

# Lecture 4: Regression Modelling for Areal Data

SISMID 2025

Howard Chang  
howard.chang@emory.edu

# Spatial Regression Models

$Y_i$  is a response variable at spatial unit  $i$ . We wish to model:

$$Y_i = \sum_{k=1}^p \beta_k x_{ik} + \epsilon_i = \mathbf{x}_i' \boldsymbol{\beta} + \epsilon_i$$

where  $\mathbf{x}_i$  is a row vector of covariate and  $\epsilon_i$  is the residual error.

**Goals:** Estimate  $\boldsymbol{\beta}$  accounting for spatial autocorrelation in  $\epsilon_i$ .

**Methods:**

1. Conditional Autoregressive Model (CAR) - this focus here
2. Simultaneous Autoregressive Model (SAR)

For now, we will focus on the case where  $Y_i$  can be modeled as a **Gaussian** outcome.

# Conditional Autoregressive Models

Consider the **conditional** specification of the random variable  $Y_i$  given all the other spatial units as:

$$[Y_i | Y_j : j \neq i] \sim N \left( \rho \sum_{j=1}^n b_{ij} Y_j, \tau_i^2 \right) .$$

The above model indicates that the *full conditional* distribution of  $Y_i$ :

- ▶ is Normal;
- ▶ has a mean proportional (via  $\rho > 0$ ) to a linear combination of all other random variables. The weights are defined by  $b_{ij}$  with  $b_{ii} = 0$ ;
- ▶ has a location-specific conditional variance  $\tau_i^2$ .

What is the induced joint distribution? What conditions on  $\rho$ ,  $b_{ij}$ , and  $\tau_i^2$  do we need to ensure a valid joint distribution?

# Conditional Autoregressive Model

To get a valid joint distribution, the conditional distribution of each  $Y_i$  must satisfy

$$[Y_i | Y_j : j \neq i] \sim N \left( \rho \frac{1}{W_{i+}} \sum_{j=1}^n W_{ij} Y_j, \frac{\tau^2}{W_{i+}} \right) .$$

We typically assume  $W_{ij}$  to be binary (indicator for being neighbors).

Then

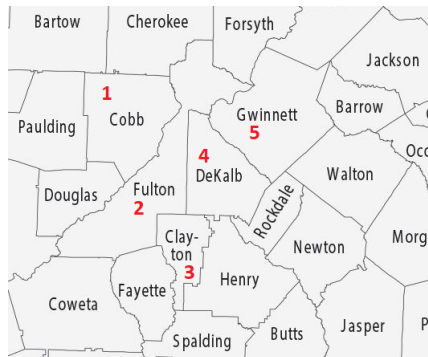
- ▶ The conditional mean of  $Y_i$  is the **weighted average** of all neighbors.
- ▶ The conditional variance is **proportional to the number** of neighbors.

The induced joint distribution is multivariate Normal:

$$[Y_1, Y_2, \dots, Y_n] \sim N(\mathbf{0}, \tau^2(\mathbf{D}_w - \rho \mathbf{W})^{-1})$$

where  $\mathbf{D}_w = \text{diag}(W_{1+}, W_{2+}, \dots, W_{n+})$  and  $\mathbf{W}$  is the adjacency matrix.

# An Example: 5-county Atlanta



$$W = \begin{bmatrix} 0 & 1 & 0 & 0 & 0 \\ 1 & 0 & 1 & 1 & 0 \\ 0 & 1 & 0 & 1 & 0 \\ 0 & 1 & 1 & 0 & 1 \\ 0 & 0 & 0 & 1 & 0 \end{bmatrix}$$

$$W_{1+} = 1 \quad W_{2+} = 3 \quad W_{3+} = 2 \quad W_{4+} = 3 \quad W_{5+} = 1$$

## An Example: 5-county Atlanta

$$[Y_i | Y_j : j \neq i] \sim N \left( \rho \frac{1}{W_{i+}} \sum_{j=1}^n W_{ij} Y_j, \frac{\tau^2}{W_{i+}} \right) .$$

