```
confint(MS,level=0.95,method="Wald")
```

- Método "Bootstrapping".

```
confint(MS,level=0.95,method="boot",nsim=500)
```

  - IC percentil. `boot.type="perc"`
  - IC percentil básico. `boot.type="basic"`
  - IC percentil normal `boot.type="norm"`

Estimación de la proporción de pobreza. Intervalos de confianza.

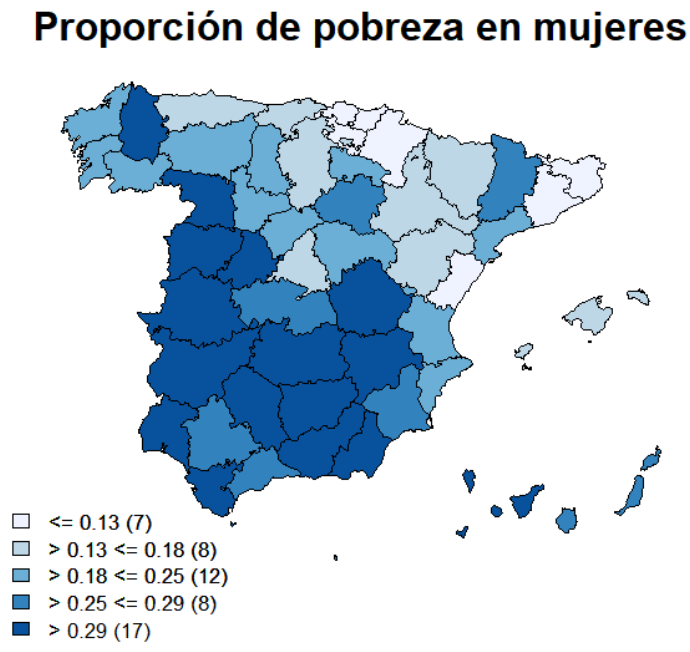
Ejemplo de obtención de Intervalos de Confianza (IC):
Intervalos de confianza bootstrap percentil básico.

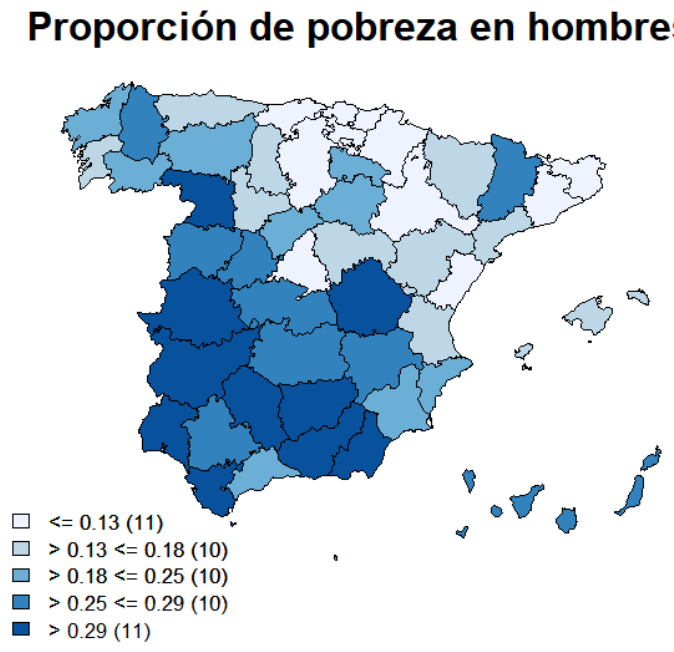
```
set.seed(281095)
confint(MS,level=0.95,method="boot",nsim=500,boot.type="basic")
```

```
                  2.5 %     97.5 %
.sig01        0.1129501  0.1899005
.sig02        0.8758710  3.9497347
.sig03       -2.7562803 -0.8863637
.sig04        0.4359008  4.9817440
(Intercept) -2.1224526 -0.6503372
edad3       -8.3148262 -0.1548832
edu1         0.2392169  1.8565559
cid1        -2.1217095 -0.4990385
lab2         3.8276708 10.4179124
```

Mapeo de la proporción de pobreza.

