BST 6200 Spatial Statistics and Disease Mapping

Chapter 7: Spatial Attribute Analysis

Steven E. Rigdon

Spring 2020

Chapter 7: Spatial Attribute Analysis

Section 7.1 Introduction (Read)

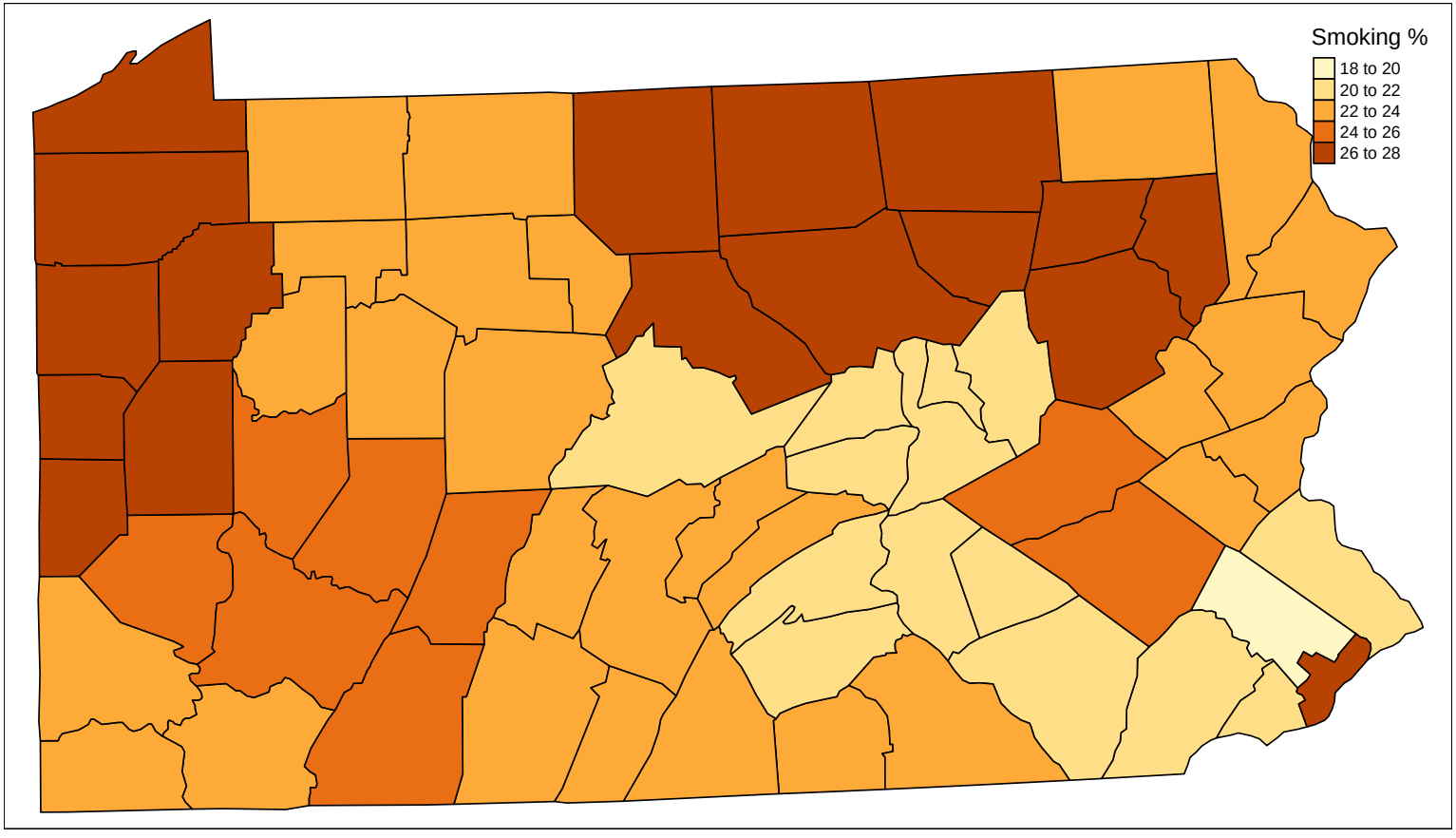
Section 7.2 Pennsylvania Lung Cancer Data

Outcomes for neighboring regions are **not independent**.