Conditional Distribution for  $Y_1$ ,  $Y_2$  and  $Y_3$ :

$$Y_1 | Y_2, Y_3, Y_4, Y_5 \sim N(\rho Y_2, \tau^2)$$

$$Y_2 | Y_1, Y_3, Y_4, Y_5 \sim N(\rho(Y_1 + Y_3 + Y_4)/3, \tau^2/3)$$

$$Y_3 | Y_1, Y_2, Y_4, Y_5 \sim N(\rho(Y_2 + Y_4)/2, \tau^2/2)$$

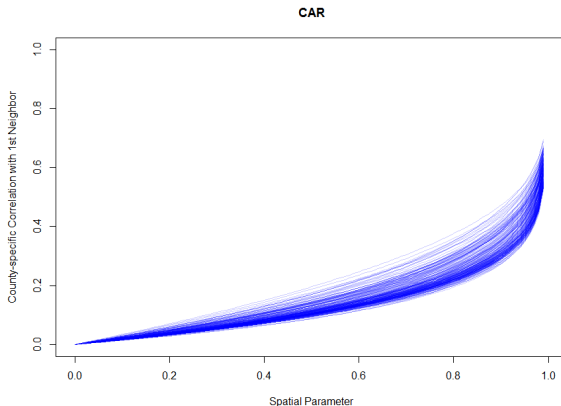
Joint Distribution for  $Y_1$ ,  $Y_2$  and  $Y_3$ :

$$\begin{bmatrix} Y_1 \\ Y_2 \\ Y_3 \\ Y_4 \\ Y_5 \end{bmatrix} \sim N \left( \begin{bmatrix} 0 \\ 0 \\ 0 \\ 0 \\ 0 \end{bmatrix}, \tau^2 \begin{bmatrix} 1 & -\rho & 0 & 0 & 0 \\ -\rho & 3 & -\rho & -\rho & 0 \\ 0 & -\rho & 2 & -\rho & 0 \\ 0 & -\rho & -\rho & 3 & -\rho \\ 0 & 0 & 0 & -\rho & 1 \end{bmatrix}^{-1} \right)$$

## Impacts of $\rho$ on Average 1st Neighbor Correlation

$$Var[\mathbf{Y}] = \tau^2(\mathbf{D}_w - \rho\mathbf{W})^{-1}, \tau^2 = 1, \mathbf{W} = \text{1st-order}$$

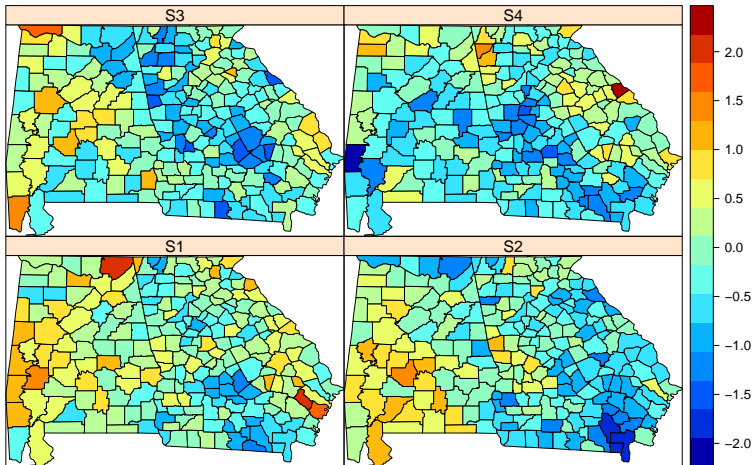
Below we plot the induced average correlation between a county and its neighbors as a function of  $\rho$ .



Note that even with  $\rho = 0.99$ , the correlation is still  $< 0.8$ !

# Example Simulations

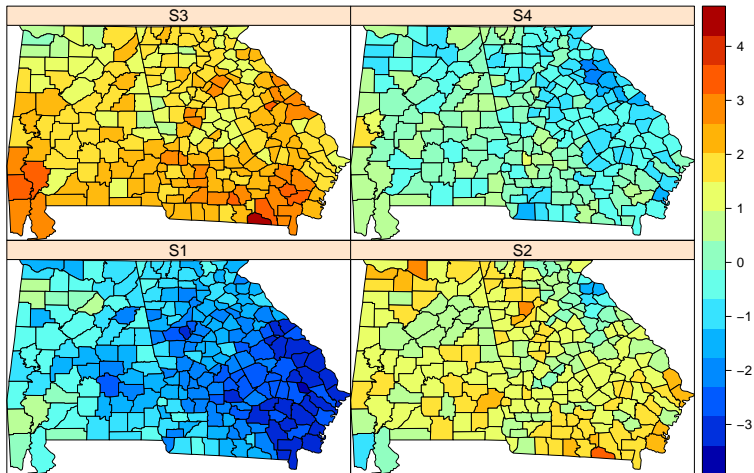
CAR ( $\rho = 0.99, \tau^2 = 1$ )





