Stat 225

Rima Izem

Areal Data Analysis

Review...

- Adjacency matrix, two ingredients: proximity and strength
- Testing for spatial association:

Review...

- Adjacency matrix, two ingredients: proximity and strength
- Testing for spatial association: Moran's I, Geary's C.
- Local tests for spatial association

SAR model

SAR: simultaneous autoregressive model, model for exponential family distribution

- Gaussian with mean zero
- Autoregressive regression (lag model and SAR model)
- General model (poisson example)

SAR model, contd

$$\begin{aligned} Y_i &= \sum_j c_{i,j} Y_j + \epsilon_i \\ \text{Observed Values} &= \text{Spatial Signal} + \text{independent residuals} \end{aligned}$$

Observed value is an average of *neighboring* observations (hence the *auto* in autoregression).

SAR model, contd

$$\begin{aligned} Y_i &= \sum_j c_{i,j} Y_j + \epsilon_i \\ \text{Observed Values} &= \text{Spatial Signal} + \text{independent residuals} \end{aligned}$$

Observed value is an average of *neighboring* observations (hence the *auto* in autoregression). Two questions: IS this model well defined?

SAR model, contd

$$
\begin{aligned}
Y_i &= \sum_j c_{i,j} Y_j + \epsilon_i \\
\text{Observed Values} &= \text{Spatial Signal} + \text{independent residuals}
\end{aligned}
$$

Observed value is an average of *neighboring* observations (hence the *auto* in autoregression). Two questions: IS this model well defined? How is matrix C related to Adjacency matrix W?

SAR model, gaussian case

Individual specification (local):

$$\begin{aligned} Y_i &= \sum_j c_{i,j} Y_j + \epsilon_i \\ \text{Observed Values} &= \text{Spatial Signal} + \text{independent residuals} \end{aligned}$$

or equivalently, in Matrix form (global)

$$(I - C)Y = \epsilon$$

where $\epsilon \sim MN(0, D)$, and $D = diag(\sigma_1^2, \ldots, \sigma_n^2)$. Is this model well defined? (i.e. does this model define a valid multivariate distribution)?

SAR model, gaussian case (contd)

In model

$$(I - C)Y = \epsilon, \text{ where } \epsilon \sim N(0, D)$$

ϵ induces the following distribution for Y

$$Y \sim N(0, (I - C)^{-1} D((I - C)^{-1})')$$

if and only if $(I - C)$ is **full rank.**

SAR model, gaussian case (contd)

Common choice for C: $C = \lambda W$

- λ is called the spatial autoregression parameter.
- Model becomes

$$Y_i = \lambda \sum_{j \text{ neighbor } i} w_{ij} Y_i + \epsilon_i$$

SAR model, gaussian case (contd)

Equivalent choice for C: $C = \alpha \tilde{W}$, where $\tilde{W}$ is the weighted adjacency matrix.

- α is called the spatial autocorrelation parameter.
- Model becomes

$$Y_i = \alpha \sum_{j \text{ neighbor } i} \frac{w_{ij}}{\sum_k w_{ik}} Y_i + \epsilon_i$$

Note on eigenvalues and eigenvectors, simple example

- Let $A = \begin{bmatrix} -1 & 2 \\ 0 & 3 \end{bmatrix}$
- If $Au = \lambda u$ (i.e. $(A - \lambda Id) * u = 0$) then u is an eigenvector associated with eigenvalue λ
- eigenvalues, solve the equation $det|A - \lambda Id| = 0$. Find $\lambda = -1$ or 3. Eigenvectors $(1, 0)$ and $(1, 2)$

How to choose λ such that $(I - \lambda W)^{-1}$ exists?

- $(I - \lambda W)$ exists iff $det(I - \lambda W) \neq 0$
- $(I - \lambda W)^{-1}$ exists if

How to choose λ such that $(I - \lambda W)^{-1}$ exists?

- $(I - \lambda W)$ exists iff $det(I - \lambda W) \neq 0$
- $(I - \lambda W)^{-1}$ exists if ($\lambda = 0$ or $\frac{1}{\lambda}$ is not an eigenvalue of W)
- Let β_l and β_s be the largest and smallest eigenvalues of W. A necessary condition for $(I - \lambda W)^{-1}$ to exist is that $\frac{1}{\lambda} < \beta_s$ or $\frac{1}{\lambda} > \beta_l$.
- Note that if W is row standardized, it has eigenvalue 1.