This chapter is about measuring and accounting for this lack of independence.

Look at R code ...

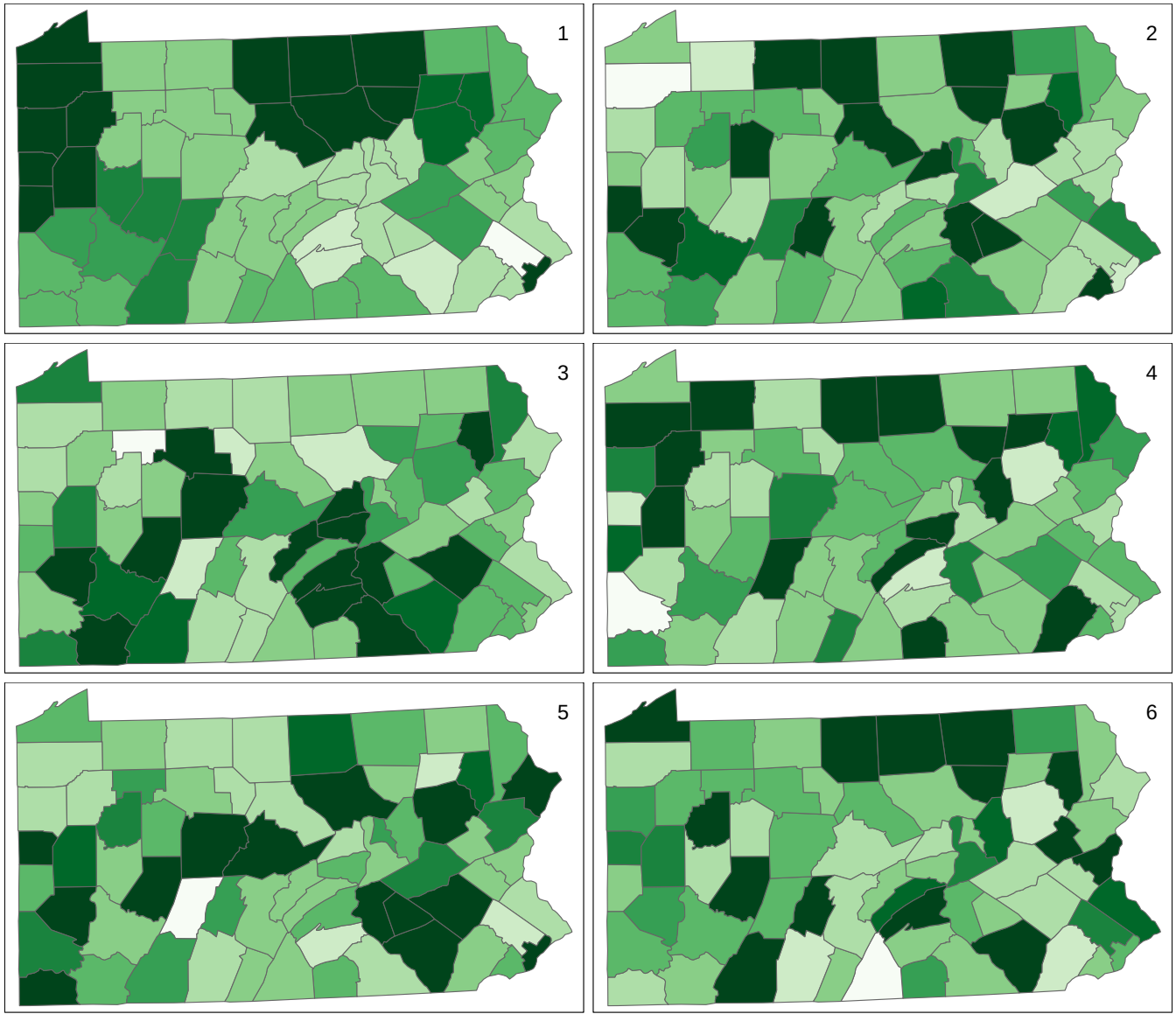
Choropleth Map for Pennsylvania Smoking



Section 7.3: Visual Exploration of Autocorrelation

Using simulation to gauge the strength of evidence against independence.