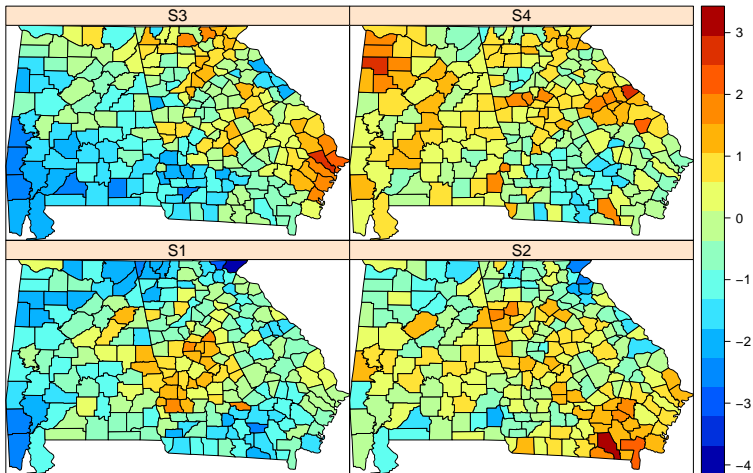
# Example Simulations

CAR ( $\rho = 0.9999, \tau^2 = 1$ )



# Example Simulations

CAR ( $\rho = 0.99, \tau^2 = 2$ )



# CAR Parameters

$$[Y_1, Y_2, \dots, Y_n] \sim N(\mathbf{0}, \tau^2(\mathbf{D}_w - \rho\mathbf{W})^{-1}) .$$

Several important properties of the covariance matrix should be noted.

- ▶  $\tau^2$  cannot be interpreted as the variance at each location. This is because  $(\mathbf{D}_w - \rho\mathbf{W})^{-1}$  is not a correlation matrix. The variance is dependent on  $\rho$  and  $\mathbf{W}$ .
- ▶  $\rho = 0$  implies that  $Y_i$  are independent; but the marginal variance  $(\tau^2\mathbf{D}_w^{-1})$  is proportional to the number of neighbors. This is not a desirable property for independent data!
- ▶ If  $\rho = 1$ , then the joint distribution is invalid because  $\mathbf{D}_w - \mathbf{W}$  is not invertible; it is of rank  $n - 1$ . Therefore,  $\rho = 1$  cannot be used to model observed data; but can be used as prior distribution for random effects (particularly in Bayesian analysis later). When  $\rho = 1$ , we often refer to it as the **intrinsic** or **improper** CAR model.

# CAR Model Example

$$y_i = \beta_0 + \beta_1 x_i + \epsilon_i \quad \epsilon \sim N(0, \tau^2(\mathbf{D}_w - \rho \mathbf{W})^{-1})$$

- ▶  $y_i$  = chlamydia incidence per 10,000.
- ▶  $x_i$  = household income (centered and divided by 1,000).

		Estimate (Standard Error)				
	<b>W</b>	$\beta_0$	$\beta_1$	$\lambda$	$\tau^2$	AIC
LM	NA	46 (1.6)	-1.31 (0.18)	NA	24	2089
CAR	1st-order	42 (3.5)	-1.33 (0.22)	0.15	21	2045
CAR	2nd-order	38 (3.9)	-1.14 (0.21)	0.08	21	2051

- ▶ Household income and chlamydia incidence are negatively associated.
- ▶ Note that standard errors for regression coefficients increase when accounting for spatial correlation.
- ▶ Using AIC as a model selection criterion, CARS with 1st-order adjacency is preferred.

# Advantages of CAR

The spatial dependence structure is contained entirely in  $(\mathbf{D}_w - \rho \mathbf{W})^{-1}$ . The matrix  $\mathbf{Q} = (\mathbf{D}_w - \rho \mathbf{W})$  is known as the **precision matrix**. The CAR model falls under a class of Markov network or undirected graphical model known as **Gaussian Markov random field** (GMRF).

