Stat 225

Rima Izem

Areal Data Analysis

Review...

- Adjacency matrix, two ingredients: proximity and strength
- Testing for spatial association: Moran's I, Geary's C.
- Fitting spatial model: SAR (simultanous autoregressive model)
  - SAR: Gaussian with mean zero
  - SAR: Model with covariates
  - Spatial-lag model.
  - EDA and SAR fits of German Example.

SAR model, contd

$$Y_i = \sum_j c_{i,j} Y_j + \epsilon_i$$

Observed Values = Spatial Signal + independent residuals

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

SAR model, gaussian case (contd)

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

- α is called the spatial autocorrelation parameter, $\alpha < 1$
- Model becomes

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

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 $\hat{\lambda}W(Y - X\hat{\beta})$

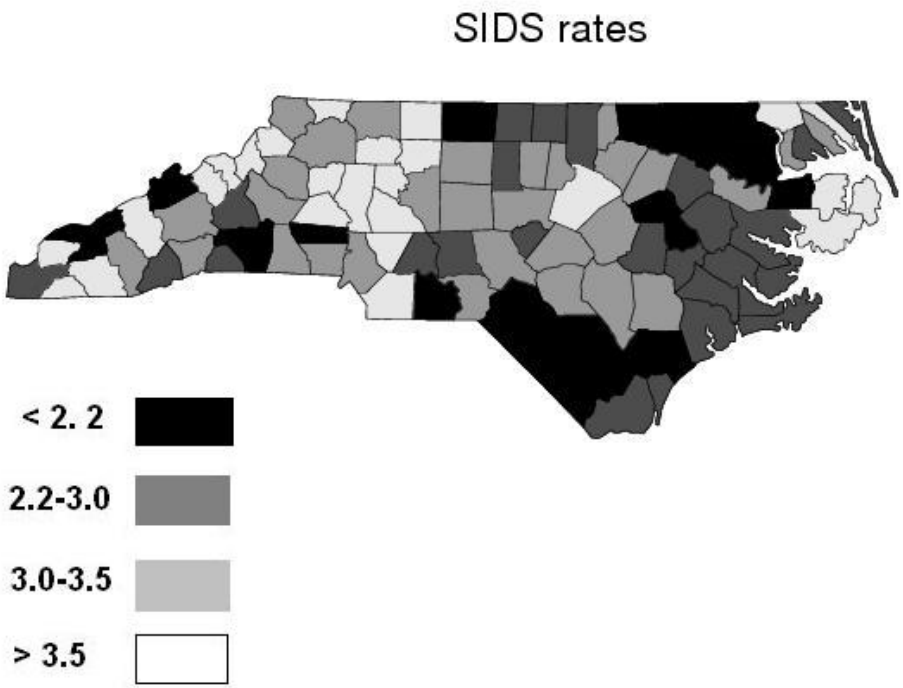
Spatial lag model

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 $\epsilon_i \sim N(0, \sigma_i^2)$

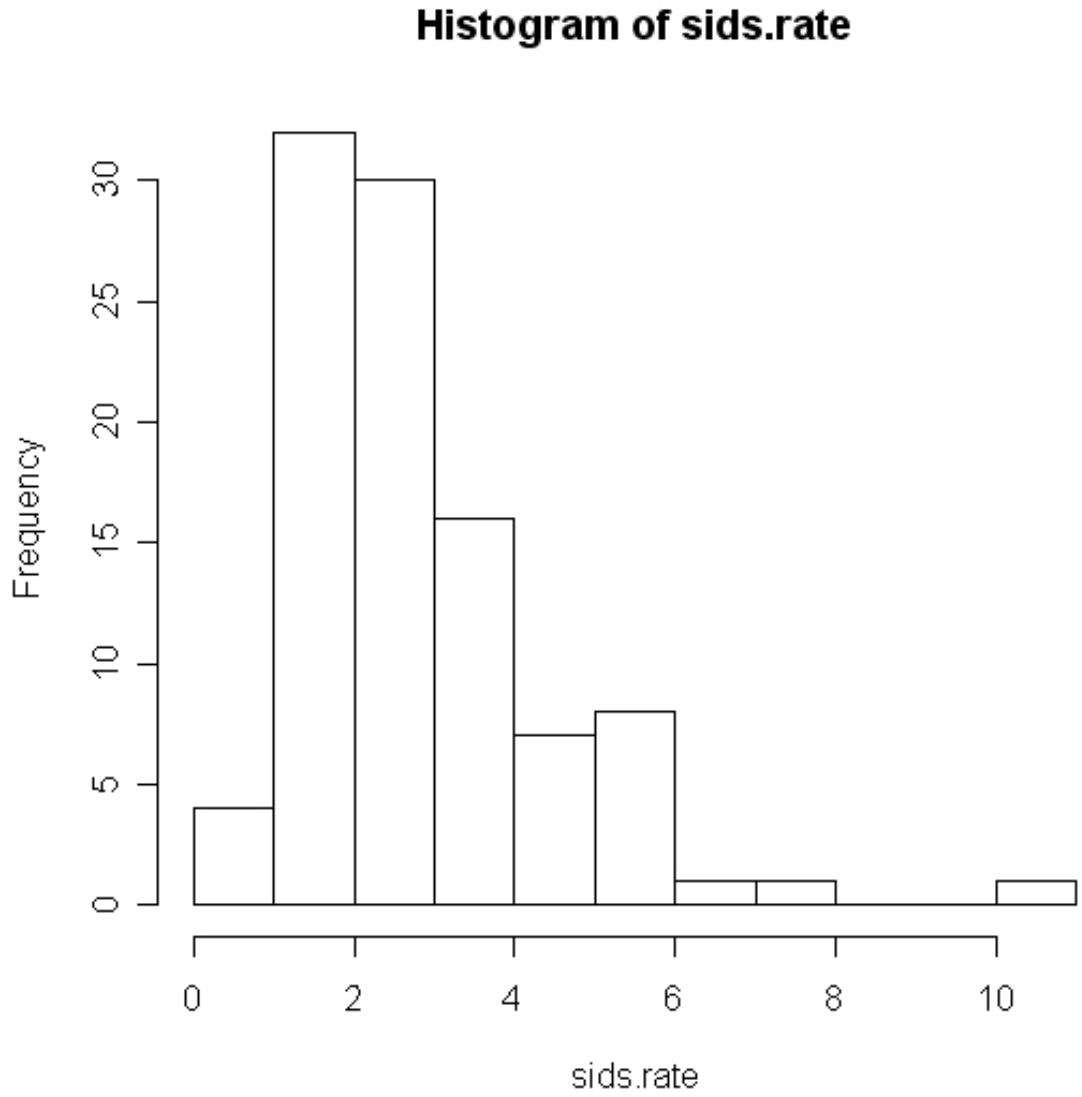
SID data



SID data

- Data: sudden infant mortality death for 1974-78 and 1979-84 in North Carolina counties.
- Source: Cressie 1992. Data also available in R package spdep.
- Explanatory variables considered by Cressie: number of births (white/non-white).

SID data, rates



What to do if the data is not normally distributed

- Approach 1: transform the scale of the data, so that on the transformed scale the variable is normally distributed. We will take this approach for the SID data in the following analysis.
- Approach 2: Assume the data has a known distribution (e.g. Poisson, Binomial, multinomial...etc) and fit a generalized linear model or generalized SAR or CAR. We will see this approach later in the course.

