Applied Multilevel Regression Modeling

Day 10: Spatial Multilevel

Constantin Manuel Bosancianu

WZB Berlin Social Science Center
Institutions and Political Inequality
bosancianu@icloud.com

August 9, 2019

The “standard” setup

$$
\begin{cases}
Y_{ij} = \beta_{0j} + \beta_{1j} * X1_{ij} + \cdots + \epsilon_{ij}, \; with\, \epsilon_{ij} \sim \mathcal{N}(0, \sigma_{\epsilon}^2) \\
\beta_{0j} = \gamma_{00} + \gamma_{01} * Z_j + \cdots + \upsilon_{0j} \\
\beta_{1j} = \gamma_{10} + \gamma_{11} * Z_j + \cdots + \upsilon_{1j} \\
\vdots \quad \vdots \quad \vdots \quad \vdots \quad \vdots \quad \vdots
\end{cases} \tag{1}
$$

Errors at all levels were distributed $\mathcal{N}$ with 0 mean and constant variance.

Implication: After controlling for group effects, L1 observations are independent of each other! *Geography* as a “space”, not a “place”...

A more realistic assumption

There is spatial association between random effects at both levels of the hierarchy.

Ignoring this leads us to believe we have a larger n than in reality.

$cor(\upsilon_{0k}, \upsilon_{0l}) = f(d)$ for all l units proximate to k, with $f(d)$ a function of distance between the units.

A similar assumption can apply to proximate ϵ_{ij}, irrespective of whether they belong to same j group or not.

Multilevel spatial analysis

CAR models (1)

To incorporate the effects of this spatial proximity, we need to include in the model a matrix of spatial weights.[1]

$$E(v_j|v_{-j}) = \frac{1}{w_{j+}} \sum_{k \sim j} v_k \tag{2}$$

We get a $J \times J$ matrix where element $w_{jk} = 1$ if j and k are neighbors ($j \sim k$).

$$LCAR: E(v_j|v_{-j}) = \frac{\lambda}{1 - \lambda + \lambda w_{j+}} \sum_{k \sim j} v_k \tag{3}$$

[1] This is incorporated either as a set of spatially weighted predictors (spatial lag model), or as spatially weighted error terms (spatial error model).

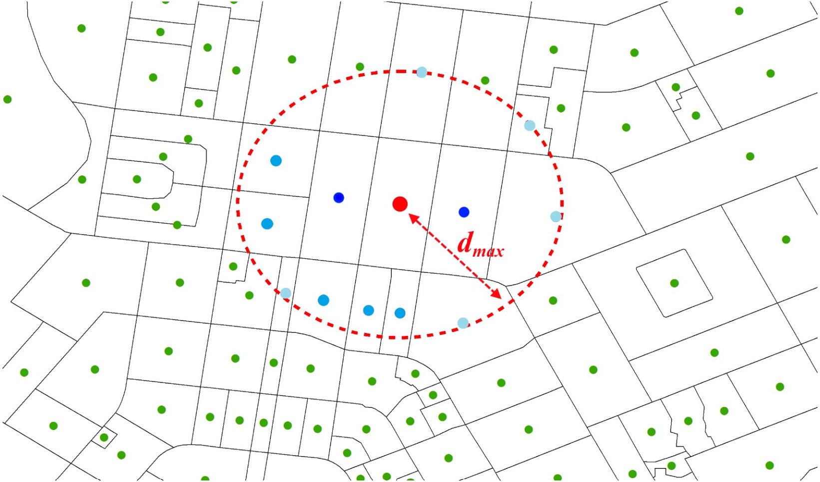
CAR models (2)

At L1, the procedure is the same, though we don't apply the "neighboring" principle anymore.

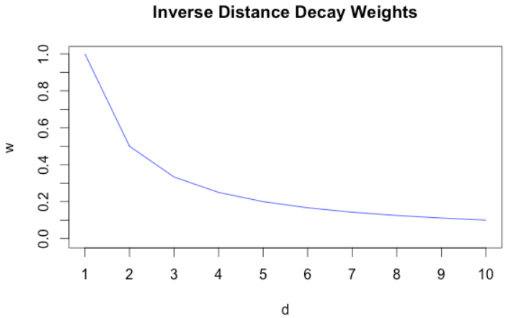
A spatial weights matrix is also incorporated, this time based on the distance between L1 units.

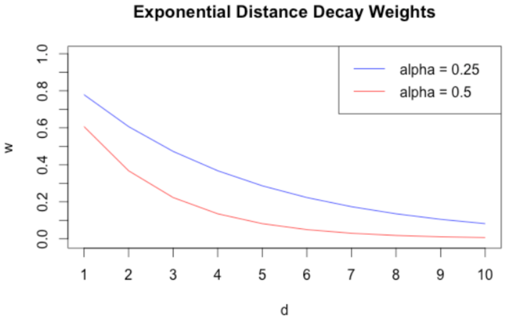
A distance decay matrix is computed, quantifying the extent of influence between units based on the distance.