- ▶ GMRF theory says that

$$\mathbf{Q}_{ij} = 0 \quad \Leftrightarrow \quad Y_i \perp\!\!\!\perp Y_j \mid \mathbf{Y}_k, k \neq i, j$$

Therefore, if the proximity matrix  $\mathbf{W}$  is zero between  $Y_i$  and  $Y_j$ , it implies that they are conditionally independent. It is often easier to describe conditional dependence than joint dependence.

- ▶ In practice,  $\mathbf{W}$  contains a large number of zeros. Computation associated with the normal density can be done efficiently with  $\mathbf{Q}^{-1}$  using sparse matrix algorithms (e.g.  $n > 20,000$ ).

# Goals of Disease Mapping

## Setup:

- ▶ Counts of disease incidence or prevalence aggregated over some contiguous administrative spatial areal units.
- ▶ Each count is associated with certain covariates (e.g. at-risk population size) at the same spatial unit.

## Goals:

1. **Regression:** explain spatial variation in diseases rates as a function of covariates.
2. **Smoothing:** provide location-specific estimates that have better precision.

Smoothing = Borrowing Information

# Poisson Generalized Linear Model

For spatial unit  $i = 1, \dots, n$ , assume the following Poisson log-linear model

$$Y_i \sim \text{Poisson}(\lambda_i)$$
$$\log(\lambda_i) = \mathbf{X}_i' \boldsymbol{\beta}.$$

- ▶  $Y_i$  is the observed case count.
- ▶  $\lambda_i$  is the location-specific Poisson mean with fixed-effect  $\boldsymbol{\beta}$  and covariates  $\mathbf{X}$ .

Problems with this model:

1.  $\lambda_i$  is completely determined by the covariates and  $\boldsymbol{\beta}$  is shared across ALL spatial units.
2. Does not account for spatial variability/dependence not explained by  $\mathbf{X}$  (often leading to over/under-dispersion).

# Poisson Generalized Linear Model

First, we can include a location-specific effect ( $u_i$ ) in the mean structure:

$$Y_i \sim \text{Poisson}(\lambda_i)$$
$$\log(\lambda_i) = \mathbf{X}_i' \boldsymbol{\beta} + u_i .$$

- ▶  $u_i$  can be interpreted as the location-specific log baseline mean count.
- ▶  $u_i$  is difficult (almost impossible) to estimate well because we only have one observation at each location!

Under a hierarchical modeling framework, we treat  $u_i$  as random effects:

$$u_i \stackrel{iid}{\sim} N(0, \sigma^2) .$$

This allows shrinkage, information-borrowing, or penalization on  $u_i$  based on the value of  $\sigma^2$ . Smaller  $\sigma^2$  results in higher shrinkage towards 0.



# A Note on Offset

In the Poisson mean, we often include an offset  $\log P_i$ :

$$\log(\lambda_i) = \log P_i + \mathbf{X}_i' \boldsymbol{\beta} + u_i .$$

This implies the Poisson model

$$Y_i \sim \text{Poisson} (P_i \exp\{\mathbf{X}_i' \boldsymbol{\beta} + u_i\}) .$$

Two typical choices for  $P_i$ :

1. at-risk population size; or
2. expected disease count.

With offsets,  $\exp\{\mathbf{X}_i' \boldsymbol{\beta} + u_i\}$  can be interpreted as

1. prevalence/incidence **rates** for the at-risk population; or
2. **relative deviation** from the expected counts.

# Spatial Random Effects

$$\log(\lambda_i) = \log P_i + \mathbf{X}_i' \boldsymbol{\beta} + u_i .$$

Often  $\mathbf{X}$  is not able to remove all the spatial dependence in  $u_i$ . So we can also consider  $u_i$ 's as spatially-dependent random effects.

Let  $\mathbf{u}' = (u_1, \dots, u_n)$ , we assume the joint distribution is Normal with covariance  $\boldsymbol{\Sigma}$ :

$$\mathbf{u} \sim N(\mathbf{0}, \boldsymbol{\Sigma}) .$$

For areal data, the most commonly used spatial random effect specification is the **conditional autoregressive** (CAR) model.