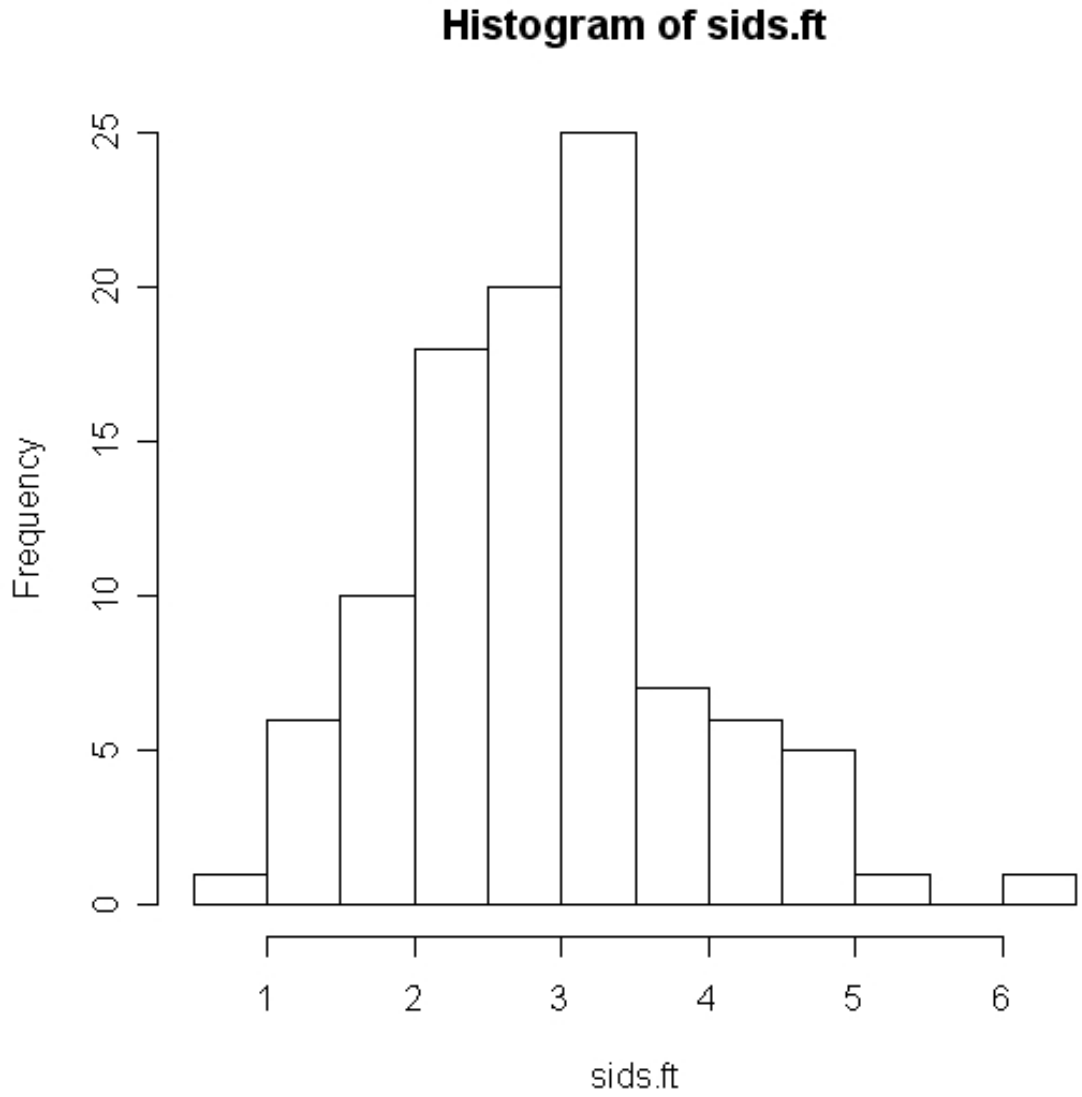
SID data, note

To reduce heteroscedasticity in the data, the following variance transformation was applied to the SID data (Freeman-Tukey square root transformation)

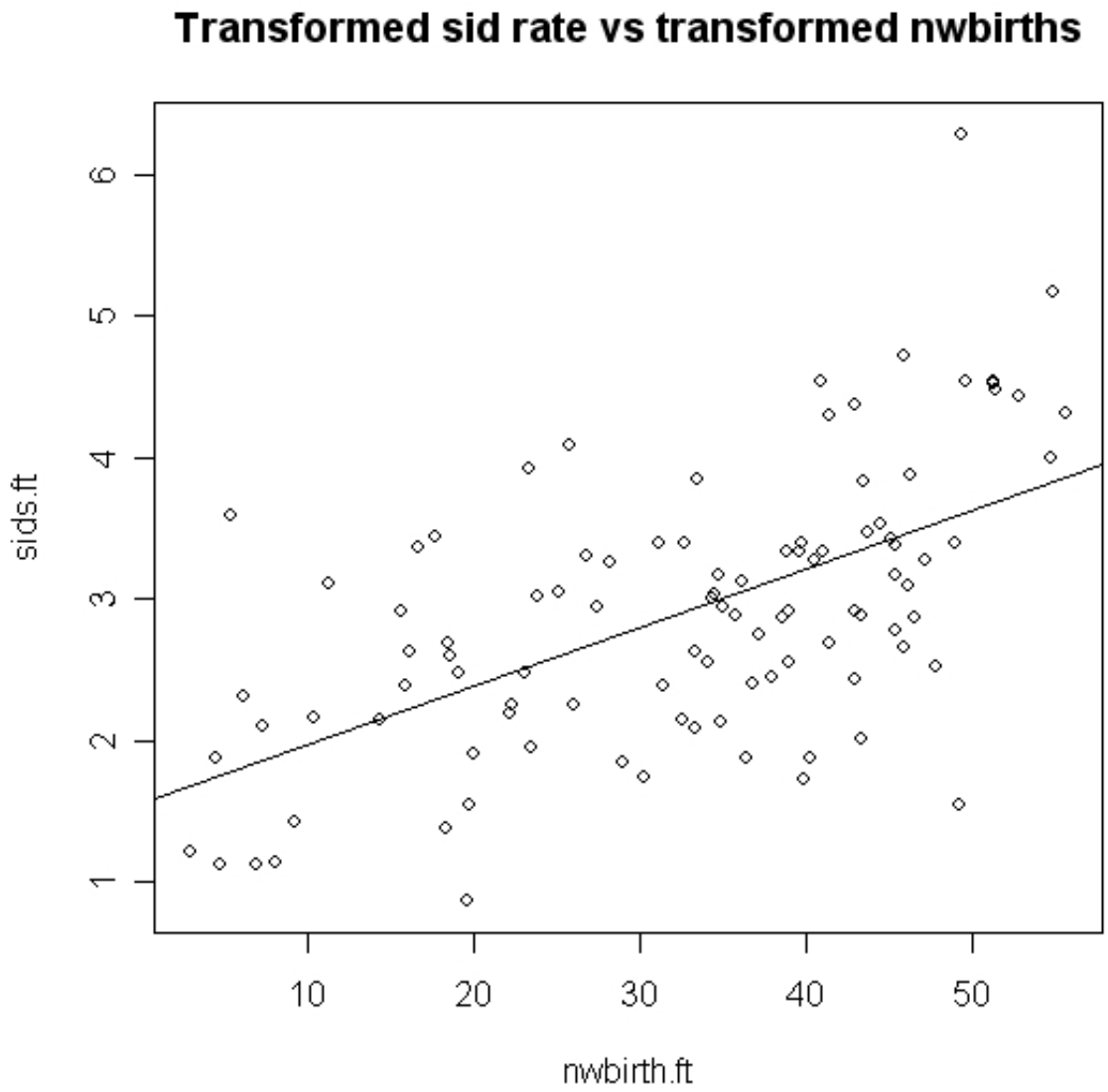
$$Z_i = (1000(S_i)/n_i)^{(1/2)} + (1000(S_i + 1)/n_i)^{(1/2)}$$

where S_i is the number of SIDS, n_i is the number of live births and Z_i is the transformed rate.