Randomly "shuffle" the 67 counties' smoking rates and plot the resulting choropleth map. Plot the **actual** choropleth map. Can you tell which one is real and which ones are shuffled?


1
2
3
4
5
6

Defining Neighbors

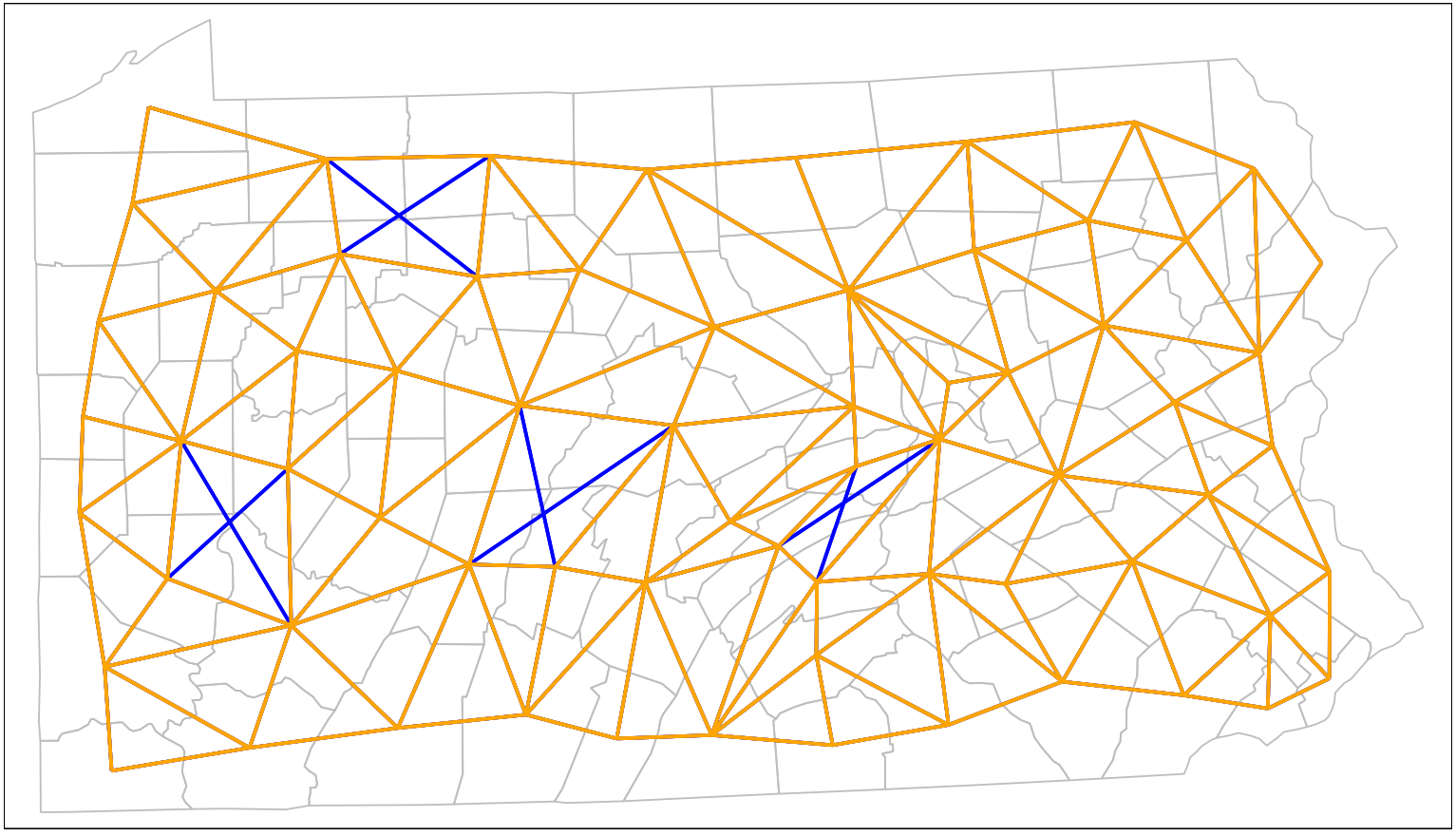
Queen adjacency: Regions are *adjacent* if their boundaries share at least one point in common.