Representación geográfica de la distribución de la proporción de pobreza en España por provincia y sexo.





Publicación de la investigación en acceso abierto

Toda la metodología y resultados ilustrados en esta presentación están recogidos en un trabajo aceptado para publicación, en acceso abierto, en la revista *Journal of Survey Statistics and Methodology*.

Diz-Rosales, N., Lombardía, M.J., y Morales, D. (2023, aceptado para publicacion). Poverty mapping under area-level random regression coefficient Poisson models. *Journal of Survey Statistics and Methodology*. DOI: 10.1093/jssam/smad036.

Nuestra línea de investigación

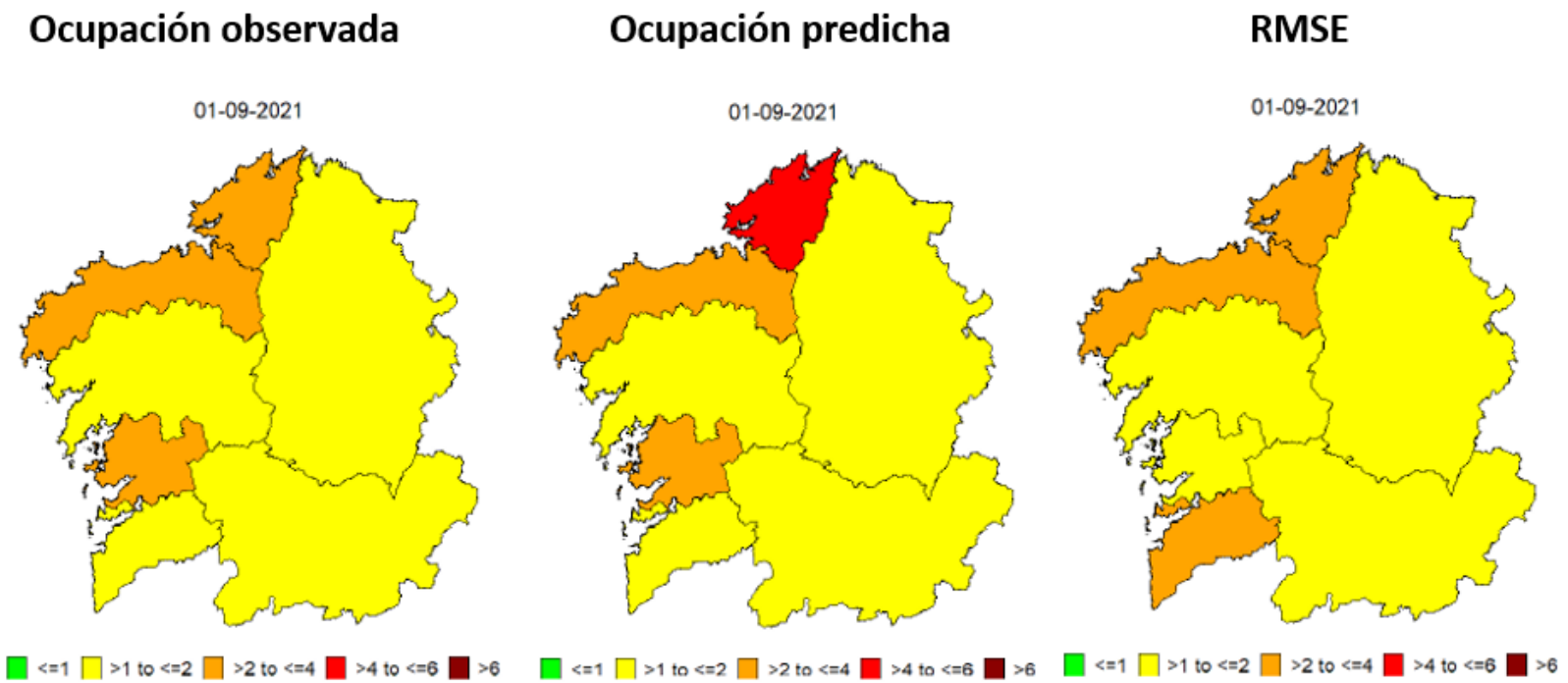
Nuestra línea de investigación se sustenta en **desarrollar** y **aplicar** modelos lineales generalizados mixtos o *Generalized linear mixed models* (GLMMs) con pendiente aleatoria a dos problemáticas de estimación en áreas pequeñas.