SID data, transformed rate

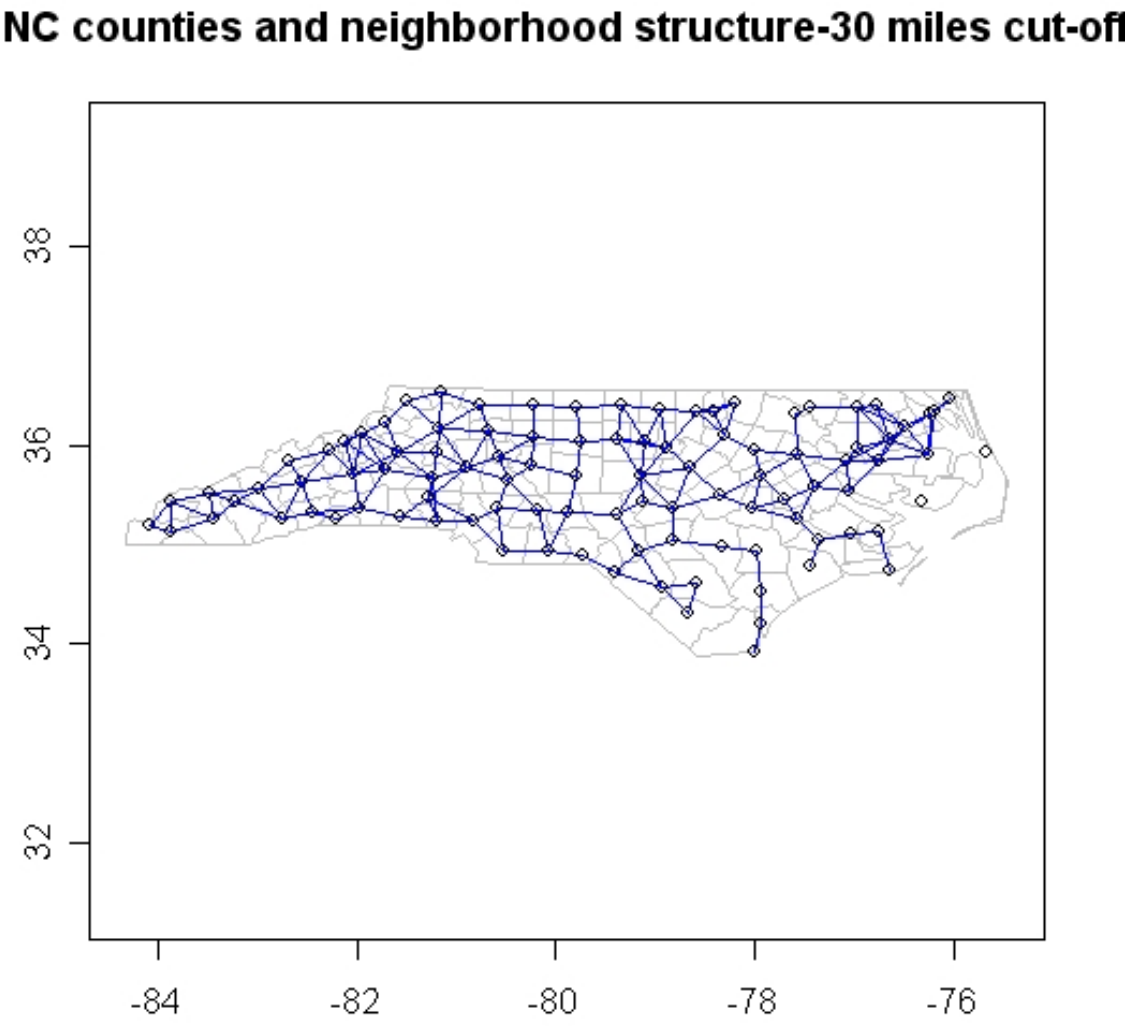


SID data


Transformed sid rate vs transformed nwbirths
sids.ft
nwbirth.ft
1
2
3
4
5
6
10
20
30
40
50

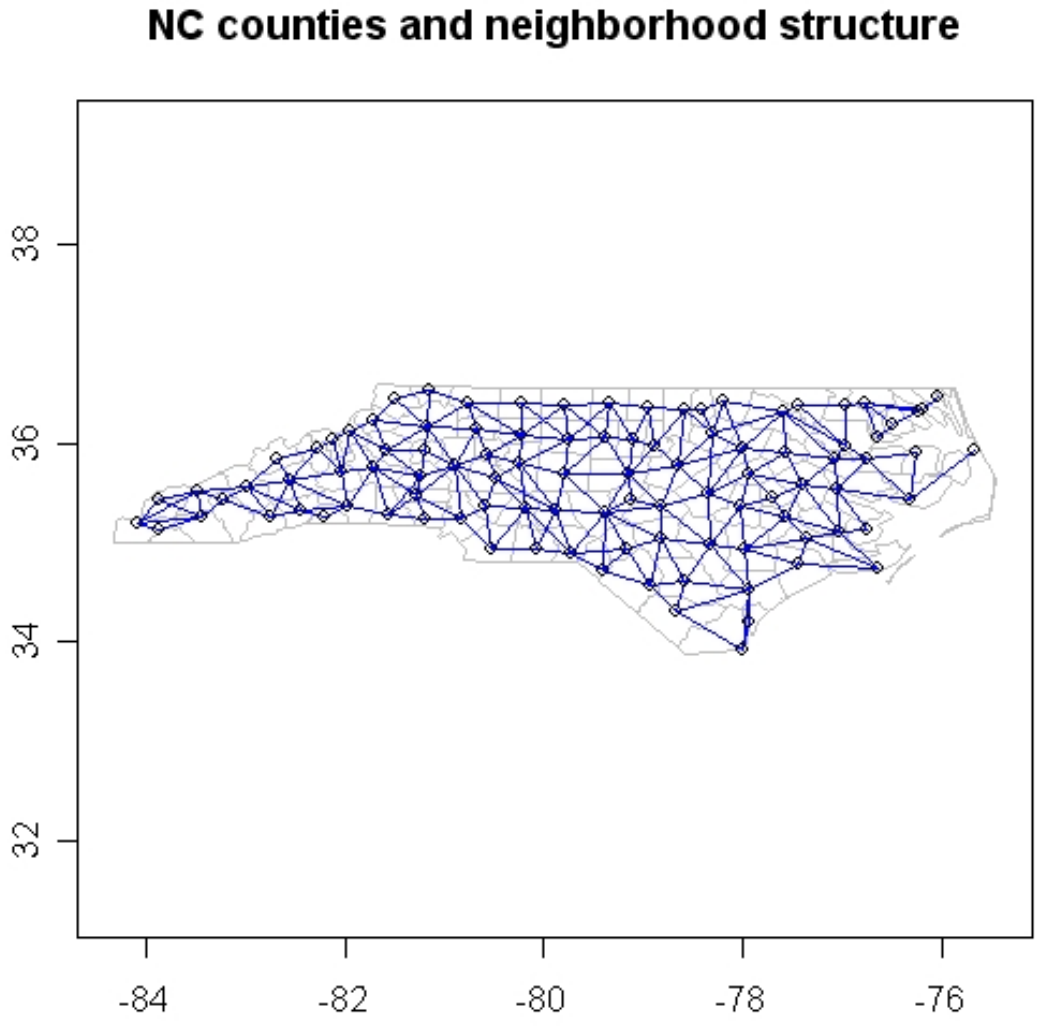
SID data

30 miles cut-off



SID data

Based on sharing a border

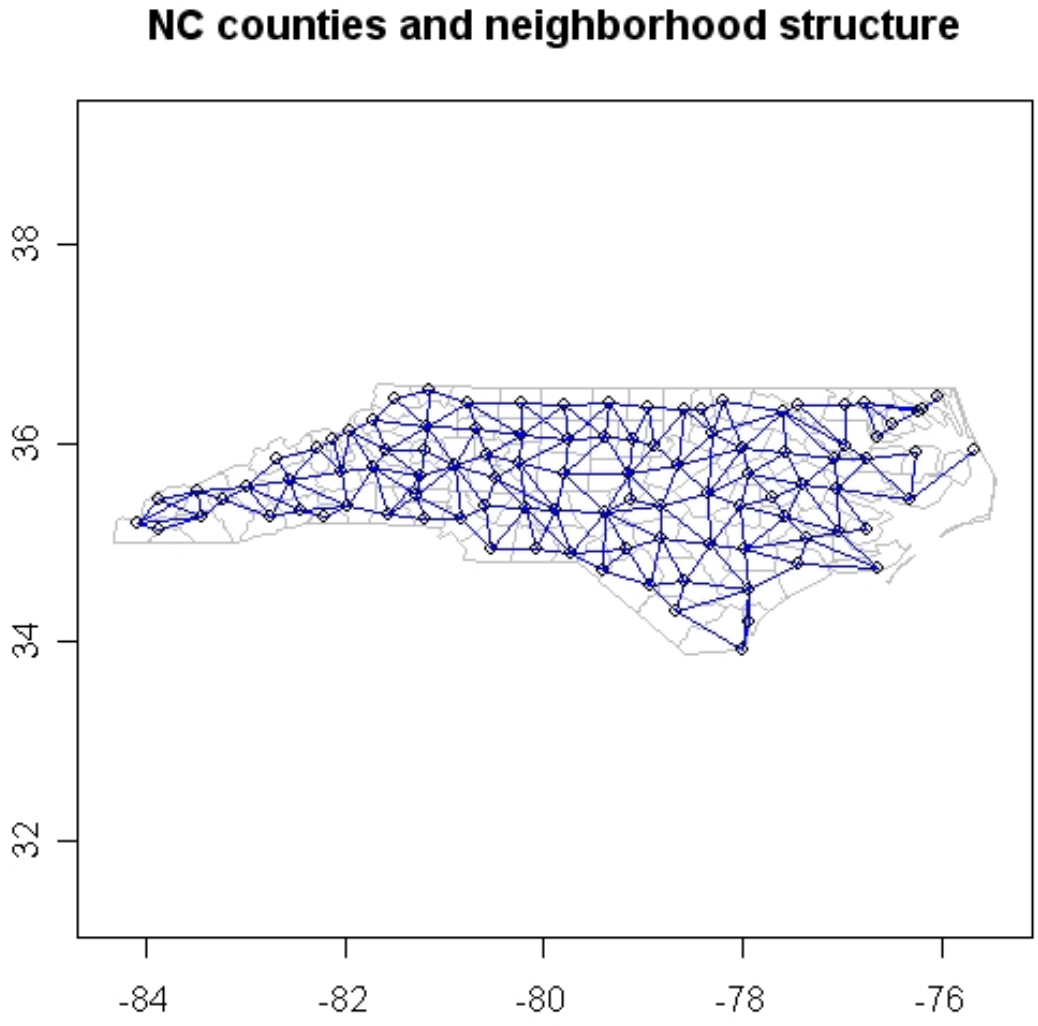


SID data, contd

- Use Freeman-Tuckey transformation
- Fit 1: response = sid 1974-1978, explanatory = nwbirths 1974-1978, Adjacency matrix = row standardized binary matrix of border based neighborhood structure, Homoscedastic model. SAR results: nwbirths is significant, spatial variation not significant.
- Results in following slides, response = sid 1979-84, explanatory = nwbirths 1979-84, Adjacency matrix = distance based or border based, heteroscedastic model.