Rook adjacency: Regions are *adjacent* if their boundaries share more than one point in common.

Adjacency Map for Pennsylvania Counties

Rook adjacency: orange edges

Queen adjacency: orange and blue edges taken together



Lagged Means

Lagged means refers to the average of all adjacent counties.

Let δ_i denote the set of neighbors of county i.

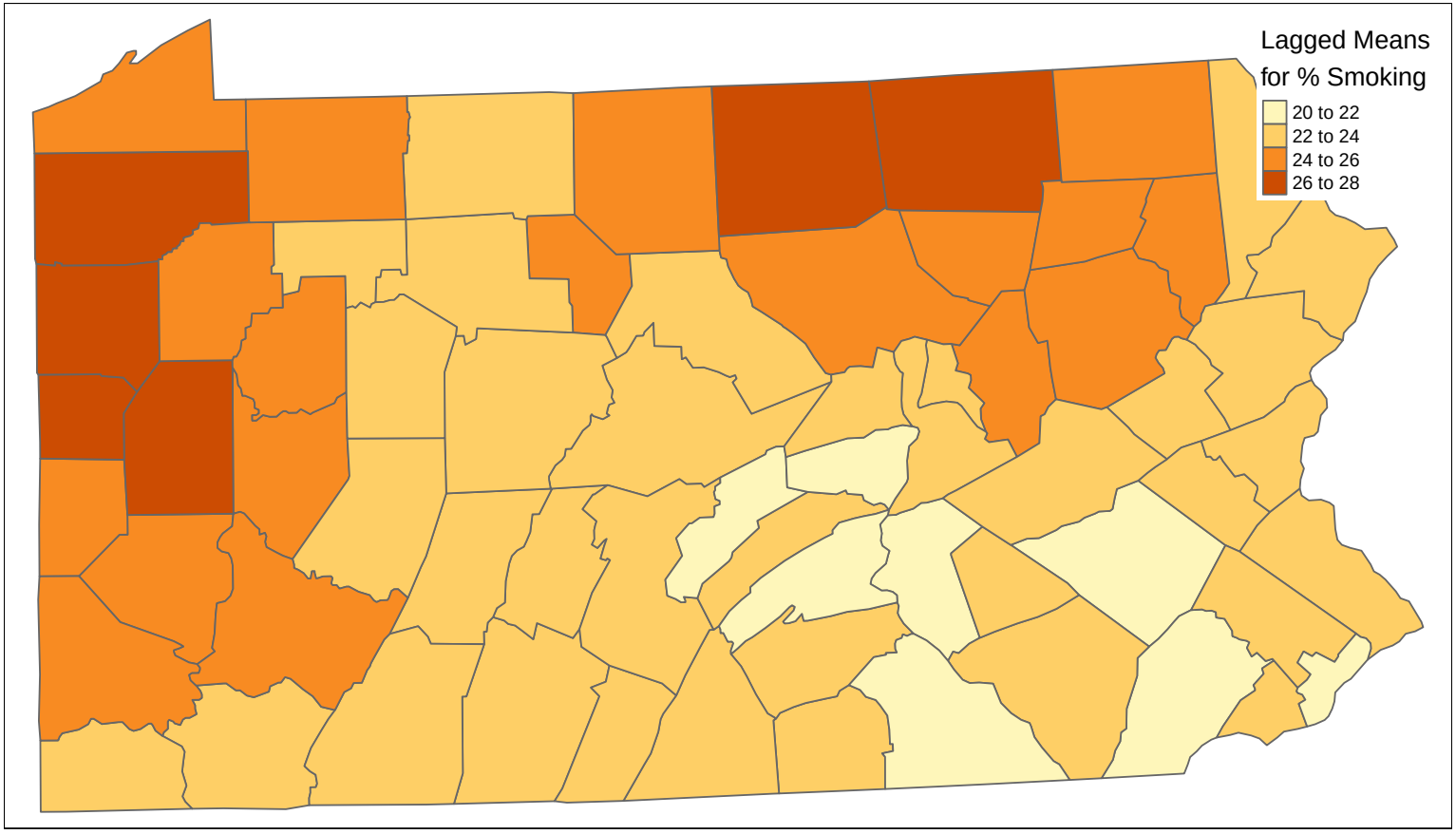
Let $|\delta_i|$ denote the number of elements in the set δ_i.