- Estimación de la proporción de pobreza por provincia y sexo en España.
- **Estimación de la ocupación por COVID-19 en las unidades de cuidados intensivos (UCIs) por área sanitaria y día en España (Diz-Rosales et al., 2023).**

Proporción de Unidades de Cuidados Intensivos (UCI) ocupadas por pacientes COVID-19

Proporción de Unidades de Cuidados Intensivos (UCI) ocupadas por pacientes COVID-19

Mapeo del riesgo alcanzado por la ocupación de las UCI por cada 100.000 habitantes a causa del COVID-19. Para cada día se representan los valores observados, los predichos a siete días por el modelo y el RMSE asociado a cada predicción.



Caja de recursos para usar lme4

1. Detallado manual explicativo con ejemplos de uso (Bates et al., 2023).
2. Numerosos tutoriales. Recomendados los del desarrollador del paquete lme4 Ben Bolker (Bolker, 2019; Bolker et al., 2023).
3. Comunidad activa y resolutiva. Consulta de la GitHub de lme4.

Implementaciones necesarias en lme4

- El paquete lme4 no permite de modo directo indicar estructuras de covarianza, lo que puede ser necesario en caso de correlación temporal, espacial o de otra índole en los efectos aleatorios. Actualmente, se ha desarrollado, y está en continúa actualización, "lme4ord" para adoptar esas funcionalidades.
- Es necesario incrementar el número de familias disponibles. Por ejemplo, la binomial negativa no se puede especificar. Actualmente se ha desarrollado y está en continuación actualización "glmmTMB".

Conclusiones

Conclusiones

- La Estimación en Áreas Pequeñas es un campo de la estadística en continuo desarrollo por sus grandes aplicaciones, siendo potenciada por organizaciones como The World Bank, Naciones Unidas, Eurostat o el Census Bureau en el marco de los ODS.
- Bajo el principio de que la unión hace la fuerza, los modelos mixtos permiten incrementar la precisión en las estimaciones en áreas pequeñas.
- En R, lme4 es el paquete por excelencia para modelos mixtos, en base a su flexibilidad, facilidad de uso, pocos requisitos computacionales, compatibilidad con un amplio abanico de paquetes clave y, especialmente, por su continuo soporte, desarrollo y comunidad activa.

- Ejemplificamos el uso de lme4 para la estimación de la proporción de pobreza, aprendiendo a definir modelos con intercepto y pendiente aleatoria, a realizar el diagnosis del modelo y a obtener métricas como el cAIC y los intervalos de confianza.
- Ejemplificamos el uso de lme4 para la estimación y predicción de la ocupación de las Unidades de Cuidados Intensivos (UCI) por pacientes con COVID-19.
- El equipo de desarrollo del paquete está trabajando para poder flexibilizar la especificación de las estructuras de covarianza en la definición de los efectos aleatorios y ampliar las familias disponibles, que a día de hoy, constituyen sus principales limitaciones.

Agradecimientos

Agradecimientos

Este trabajo es parte de la ayuda PRE2021-100857, financiada por MCIN/AEI/10.13039/501100011033 y por el FSE+.

MINISTERIO
DE CIENCIA
E INNOVACIÓN

Cofinanciado por
la Unión Europea

AGENCIA
ESTATAL DE
INVESTIGACIÓN

Esta investigación también ha contado con el apoyo de dos subvenciones nacionales, PID2022-136878NB-I00 y PID2020-113578RB-I00, la financiación de la Generalitat Valenciana vía Prometeo/2021/063, y la financiación de la Xunta de Galicia (Grupos de Referencia Competitiva ED431C/2020/14 y Centro Singular de Investigación de Galicia ED431G/2019/01), todas ellas a través de ERDF.