Let  $\mathbf{W}$  denote the symmetric adjacency matrix. The CAR model is given by two parameters  $\rho$  and  $\tau^2$ :

$$\mathbf{u} \sim N(\mathbf{0}, \tau^2 [\mathbf{D} - \rho \mathbf{W}]^{-1})$$

is equivalent to a conditional specification, for  $i = 1, \dots, n$

$$u_i | u_{j \neq i} \sim N \left( \rho \frac{\sum_{j \neq i} W_{ij} u_j}{D_{ii}}, \frac{\tau^2}{D_{ii}} \right)$$

# Proper versus Improper CAR

An **improper** or **intrinsic** CAR is a special case where  $\rho = 1$

$$u_i | u_{j \neq i} \sim N \left( \frac{\sum_{j \neq i} W_{ij} u_j}{D_{ii}}, \frac{\tau^2}{D_{ii}} \right) .$$

However, this does not correspond to a valid joint distribution, i.e.

$\mathbf{u} \sim N(\mathbf{0}, \tau^2 [\mathbf{D} - \mathbf{W}]^{-1})$ , because  $\mathbf{D} - \mathbf{W}$  is not invertible (it has rank  $n - 1$ ).

We can see that if we add a constant  $k$  to all the  $u_i$ , the above distribution still holds.

$$u_i + k | u_{j \neq i} + k \sim N \left( \frac{\sum_{j \neq i} W_{ij} (u_j + k)}{D_{ii}}, \frac{\tau^2}{D_{ii}} \right) .$$

To get around this, we often impose a constraint for improper CAR:

$$\sum_{i=1}^n u_i = 0 .$$

# Improper CAR in Bayesian Analysis

The improper CAR is usually used under a Bayesian framework, which makes inference on the posterior distribution of  $u_i$ :

$$[\mathbf{u} | \mathbf{Y}] = \frac{[\mathbf{Y}, \mathbf{u}]}{[\mathbf{y}]} = \frac{[\mathbf{Y} | \mathbf{u}] \times [\mathbf{u}]}{[\mathbf{y}]} \propto [\mathbf{Y} | \mathbf{u}] \times [\mathbf{u}].$$

So we can use CAR as a prior for  $[\mathbf{u}]$  and the posterior distribution will be valid as long as we have data  $\mathbf{Y}$ .

**Note:** The reason we can do this here is because

- ▶ We assume  $u_i$  is a random effect.
- ▶ **Conditioned** on  $u_i$ , the stochastic nature of  $Y_i$  is given by the Poisson likelihood - NOT by the CAR prior.

# Leroux's Proper CAR

Recall that one disadvantage of the proper CAR model

$$\mathbf{u} \sim N(\mathbf{0}, \tau^2[\mathbf{D} - \rho\mathbf{W}]^{-1})$$

is that it does not reduce to an independent model with  $\rho = 0$ .

Another commonly used structure is the Leroux's form:

$$\mathbf{u} \sim N(\mathbf{0}, \tau^2[(1 - \rho)\mathbf{I} + \rho(\mathbf{D} - \mathbf{W})]^{-1})$$

The above reduces to iCAR when  $\rho = 1$  and an exchangeable model when  $\rho = 0$ . One challenge is that the conditional distribution becomes a bit less intuitive:

$$u_i | u_{j \neq i} \sim N\left(\frac{\rho \sum_{j \neq i} W_{ij} u_j}{1 - \rho + \rho D_{ii}}, \frac{\tau^2}{1 - \rho + \rho D_{ii}}\right).$$

The above conditional mean/variance can be seen as weighted mean between the iCAR and exchangeable model with weight (for iCAR)  $= \rho D_{ii} / (1 - \rho + \rho D_{ii})$ .

# Convolution Model

A convolution model (aka Besag-York-Mollie, BYM) includes both **spatially-dependent** and **unstructured/exchangeable** random effects.

$$\log(\lambda_i) = \log P_i + \mathbf{X}_i' \boldsymbol{\beta} + u_i + v_i .$$

$$\mathbf{u} \sim N(\mathbf{0}, \tau^2 [\mathbf{D} - \mathbf{W}]^{-1})$$

$$v_i \stackrel{iid}{\sim} N(0, \sigma^2) .$$

**Challenge!!!** We are asking our model to partition a location-specific random effect into a spatial and a non-spatial component. With only one  $Y_i$  at each location, this is a very difficult task. In this setting, our main inferential interest may be on  $u_i + v_i$ .