The lagged mean $\tilde{z}_i$ is defined as

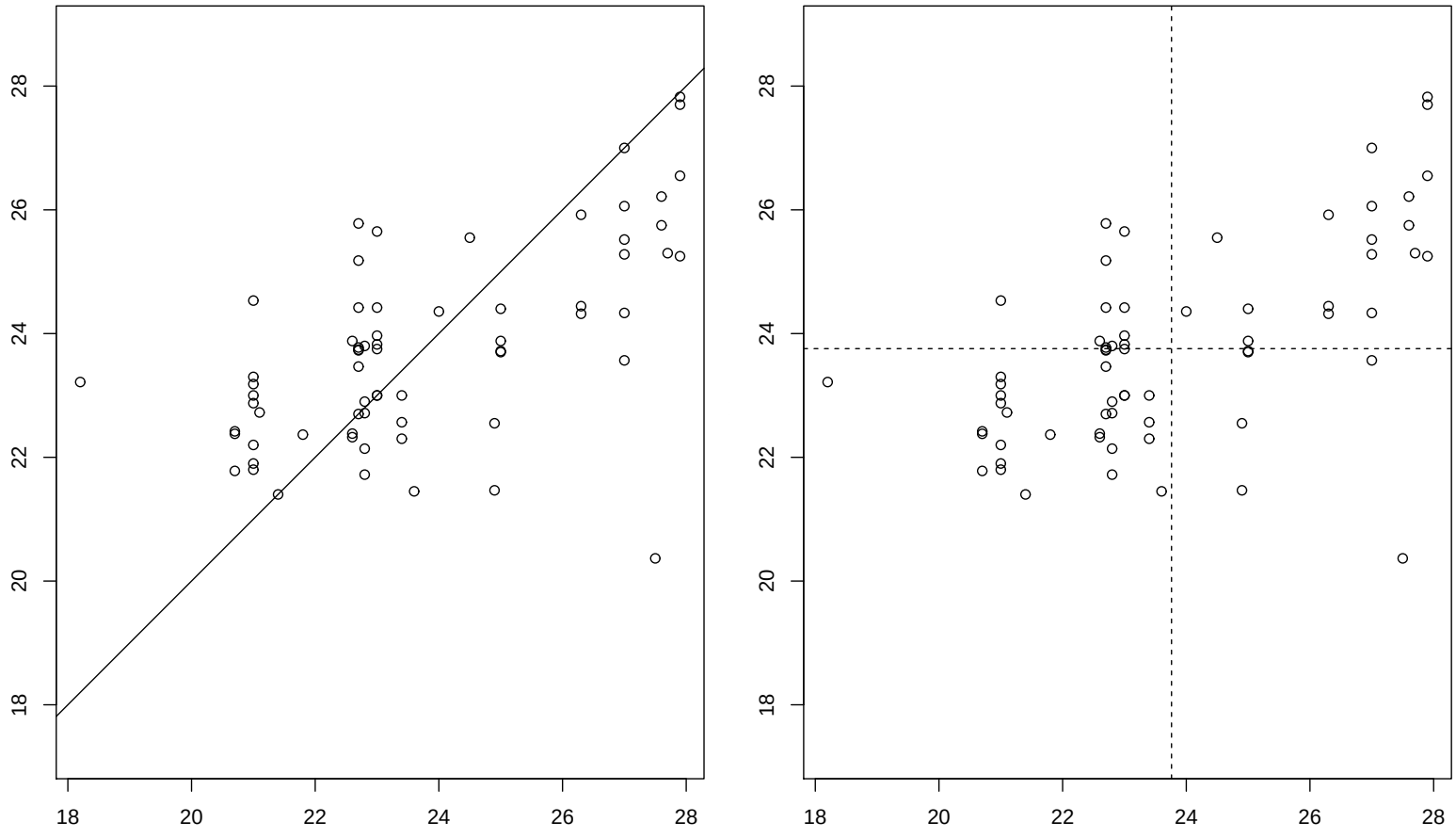
$$\tilde{z}_i = \frac{1}{|\delta_i|} \sum_{j \in \delta_i} z_j$$