SID data

Based on sharing a border



SID data, SAR fitting (border based, row standardized binary weights, heteroscedastic, output from R)

Residuals:

Min	1Q	Median	3Q	Max
-2.16119	-0.36993	0.15399	0.54580	2.54710

Coefficients:

	Estimate	Std. Error	z value	Pr($>\|z\|$)
(Intercept)	2.5887933	0.2540358	10.1907	<2e-16
nwbirth.ft.all	0.0059366	0.0070120	0.8466	0.3972

Lambda: 0.22724 LR test value: 2.8547 p-value: 0.091107

Log likelihood: -116.2595

ML residual variance (sigma squared):1529.1, (sigma: 39.104)

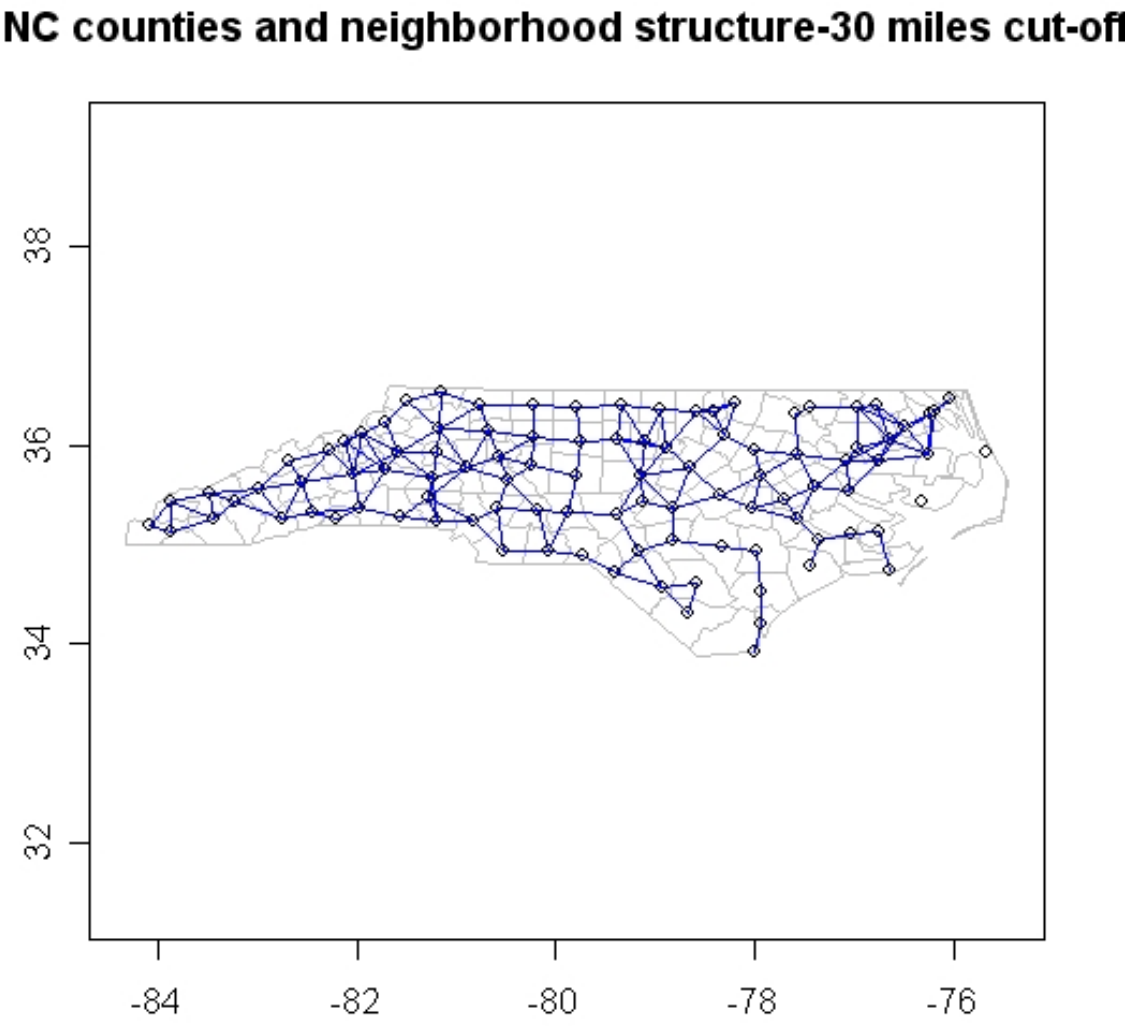
Number of observations: 100

Number of parameters estimated: 4

AIC: 240.52

SID data

Based on 30 miles cut off



SID data

In North Carolina SID data (Cressie, 1992) for the 30 miles cutoff matrix

$$w_{ij} = \frac{min\{d_{ij} : i, j \text{ neighbor of} i\}}{d_{ij}} (\frac{n_j}{n_i})^{0.5}$$

where n_i is number of births of unit i, $d_{i,j}$ is the distance between counties i and j. Note that this adjacency matrix is not symmetric.
In addition, the model was fit with the following heteroscedastic weights

$$\sigma_i^2 = d_i * \sigma^2 = \sigma^2 / n_i$$

(specify option weights equal to number of live births for the spautolm function in R)

SID data, SAR fitting (30 miles cut-off, R output, with options described in previous slide)

Residuals:

Min	1Q	Median	3Q	Max
-2.10009	-0.36822	0.13819	0.57126	2.48276

Coefficients:

	Estimate	Std. Error	z value	Pr($>\|z\|$)
(Intercept)	2.6243889	0.2485583	10.5584	<2e-16
nwbirth.ft	0.0056845	0.0068164	0.8339	0.4043