# Chlamydia Application

First consider models without covariates:

$$Y_i \sim \text{Poisson}(\lambda_i) \quad \log(\lambda_i) = \log P_i + \beta_0 + u_i.$$

where  $Y_i$  is the chlamydia counts and  $P_i$  is the population size.

Model 1: No Model (MLE estimates):

$$\hat{\beta}_0 + \hat{u}_i = \log(Y_i/P_i).$$

Model 2: Exchangeable (independent)

$$u_i \sim N(0, \sigma^2)$$

Model 3: Only Spatial (improper CAR)

$$u_i \sim iCAR(\tau^2)$$

Model 4: Only Spatial (proper Leroux)

$$u_i \sim pCAR(\tau^2)$$

Model 5: Convolution (iCAR + exchangeable)

$$u_i = \gamma_i + \theta_i$$

$$\gamma_i \sim iCAR(\tau^2) \quad \theta_i \sim N(0, \sigma^2)$$

# Bayesian Inference

We often frame estimation in disease mapping under a Bayesian framework to get uncertainty estimates for random effects.

Typical priors (we will use later)

- ▶  $\beta_0 \sim \text{Normal}(0, \sigma_\beta^2)$  where  $\sigma_\beta^2$  can be 0 (improper flat prior) or be very large (e.g.,  $100^2$ ).
- ▶ Variance components  $\tau \sim \text{Inv-Gamma}(a, b)$  with small  $a$  and  $b$  (e.g. 0.0001).
- ▶  $\text{Logit}(\rho) \sim \text{Normal}(0, \sigma_\rho^2)$ .

Estimation can be done using MCMC-based methods (JAGS, Stan, custom code) or approximation methods (INLA).



# Model Estimates

## Posterior Median and 95% Posterior Interval

	$\beta_0$	$\tau^2$	$\sigma^2$
Exch.	-5.57 (-5.66, -5.57)		0.42 (0.35, 0.51)
iCAR	-5.57 (-5.58, -5.55)	1.21 (1.13, 1.48)	
pCAR	-5.57 (-5.95, -5.19)	1.08 (0.85, 1.37)	
Conv.	-5.57 (-5.62, -5.52)	0.60 (0.34, 1.03)	0.12 (0.06, 0.19)

For pCAR, estimate of  $\rho$  is 0.81 (0.59, 0.95).

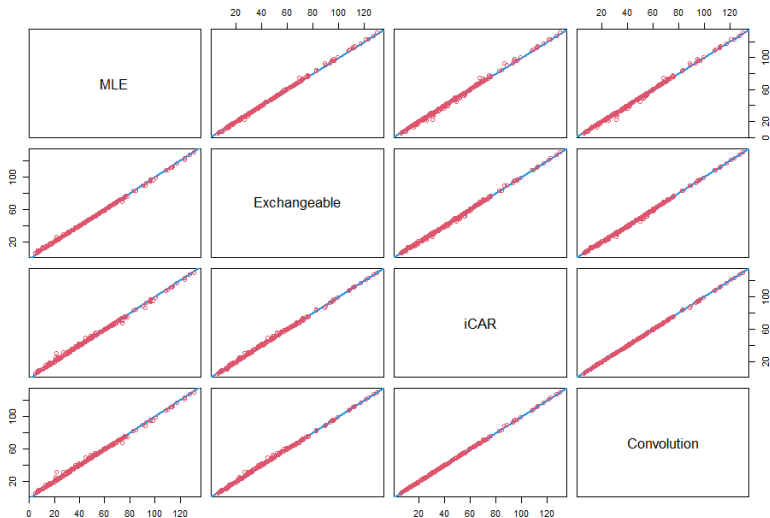
## Model Comparison

	DIC	pD	WAIC	pWAIC
Exch.	1926	218	1873	118
iCAR	1917	211	1864	114
pCAR	1920	213	1870	117
Conv.	1919	213	1868	116