[Note typo in equation (7.3) in book.]

A choropleth map of lagged means can sometimes give more information about spatial autocorrelation than a choropleth map of the actual outcomes.


Lagged Means
for % Smoking
20 to 22
22 to 24
24 to 26
26 to 28

Scatterplot of Actual vs. Lagged Smoking Rates



Section 7.4 Moran's I: An Index of Autocorrelation

Definition:

$$I = \frac{n}{\sum_{i=1}^{n}\sum_{j=1}^{n} w_{ij}} \frac{\sum_{i=1}^{n}\sum_{j=1}^{n} w_{ij}(z_i - \overline{z})(z_j - \overline{z})}{\sum_{i=1}^{n}(z_i - \overline{z})^2}$$

where there are many choices for w_{ij}, including

$$w_{ij} = \begin{cases} 1 & \text{if regions } i \text{ and } j \text{ are adjacent} \\ 0 & \text{if otherwise} \end{cases}$$

or

$$w_{ij} = \begin{cases} \frac{1}{|\delta_i|} & \text{if regions } i \text{ and } j \text{ are adjacent} \\ 0 & \text{if otherwise} \end{cases}$$

This latter choice guarantees that each row of the matrix W with entries w_{ij} sums to one. The sum of *all* entries in W is therefore 1.

Simple Example

If the rows of W sum to 1 then

$$I = \frac{\sum_{i=1}^{n}\sum_{j=1}^{n} w_{ij}(z_i - \overline{z})(z_j - \overline{z})}{\sum_{i=1}^{n}(z_i - \overline{z})^2} = \frac{\sum_{i=1}^{n}\sum_{j=1}^{n} w_{ij}\ q_i\ q_j}{\sum_{i=1}^{n} q_i^2} = \frac{\mathbf{q}^T\mathbf{W}\mathbf{q}}{\mathbf{q}^T\mathbf{q}}$$

where $q_i = (z_i - \overline{z})$ and $\mathbf{q} = [q_1, q_2, \ldots, q_n]^T$.

Using R to Get Moran's I

Bounds for Autocorrelation

The usual correlation must satisfy $-1 \leq R \leq 1$, but the range for I is more complicated.

The range for Moran's I depends on the weight matrix W.