Lambda: 0.039313 LR test value: 3.4315 p-value: 0.063964

Log likelihood: -112.9878

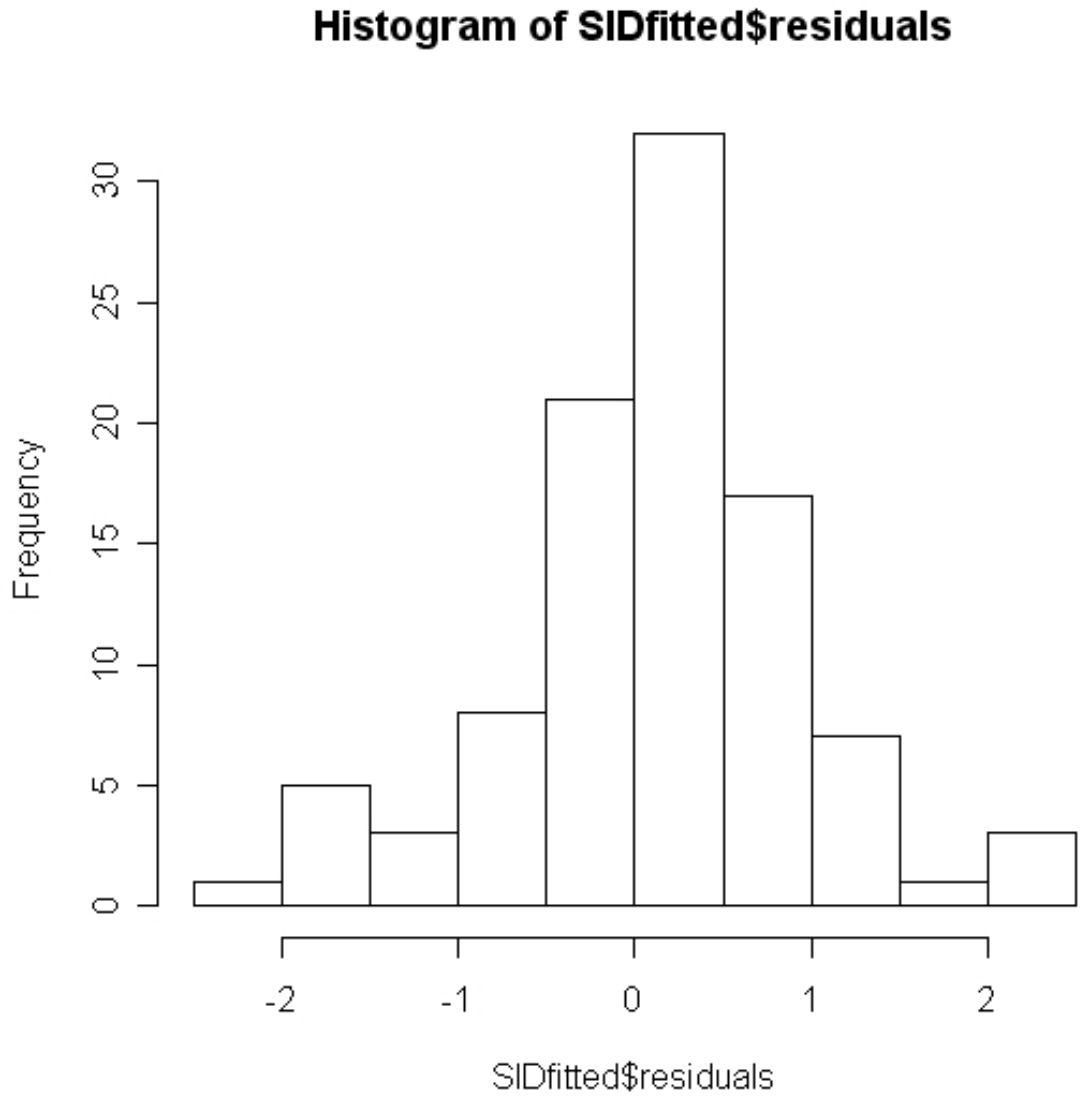
ML residual variance (sigma squared):1545, (sigma: 39.306)

Number of observations: 98

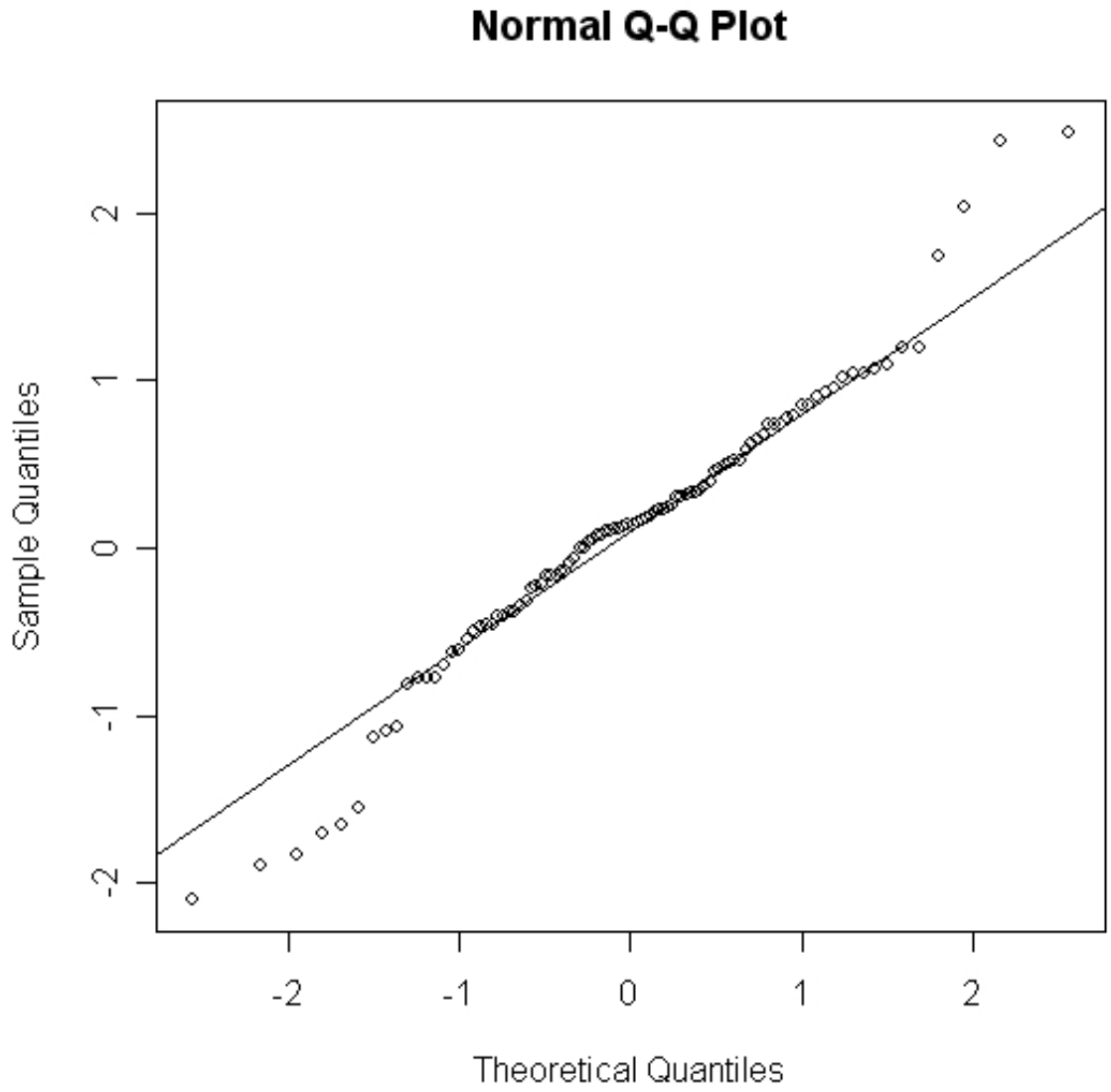
Number of parameters estimated: 4

AIC: 233.98

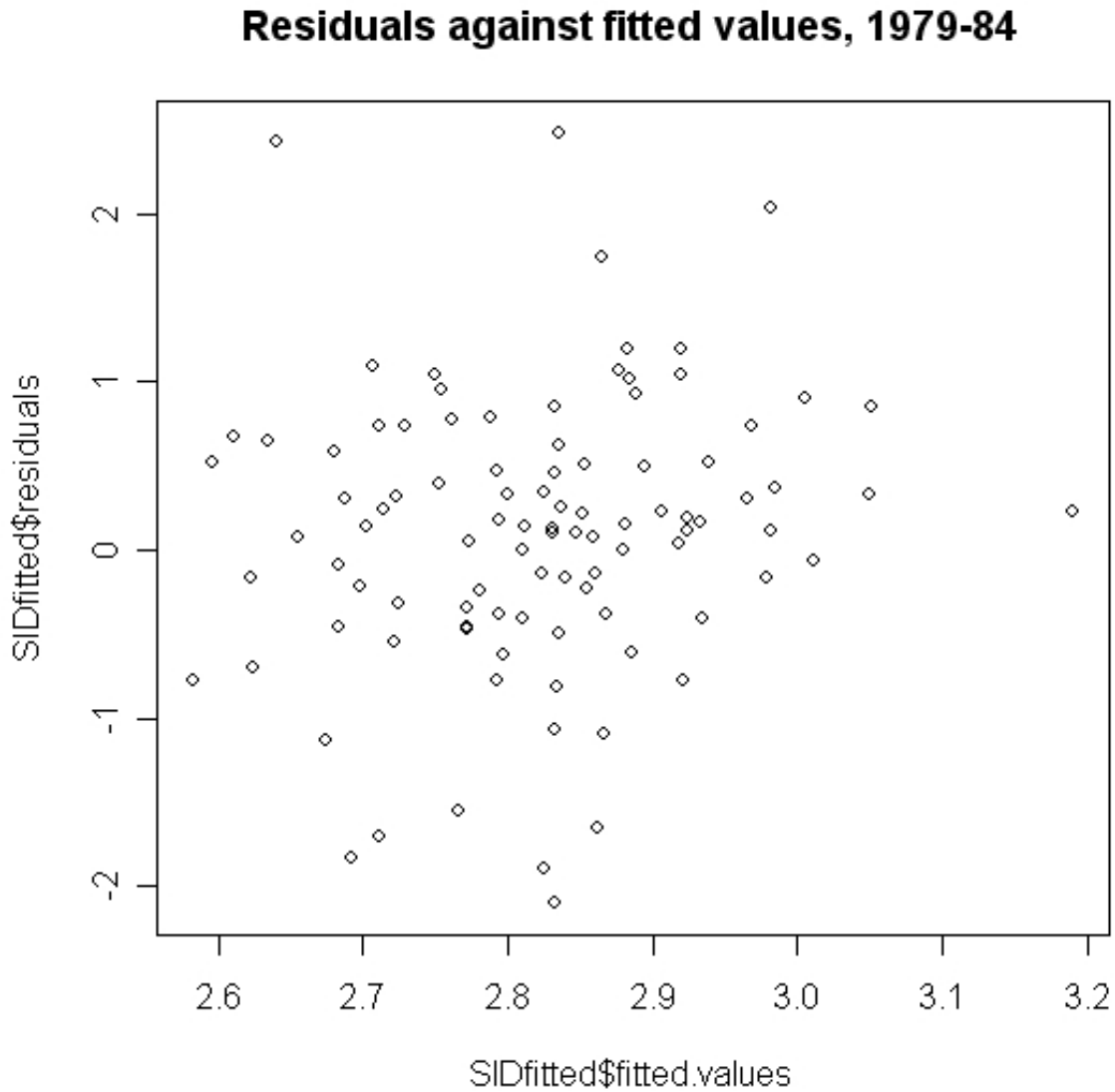
SID data, inspection of residuals (SAR)