Referencias

- Bates, D., Maechler, M., Bolker, B., Walker, S., Christensen, R., Singmann, H., Dai, B., Scheipl,F., Grothendieck, G., y Green, P. (2023). Package lme4, version 1.1-34. Disponible en: https://cran.r-project.org/web/packages/lme4/lme4.pdf. (Acceso: 16/10/2023).
- Battese, G.E., Harter, R.M., y Fuller, W.A. (1988). An error-components model for prediction of county crop areas using survey and satellite data. *Journal of the American Statistical Association*, **83**, 28–36.
- Bolker, B. (2019). RPubs bbolker. RPubs by RStudio. Disponible en: https://rpubs.com/bbolker. (Acceso: 16/10/2023).
- Bolker, B., y otros. (2023). GLMM FAQ. Disponible en: http://bbolker.github.io/mixedmodels-misc/glmmFAQ.html. (Acceso: 16/10/2023).

- Diz-Rosales, N., Lombardía, M.J., y Morales, D. (2023). Modelling the COVID-19 ICU occupancy with area-level random regression coefficient Poisson models. *En XIX Conferencia Española y VIII Encuentro Iberoamericano de Biometría (CEB-EIB 2023).* Vigo, España. Disponible en: http://cebeib2023.webs2.uvigo.es/wp-content/uploads/2023/06/LdR-CEB-EIB_2023.pdf
- Diz-Rosales, N., Lombardía, M.J., y Morales, D. (2023, aceptado para publicación). Poverty mapping under area-level random regression coefficient Poisson models. Journal of Survey Statistics and Methodology. DOI: 10.1093/jssam/smad036.
- Fay, R. E., y Herriot, R. A. (1979). Estimates of income for small places: an application of James-Stein procedures to census data. *Journal of the American Statistical Association*, **74**, 269–277.

- INE, 2023. Risk of poverty: At-risk-of-poverty rate, by Autonomous Community. https://www.ine.es/jaxiT3/Tabla.htm?t=9963&L=1. (Acceso: 30/09/2023).
- Jiang, J., y Lahiri, P. (2001). Empirical best prediction for small area inference with binary data. *Annals of the Institute of Statistical Mathematics*, **53**, 217–243.
- Kreutzmann, A.K., y Chen, H. (2022). Why is SAE important for SDG data disaggregation.United Nations Division Statistics. Disponible en:https://unstats.un.org/wiki/display/SAE4SDG/Why+is+SAE+important+for+SDG+data+disaggregation. (Acceso: 15/10/2023).
- Morales, D., Esteban, M.D., P´erez, A., y Hobza, T. (2021). *A course on small area estimation and mixed models. Methods, theory and applications in R*. Springer, Switzerland.

- Prasad, N.G.N., y Rao, J.N.K. (1990). The estimation of the mean squared error of small area estimators. *Journal of the American Statistical Association*, *85*, 163–171.
- Rao, J.N.K., Molina, I.(2015). *Small area estimation. (2nd ed.)*. Wiley, Hoboken.
- Säfken, B., Rügamer, D., Kneib, T., y Greven, S. (2021). Conditional Model Selection in Mixed-Effects Models with cAIC4. *Journal of Statistical Software*, **99**(8), 1–30.
- Tirado, D. A., Díez-Minguela, A., y Martínez-Galarraga, J. (2016). Regional inequality and economic development in spain, 1860–2010. *Journal of Historical Geography*, **54**, 87–98.

Muchas gracias por vuestra atención

R y lme4: Uniendo fuerzas para la Estimación en Áreas Pequeñas

Naomi Diz-Rosales[1], María José Lombardía[1], Domingo Morales[2].

[1]Universidade da Coruña, CITIC, Spain.
[2]Universidad Miguel Hernández de Elche, IUCIO, Spain.

X Xornada de Usuarios de R en Galicia.
18/10/2023.