Simultaneous Autoregressive **Regression** Model

$$
\begin{aligned}
Y &= X\beta + C(Y - X\beta) + \epsilon; \text{ or equivalently} \\
\text{Data} &= \text{ Linear trend } + \text{ Spatial signal } + \text{ error} \\
Y &= CY + (I - C)X\beta + \epsilon
\end{aligned}
$$

where X is a set of covariates, ϵ_i's are independent and $\epsilon_i \sim N(0, \sigma_i^2)$

Fitting SAR model

- find parameters $\hat{\beta}$, $\hat{\lambda}$ and $\hat{\sigma}_i^2$'s which maximize the likelihood.
- In spdep, $\sigma_i^2 = d_i * \sigma^2$, where the weights d_i's are provided by user and parameter σ^2 is fitted. In following results, $d_i = 1$ for all i.

Fitting SAR model

$$Y = X\beta + C(Y - X\beta) + \epsilon$$

- Fitted Values: $\hat{Y} = X\hat{\beta} + \hat{\lambda}W(Y - X\hat{\beta})$
- Residuals: $Y - \hat{Y}$
- Fitted linear trend $X\hat{\beta}$
- Fitted spatial signal $\hat{\lambda}W(Y - X\hat{\beta})$

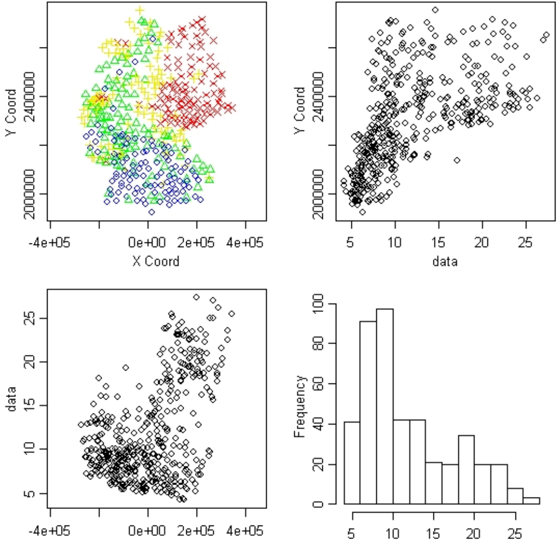
Spatial lag model

Model considered by economists (Anselin, 1988)

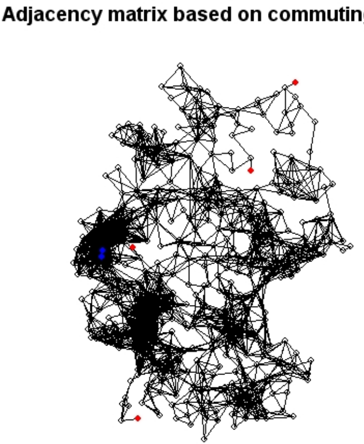
$$
\begin{aligned}
Y - X\beta &= CY + \epsilon; \text{ or equivalently} \\
Y &= CY + X\beta + \epsilon \\
\text{Data} &= \text{Spatial Signal} + \text{Linear Trend} + \text{error}
\end{aligned}
$$

where X is a set of covariates, ϵ_i's are independent and $\epsilon_i \sim N(0, \sigma_i^2)$

German data, EDA



German data, graph used (commuting time less than 60min)



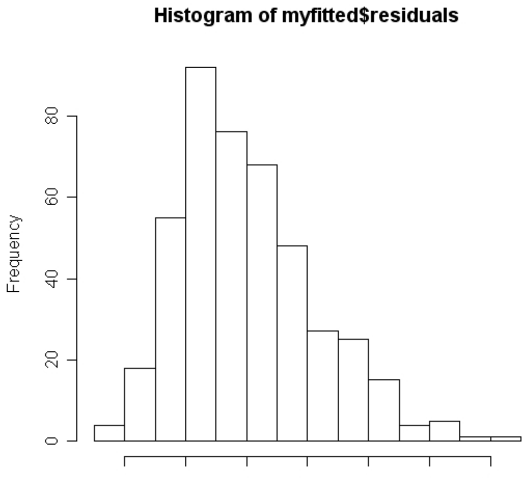
German data, output of spautolm in spdep in R

```
> summary(myfit)

Call: spautolm(formula = URdata[, 4] ~ WE, data = URdata, listw = newMat,
    family = "SAR")

Residuals:
     Min       1Q   Median       3Q      Max
-8.0002 -1.9455 -0.3046  1.5009 10.2143

Coefficients:
             Estimate Std. Error z value  Pr(>|z|)
(Intercept)  19.75545    0.30110  65.610 < 2.2e-16
WE          -10.86587    0.36206 -30.012 < 2.2e-16

Lambda: 0.83165 LR test value: 73.961 p-value: < 2.22e-16

Log likelihood: -1053.717
ML residual variance (sigma squared): 6.9414, (sigma: 2.6347)
Number of observations: 439
Number of parameters estimated: 4
AIC: 2115.4
```