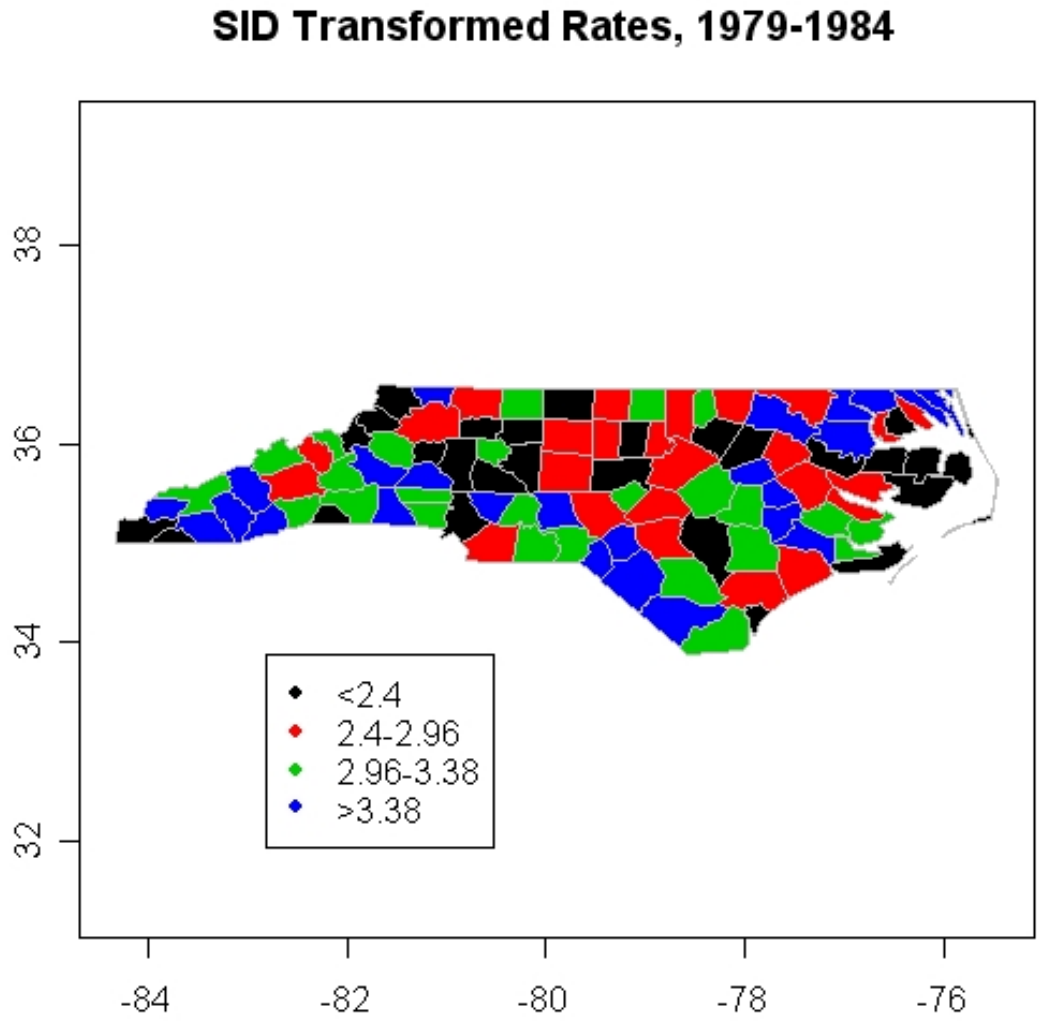
SID data, inspection of residuals (SAR)



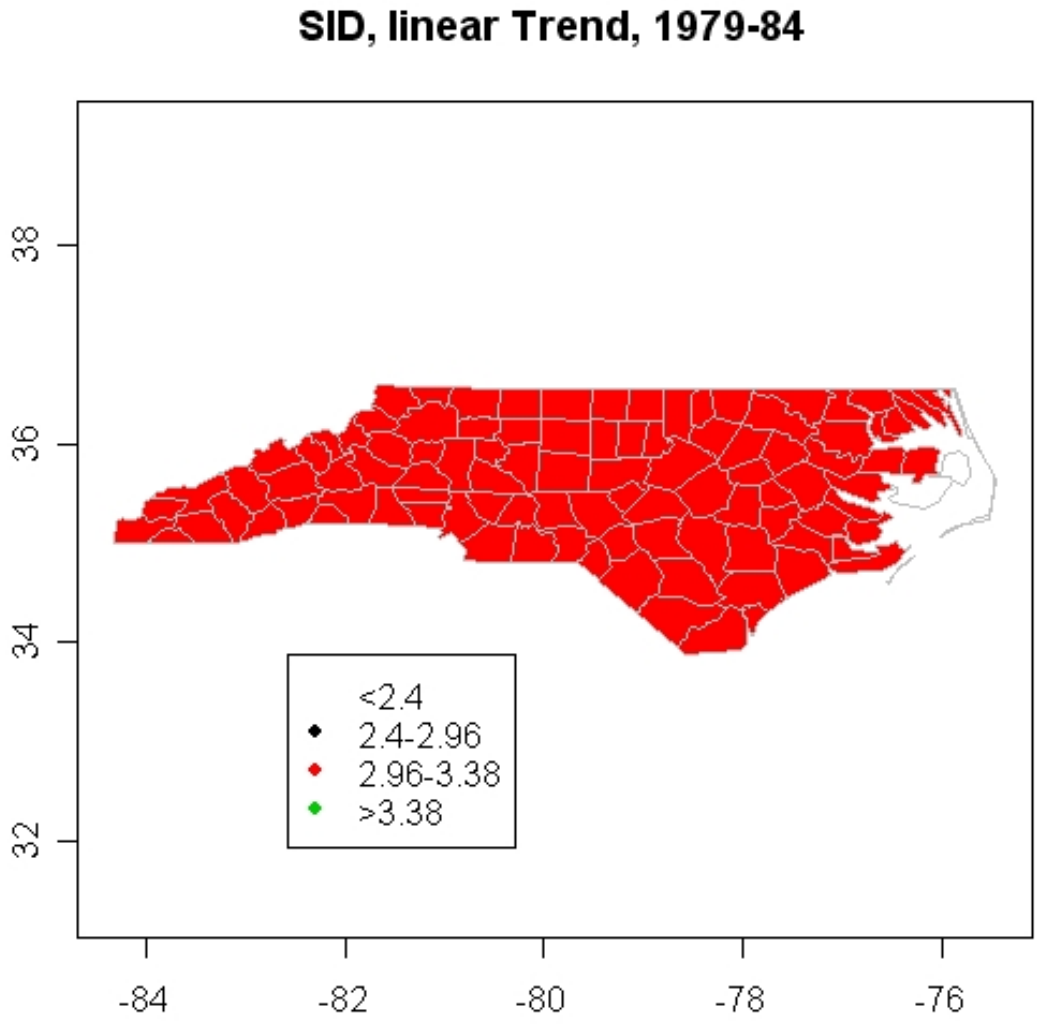
SID data, residuals vs linear trend (SAR)