Decay matrix



Decay matrix





Representations obtained from AURIN (Australian Urban Research Infrastructure Newtwork)

Interpretation

Because of allowing for spatial influence, we now have *direct* and *indirect* effects.

Spatial spillovers: impact passing through neighboring units, and then back to the unit itself.

These operate both for L1 and L2 units. *Indirect effects*: increases in the unit's outcome produced by increases in the predictor in surrounding units, averaged across all units.

Bayesian estimation

Bayes' Theorem

A simple device to invert conditional probabilities.

$$p(A|B) = \frac{p(B|A) * p(A)}{p(B)} \tag{4}$$

Immense possibilities emerge if we replace B with *data*, and A with an unknown parameter, θ.

$$\underbrace{p(\theta|data)}_{\text{posterior}} = \frac{\overbrace{p(data|\theta)}^{\text{likelihood}} \quad * \quad \overbrace{p(\theta)}^{\text{prior}}}{\underbrace{p(data)}_{=1}} \tag{5}$$

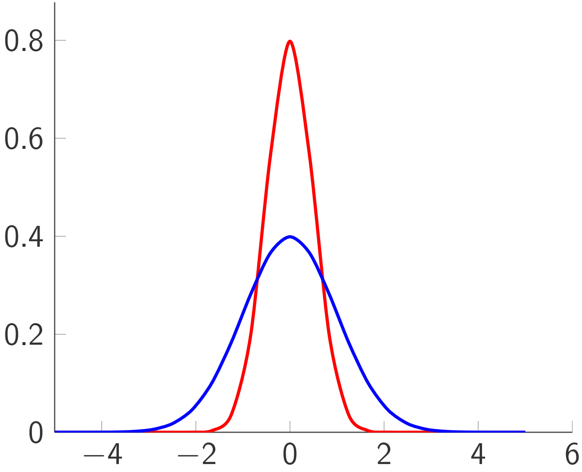
Prior distribution

We're making the switch from probabilities understood in the Frequentist sense to Bayesian probabilities, which denote our degree of (un)certainty.

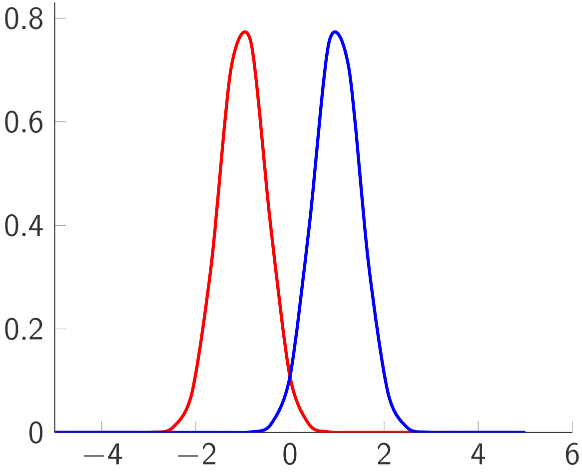
The jump is not really that large, if you consider the everyday meaning of "probable" or "likely".

A prior simply expresses our degree of uncertainty about the value of a parameter.

Different degree of confidence...



Different parameters

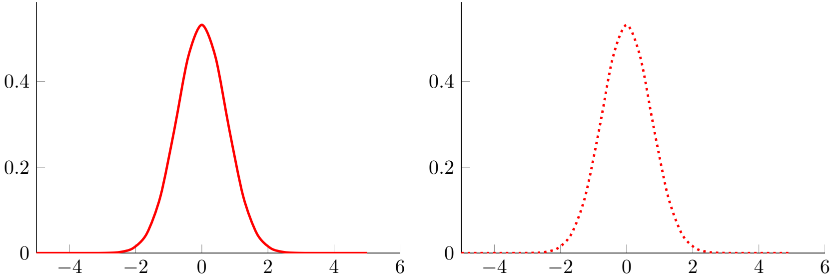


Monte Carlo methods

The solution in this case is to get a rough “image” of this distribution, and to use this “copy” for the purpose of computing whatever parameters we’re interested in.

Monte Carlo methods are algorithms through which this “image” can be obtained.

Role of the "image"



It's like the cinema-ripped fuzzy copy of a movie: not crystal-clear, but enough to be able to tell what the action is all about

The “image”

This is called a Markov chain because it has a short memory: the next point in the chain is only determined by the current point.

The path travelled by the algorithm represents a pretty good image of the distribution we’re interested in.[2]

[2] As long as we exclude the first steps, which award the starting point too great of an influence compared to the others.

Thank you for joining the course!

Let's keep in touch:

bosancianu@icloud.com