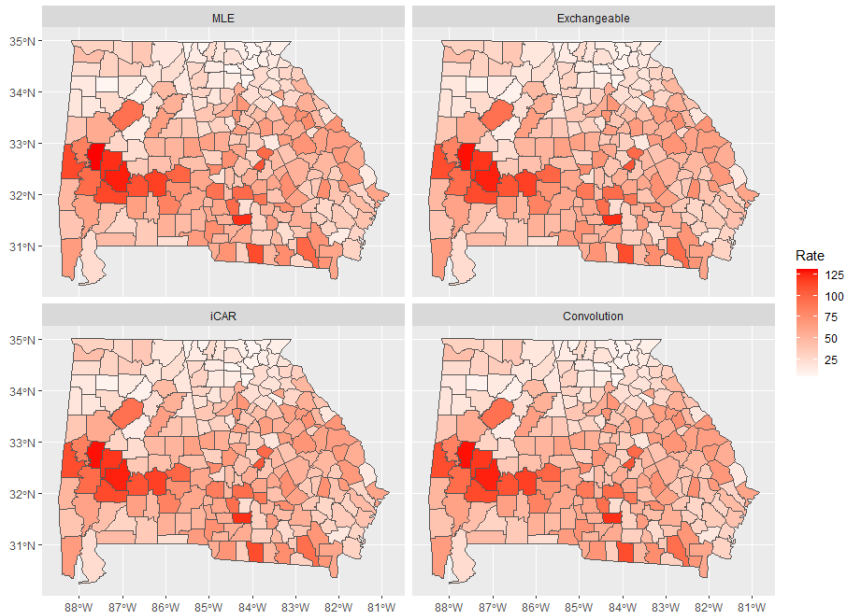
DIC/WAIC prefers the spatial models, which also gives the smallest number of effective degrees of freedom (pD).

# Estimated Relative Risk (per 10,000)

County-specific estimates are very similar.

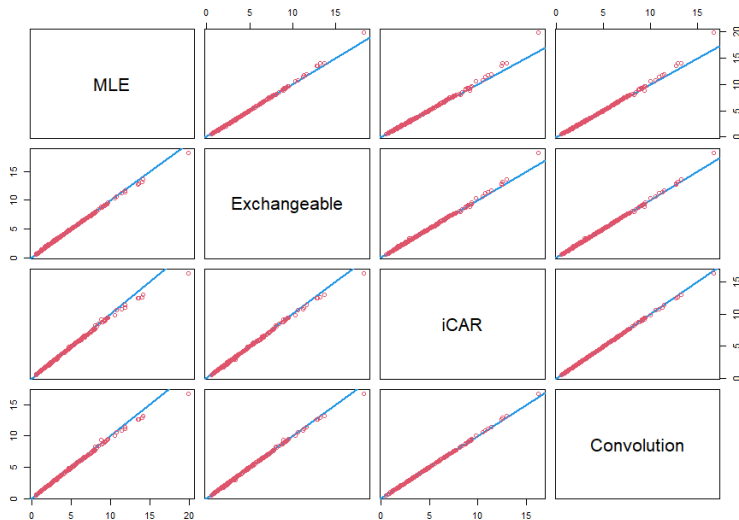


# Estimated Relative Risk (per 10,000)



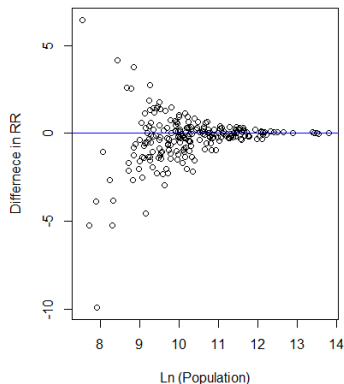
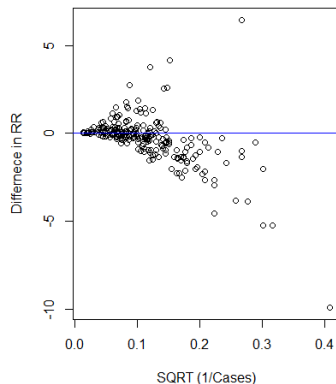
# Comparison of County-Specific Standard Error

Larger SE's from MLE are reduced in random effect models.