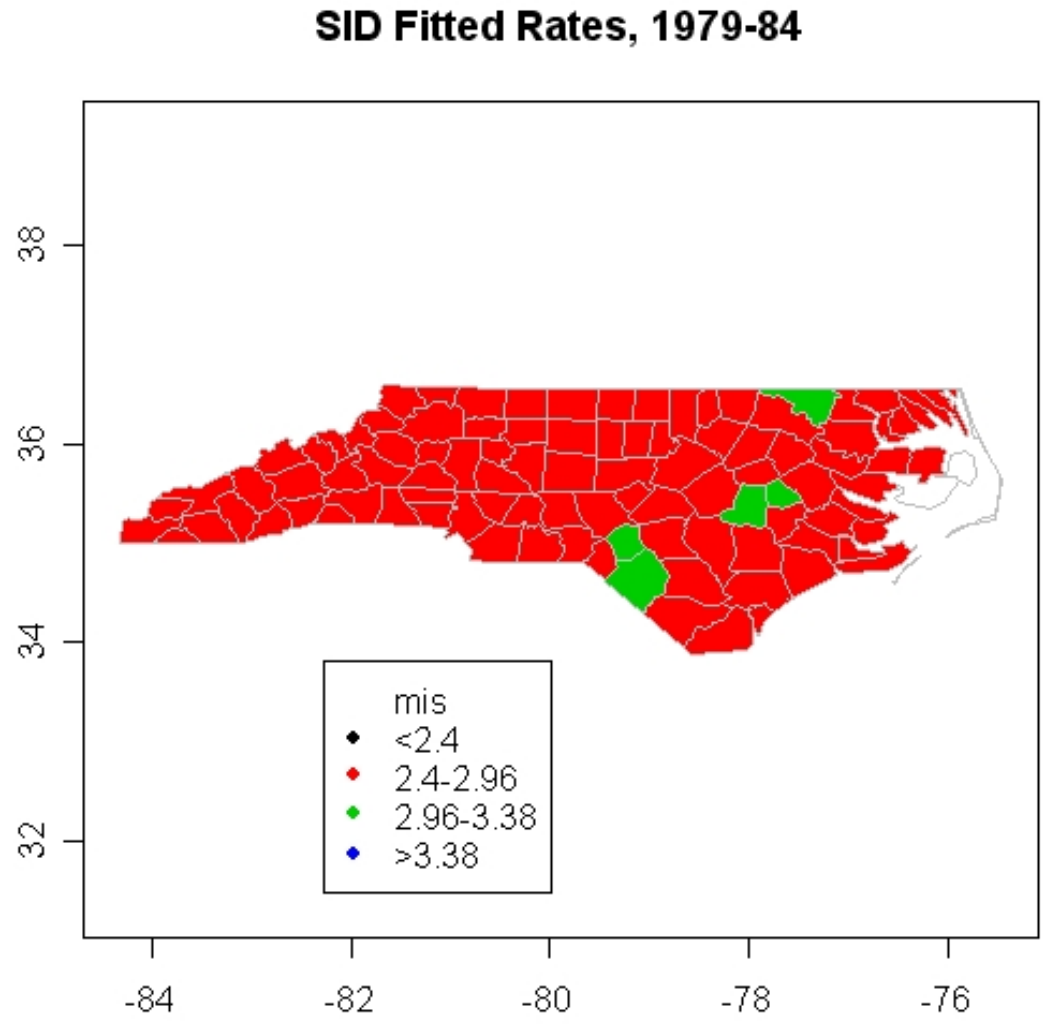
SID data, (transformed) rates (SAR)



SIDS, Linear regression (Fitted) part (SAR)



SIDS, Predictions (SAR)



CAR model, gaussian case

$$Y_i|(y_j, j \neq i) \sim N(\sum_j b_{ij} y_j, \sigma_i^2)$$

- We can recover the joint distribution of Y (if exists) from full conditionals (Brook's Lemma)
- When does this model define a *proper* joint distribution of Y?

CAR model, gaussian case (contd)

Joint density of Y for CAR model (if exists) is

$$Y \sim N(0, (I - B)^{-1}\Delta)$$

where $B = (b_{ij})$, and $\Delta = diag(\sigma_1^2, \ldots, \sigma_n^2)$. The joint density is proper iff

CAR model, gaussian case (contd)

Joint density of Y for CAR model (if exists) is

$$Y \sim N(0, (I - B)^{-1}\Delta)$$

where $B = (b_{ij})$, and $\Delta = diag(\sigma_1^2, \ldots, \sigma_n^2)$. The joint density is proper iff

- $(I - B)$ is invertible, and
- $(I - B)^{-1}\Delta$ is symmetric and of full rank.

CAR Model, gaussian case (contd)

Particular case: CAR model with $B = \tilde{W}$, and $\Delta = \sigma^2 * diag(1/w_{1+}, \ldots, 1/w_{n+})$.

- This case gives an improper joint density because $(I - B)^{-1}\Delta$ is not full rank.
- Corresponding model called the Intrinsically Autoregressive model (IAR)
- Model could be used as a *prior* distribution in a Bayesian model.

CAR Model, gaussian case (contd)

Common choice: take $B = \lambda \tilde{W}$, and
$\Delta = \sigma^2 * diag(1/w_{1+}, \ldots, 1/w_{n+})$. $(w_{i+} = \sum_{i,j} w_{i,j})$

- These cases specify proper joint distributions.
- Full conditionals are

$$Y_i | y_j, j \neq i \sim N(\lambda w_{ij} y_j / w_{i+}, \sigma^2 / w_{i+})$$

Note that the conditional mean is a proportion of the average of neighbors.

CAR, symmetry condition

Symmetry condition of $(I - B)^{-1}\Delta$ is equivalent to

$$\frac{b_{i,j}}{\sigma_i^2} = \frac{b_{j,i}}{\sigma_j^2} \text{ for all } i, j$$

If we choose B s.t. $B = \lambda W$, for some λ, then which W and Δ would satisfy the above condition?

- If $w_{i,j} = 0$ then

CAR, symmetry condition

Symmetry condition of $(I - B)^{-1}\Delta$ is equivalent to

$$\frac{b_{i,j}}{\sigma_i^2} = \frac{b_{j,i}}{\sigma_j^2} \text{ for all } i, j$$

If we choose B s.t. $B = \lambda W$, for some λ, then which W and Δ would satisfy the above condition?

- If $w_{i,j} = 0$ then $w_{j,i} = 0$. So, the neighborhood structure has to be symmetric, i.e. if i is neighbor of j then j has to be a neighbor of i.
- If W is symmetric, then

CAR, symmetry condition

Symmetry condition of $(I - B)^{-1}\Delta$ is equivalent to

$$\frac{b_{i,j}}{\sigma_i^2} = \frac{b_{j,i}}{\sigma_j^2} \text{ for all } i, j$$

If we choose B s.t. $B = \lambda W$, for some λ, then which W and Δ would satisfy the above condition?

- If $w_{i,j} = 0$ then $w_{j,i} = 0$. So, the neighborhood structure has to be symmetric, i.e. if i is neighbor of j then j has to be a neighbor of i.
- If W is symmetric, then $\sigma_i = \sigma$ for all i (homoscedastic model)
- If use standardized weights , i.e. $\tilde{w}_{ij} = \frac{w_{i,j}}{w_{i+}}$, then $\sigma_i^2 =$

CAR, symmetry condition

Symmetry condition of $(I - B)^{-1}\Delta$ is equivalent to

$$\frac{b_{i,j}}{\sigma_i^2} = \frac{b_{j,i}}{\sigma_j^2} \text{ for all } i, j$$

If we choose B s.t. $B = \lambda W$, for some λ, then which W and Δ would satisfy the above condition?

- If $w_{i,j} = 0$ then $w_{j,i} = 0$. So, the neighborhood structure has to be symmetric, i.e. if i is neighbor of j then j has to be a neighbor of i.
- If W is symmetric, then $\sigma_i = \sigma$ for all i (homoscedastic model)
- If use standardized weights , i.e. $\tilde{w}_{ij} = \frac{w_{i,j}}{w_{i+}}$, then $\sigma_i^2 = \frac{\sigma^2}{w_{i+}}$

CAR, symmetry condition

Symmetry condition equivalent to

$$\frac{b_{i,j}}{\sigma_i^2} = \frac{b_{j,i}}{\sigma_j^2} \text{ for all } i, j$$

Other example, for SID data,

$$w_{ij} = \frac{min\{d_{ij} : i, j \text{ neighbor of} i\}}{d_{ij}} \left(\frac{n_j}{n_i}\right)^{0.5}$$

and $\sigma_i^2 = \sigma^2 / n_i$, which satisfies the condition above.

