Distributional Modelling in R

Case Studies I – Continuous Distributions

Thomas Kneib, Nikolaus Umlauf

https://nikum.org/dmr.html

Normal Location-Scale Model

- The normal location-scale model

$$y_i | \boldsymbol{x}_i \sim \mathrm{N}(\mu(\boldsymbol{x}_i), \sigma^2(\boldsymbol{x}_i))$$

  with

$$\begin{aligned} \mu(\boldsymbol{x}_i) &= \eta_\mu(\boldsymbol{x}_i) \\ \sigma^2(\boldsymbol{x}_i) &= \exp(\eta_{\sigma^2}(\boldsymbol{x}_i)) \end{aligned}$$

  is the most commonly known distributional regression model for continuous responses.

- However, there is a huge variety of distributions to pick from!

Example: Income Inequality

- Utilise information from the German Socio-Economic Panel to study real gross annual personal labour income in Germany for the years 2001 to 2010.
- Specific focus on changes in spatial differences in income inequality.
- Response: income of males in full time employment in the age range 20–60.
- Information available on $7,216$ individuals with a total of $n = 40,965$ observations.
- Potential response distributions:
  - Log-normal $\text{LN}(\mu, \sigma^2)$.
  - Gamma $\text{Ga}(\mu, \sigma)$.
  - Inverse Gaussian $\text{IG}(\mu, \sigma^2)$.
  - Dagum $\text{Da}(a, b, c)$.

  with covariate effects on potentially all distributional parameters.

Covariates

- `educ`: Educational level measured as a binary indicator for completed higher education (according to the UNESCO International Standard Classification of Education 1997).
- `age`: age in years.
- `lmexp`: previous labour market experience in years.
- `t`: calendar time.
- `r`: area of residence in terms of geographical district (*Raumordnungsregion*).
- `east`: indicator in effect coding for districts belonging to the eastern part of Germany.

Model Specification

- Hierarchical predictor structure:

$$\eta_i = \beta_0 + \texttt{educ}_i \beta_1 + s_1(\texttt{age}_i) + \texttt{educ}_i s_2(\texttt{age}_i) + s_3(\texttt{lmexp}_i) + s_{spat}(r_i) + s_{time}(\texttt{t}_i)$$

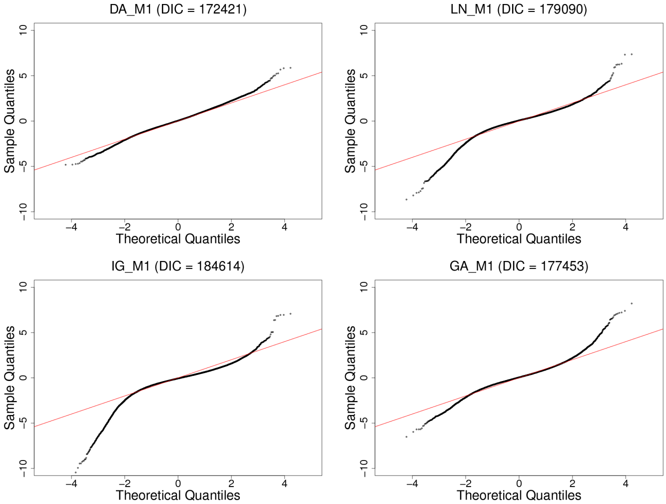
where the spatial effects is decomposed as

$$s_{spat}(\texttt{r}) = \texttt{east}_r \gamma_1 + s_{str}(\texttt{r}) + s_{unstr}(\texttt{r})$$

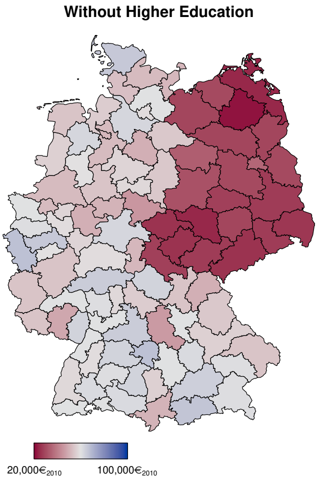
Model Comparison by DIC and Scoring Rules

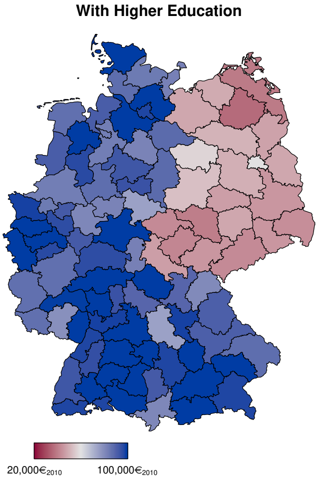
Distribution	DIC	Quadratic	Logarithmic	Spherical	CRPS
LN	179,090	0.1304	-2.4363	0.3621	-2.1581
IG	184,614	0.1464	-2.2741	0.3777	-1.6195
GA	177,453	0.1609	-2.1715	0.3963	-1.2735
DA	**172,421**	**0.1684**	**-2.1034**	**0.4053**	**-1.2662**

Quantile Residuals (after some model selection)

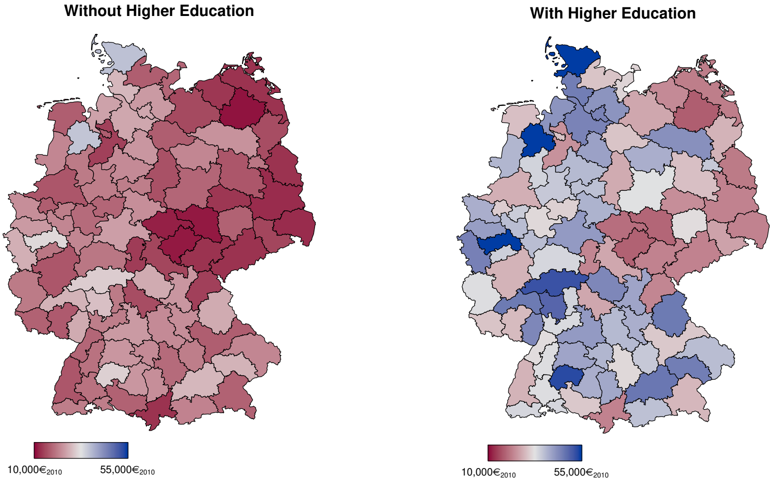


Expected Income for an “Average Man”





Income Standard Deviation for an “Average Man”



Gini Coefficient for an “Average Man”