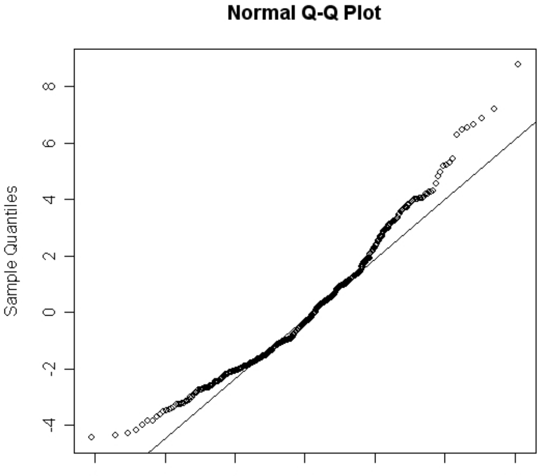
German data, output of spautolm in spdep in R with row standardized matrix

```
Call: spautolm(formula = URdata[, 4] ~ WE, data = URdata, listw = listcomm2)

Residuals:
     Min       1Q   Median       3Q      Max
-4.43755 -1.65348 -0.37015  1.20400  8.78432

Coefficients:
            Estimate Std. Error z value  Pr(>|z|)
(Intercept) 14.87249    0.96673 15.3843 < 2.2e-16
WE          -4.37039    0.79127 -5.5232 3.328e-08

Lambda: 0.86278 LR test value: 160.19 p-value: < 2.22e-16

Log likelihood: -1010.601
ML residual variance (sigma squared): 4.9872, (sigma: 2.2332)
Number of observations: 439
Number of parameters estimated: 4
AIC: 2029.2
```

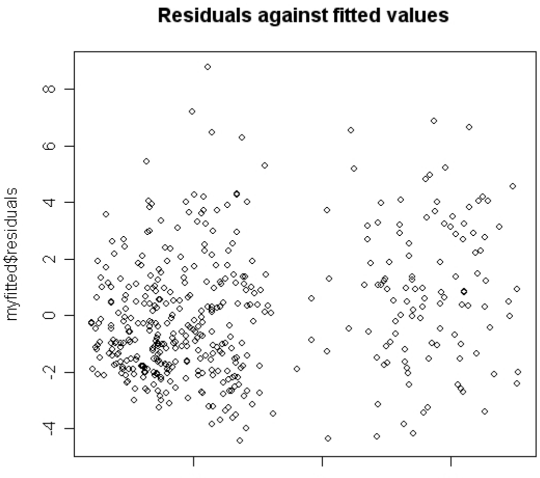
German data, inspection of residuals (SAR) (output with row standardized matrix)



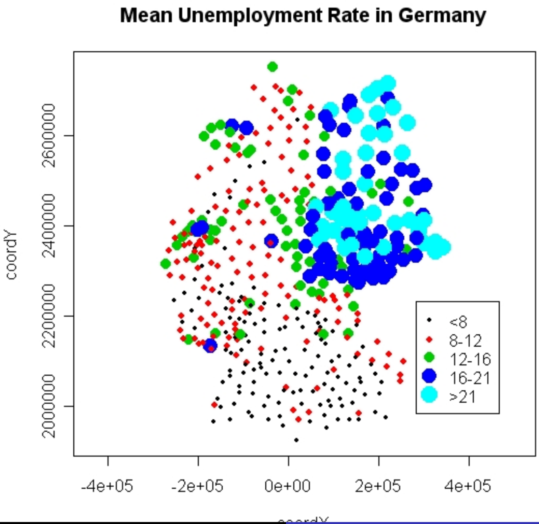
German data, inspection of residuals (SAR) (output with row standardized matrix)



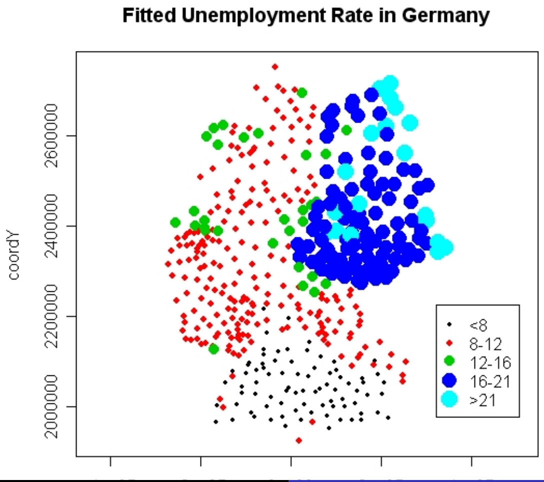
German data, residuals vs fitted values (SAR) (output with row standardized matrix)



German data, observed values



German data, Fitted values (SAR) (output with row standardized matrix)



German data, Residuals (SAR) (output with row standardized matrix)