CAR Model, gaussian case with linear trend

$$Y_i|y_j, j \neq i, x_i, \beta \sim N(\sum_j b_{ij} y_j + x_i \beta, \sigma_i^2)$$

where x_i are values of covariates in block i.

CAR Model, fitting

- As for SAR, the parameters are: the regression parameters β, the autoregression parameter λ, and the variance parameter σ_i^2.
- As in SAR, in spdep, $\sigma_i^2 = d_i * \sigma^2$ where d_i's are provided by user. (e.g: $d_i = 1/n_i$)
- WARNING: check that $(I - \lambda W)^{-1}\Delta$ is symmetric

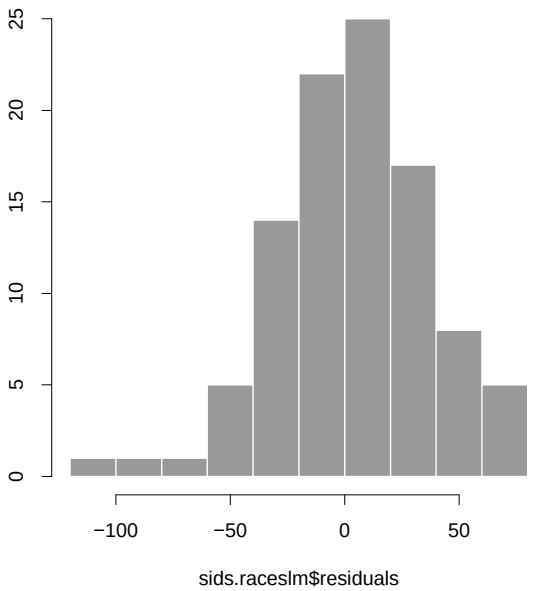
CAR Model, fitting

- As for SAR, the parameters are fit to maximize the likelihood.
- SIDS data, $B = W$, and $\Delta = diag(1/n_1, \ldots, 1/n_n)$, where n_i's are the number of births in county i. This is a proper model.
- In S+Spatial, model fitted for $B = \rho W$ and $\Delta = \sigma^2 Q$ with Q specified by the user to make $(I - B)^{-1}(\Delta)$ symmetric.

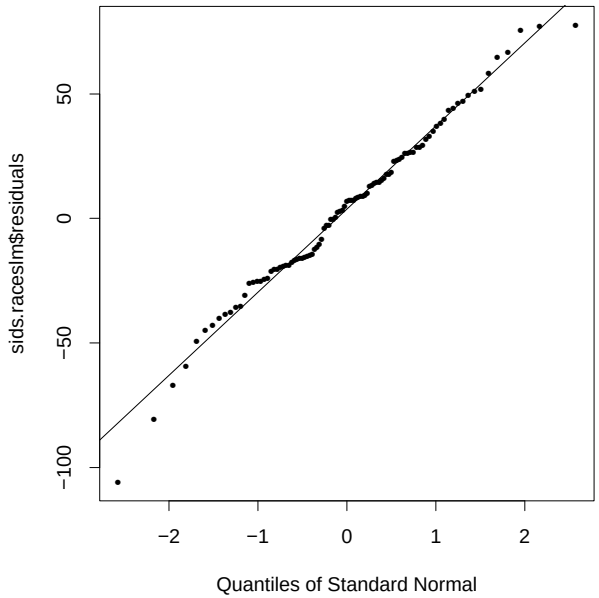
SIDS data, CAR fitting, S+spatial

```
slm(formula = sid.ft ~ nwbirths.ft, cov.family = CAR, data = sids, subset = -4
        , spatial.arglist = list(neighbor = sids.neighbor, region.id = 1:100,
        weights = 1/sids$births))
Residuals:
  Min     1Q Median    3Q  Max
 -106 -18.79   7.01 26.27 77.8

Coefficients:
             Value Std. Error t value Pr(>|t|)
(Intercept) 1.6456 0.2385     6.8990  0.0000
nwbirths.ft 0.0345 0.0066     5.2068  0.0000

Residual standard error: 34.525 on 96 degrees of freedom

Variance-Covariance Matrix of Coefficients
             (Intercept)   nwbirths.ft
(Intercept)  0.056896258 -1.515965e-03
nwbirths.ft -0.001515965  4.402731e-05

Correlation of Coefficient Estimates
             (Intercept) nwbirths.ft
(Intercept)   1.0000000  -0.9578252
nwbirths.ft  -0.9578252   1.0000000

rho =  0.6454

Iterations =  9
Gradient norm =  4.466e-7
Log-likelihood =  -200.4

Convergence:  RELATIVE FUNCTION CONVERGENCE

> sids.raceslm$tau2
[1] 1191.977
```

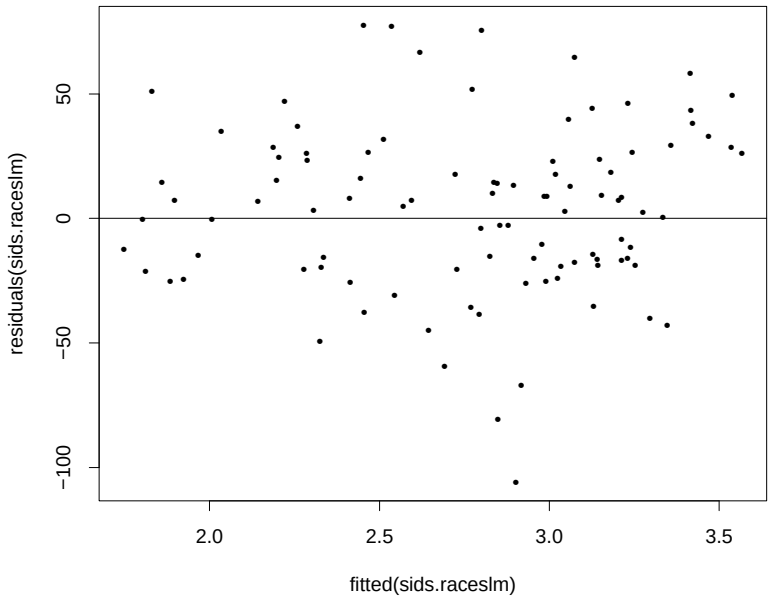
SID data, inspection of residuals (CAR)



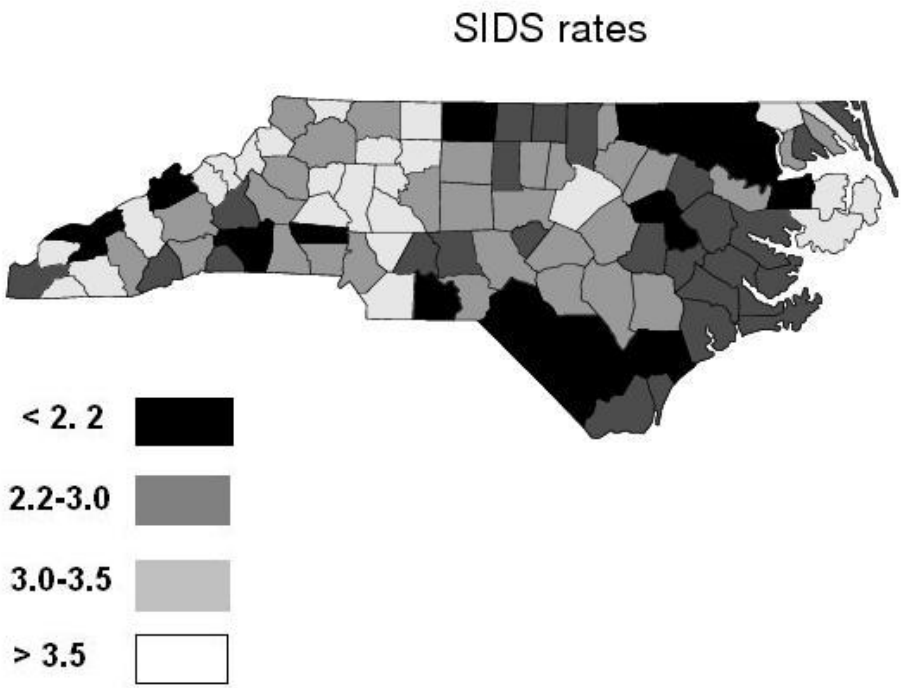
SID data, inspection of residuals (CAR)



SIDS (transformed) rates versus predicted, CAR model



SIDS (transformed) rates



SIDS Linear regression (linear trend) part, CAR model