# Difference in $\hat{\beta}_0 + \hat{u}_i$

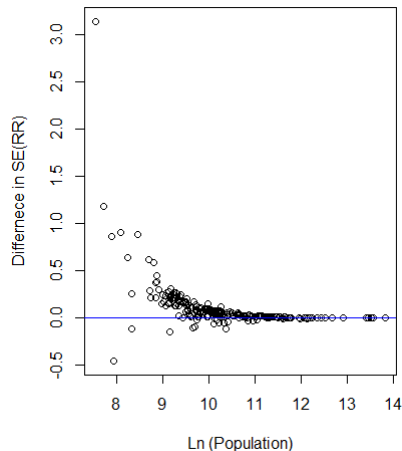
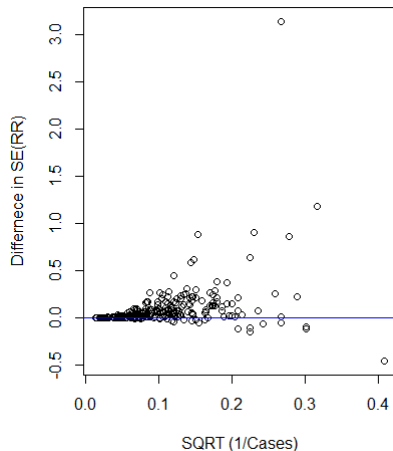
## MLE (Model 1) - Convolution (Model 5)



Larger differences observed when the population is smaller or the cases are larger.

# Difference in $SE(\hat{\beta}_0 + \hat{u}_i)$

MLE (Model 1) - Convolution (Model 5)



# Convolution Model with Covariate

$$Y_i \sim \text{Poisson}(\lambda_i)$$

$$\log(\lambda_i) = \log P_i + \beta_0 + \beta_1 \text{Income}_i + u_i + v_i$$

$$u_i \sim \text{ICAR}(\tau^2) \quad v_i \sim N(0, \sigma^2)$$

	Posterior Mean	95% Post Int
$\beta_0$	-5.56	(-5.60, -5.53)
$\beta_1 (\times 1,000)$	-0.034	(-0.043, -0.024)
$\tau^2$	0.69	(0.41, 1.15)
$\sigma^2$	0.06	(0.02, 0.11)

- ▶  $\sigma^2$  decreased with the inclusion of income in the model.
- ▶ We found that a \$1,000 increase in county-level median household income was associated with a 3.4% (95% PI: 2.4%, 4.3%) decrease in chlamydia incidence, **controlling for county-level baseline heterogeneity**.

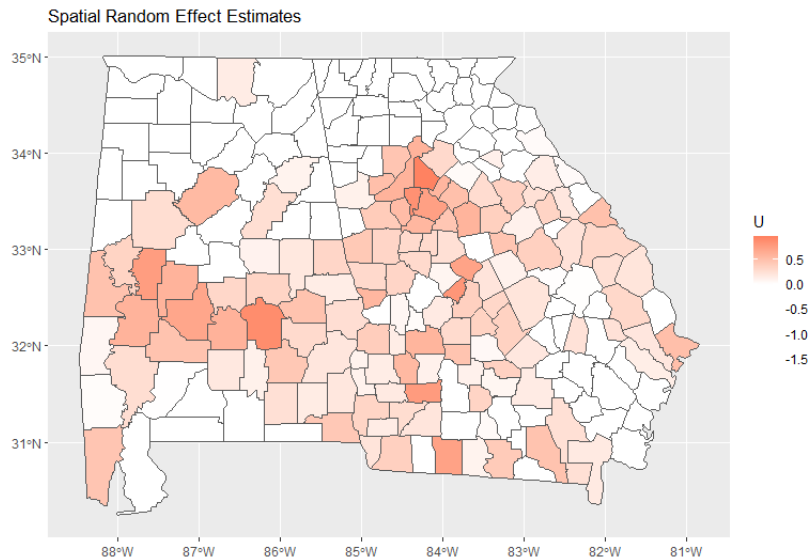
# Comparison with Other Models

	Estimates of $\beta_1 (\times 1,000)$		
	Posterior Mean	SE	95% PI/CI
Spatial Convolution	-0.034	0.0049	(-0.043, -0.024)
Poisson Regression	-0.023	0.0004	(-0.024, -0.022)
Quasi-poisson Regression	-0.023	0.0033	(-0.016, -0.029)

- ▶ A standard Poisson regression under-estimates the uncertainty considerably.
- ▶ Using quasi-Poisson to account for over-dispersion increases confidence interval length (by a lot).
- ▶ The point estimates differ as well.



# Posterior Mean of Spatial Residuals $\gamma_i$



# Posterior Mean of Independent Residuals $\theta_i$