The range for Moran's I is equal to the range of the eigenvalues of the matrix

$$(W + W^T)/2.$$

Testing for Autocorrelation

Under the null hypothesis of no spatial autocorrelation and assuming a normally distribute outcome z_i, the expected value of Moran's I is

$$E(I) = -\frac{1}{n-1}$$

and the estimate of the variance is

$$\hat{V}(I) = \frac{n^2 S_1 - n S_2 + 3 S_0^2}{(n-1)(n+1)^2 S_0^2}$$

where

$$\begin{aligned} S_0 &= \sum_{i=1}^{n} \sum_{j=1}^{n} w_{ij} \qquad [\text{remember } w_{ii} = 0] \\ S_1 &= \frac{1}{2} \sum_{i=1}^{n} \sum_{j=1}^{n} (w_{ij} + w_{ji})^2 \\ S_2 &= \sum_{k=1}^{n} \left(\sum_{j=1}^{n} w_{kj} + \sum_{i=1}^{n} w_{ik} \right)^2 \end{aligned}$$

Testing for Autocorrelation

The null hypothesis is H_0 : no autocorrelation.

The test statistic is

$$\frac{I - E(I)}{\sqrt{\hat{V}(I)}} \overset{\text{approx.}}{\sim} \mathrm{N}(0,1)$$

Note error in equations (7.7) and (7.8) in book; denominators should have a square root.

For the Pennsylvania smoking data, $n = 67$, $E(I) = -1/(n-1) = -0.01515$, and $\hat{V}(I) = 0.005998405$. The standardized test statistic is

$$\frac{0.404431265 - (-0.015151515)}{\sqrt{0.005998405}} = 5.417.$$

Using Simulation to Test Significance

The null hypothesis is H_0 : no autocorrelation. The alternative is H_1 : positive autocorrelation.

Step 1: Given fixed values for the outcomes $z_1,\ z_2,\ \ldots,\ z_n$ shuffle these among all the regions (i.e., randomly assign each of the numbers to each of the regions).

Step 2: Compute Moran's I for this particular shuffle.

Step 3: Go back to Step 1 and repeat until we've done M simulations (shuffle + compute Moran's I), keeping track of I each time.

Step 4: Count how many of the simulated values of I exceed our observed I. Call this m.

Step 5: Estimate of P-value is $\dfrac{m+1}{M+1}$.

Section 7.5: Spatial Autoregression

Simultaneous Autoregressive (SAR) Model

$$z_i = \mu + \sum_{j=1}^{n} b_{ij}(z_j - \mu) + \epsilon_i$$

where $\epsilon_i \overset{\text{indep.}}{\sim} \mathrm{N}(0, \sigma_i^2)$ and usually $b_{ij} = \lambda w_{ij}$. Often $\sigma_i^2 = \sigma^2$ for all i.

Parameters to estimate: $\mu,\ \sigma^2,\ \lambda$.

Spatial Autoregression

Conditional Autoregressive (CAR) Model

$$z_i \left| \{ z_j : j \neq i \} \right. \sim \mathrm{N} \left(\mu + \sum_{j=1}^{n} c_{ij}(z_i - \mu), \tau_i^2 \right)$$

where usually $c_{ij} = \lambda w_{ij}$. Often $\tau_i^2 = \tau^2$ for all i.

Parameters to estimate: μ, τ, λ.

Likelihood based statistical methods use the joint (not the conditional) distribution of the observed random variables. Getting from the conditional specification given above to the likelihood involves some deep theory.

Bayesian Statistics

We can add predictor variables to the SAR and CAR models, but these models become intractable. It's very hard to estimate the parameters, and standard errors are based on some rough approximations.

Complex models can often be fit in the Bayesian paradigm that are intractable in the classical framework.

The Bayesian approach to statistics expresses uncertainty in parameters in probabilistic terms.

Bayesian model fitting for complex models involves simulation.

We'll cover Bayesian statistics next.