Introduction - Some Basics

- Statistical analysis of data starts with assumptions. A common one, almost universally adopted in introductory statistics courses, is that data are realizations of independent and identically distributed random variables. Violation of this assumption is not uncommon but dealt with superficially, if at all, in most introductory courses.

- Data collected in a spatial domain of some type frequently do not meet this assumption; the observations are spatially autocorrelated (or correlated); a result nicely summarized in Tobler's First Law of Geography: *everything is related to everything else, but near things are more related than distant things.*

- Things can be near in both space, time, or both. The methods of time series analysis have analogs in spatial analysis but there are some fundamental differences between the 2, which we touch on later.

Types of Spatial Data

- We follow the classification of Noel Cressie's now classic 1993 text, *Statistics for Spatial Data.*

  - The General Spatial Model: Let $\mathbf{s} \in \mathcal{R}^d$ represent a spatial location in a d-dimensional Euclidean space, $d = 1, 2, 3, \cdots$. Associated with each location is an attribute Z (e.g. concentration of a toxic metal, number of deaths, presence/absence of some characteristic of interest, etc.). The potential value of an attribute at location $\mathbf{s}$ is a random quantity, $Z(\mathbf{s})$. We let $\mathbf{s}$ vary over a spatial domain D generating a random process
  $$\left\{ Z(\mathbf{s}) : \mathbf{s} \in D \subset \mathcal{R}^d \right\}.$$
  Different types of spatial data are determined by the nature of D.

  - Geostatistical Data: The spatial index $\mathbf{s}$ varies continuously in D, a fixed subset of $\mathcal{R}^d$. We do not observe all possible values of $Z(\mathbf{s})$ but only the sample $\{Z(\mathbf{s}_1), Z(\mathbf{s}_2), \cdots, Z(\mathbf{s}_n)\}$ at known spatial locations $\{\mathbf{s}_1, \mathbf{s}_2, \cdots, \mathbf{s}_n\}$. Note that the attribute Z can be continuous or discrete or even categorical. It is D that is continuous, not necessarily Z.

  - Lattice (Regional) Data: The spatial domain D is considered to be a fixed and countable set of sites at which data are observed. Site coordinates ($\mathbf{s}$) typically index areas or regions and not a particular point in space. Coordinates are chosen in some appropriate manner, e.g. the center of a quadrat, the centroid of an irregularly shaped region, etc. Thus, for example, the "value" of an attribute at a site $\mathbf{s}_i$ represents an aggregation (e.g. an average or total) of values of the attribute in an area. For example, the number of deaths in a county, the number of plants in a quadrat, etc. Lattices may be regular or irregular. It will also often be the case that one has information for every site in the lattice (e.g. the numbers of deaths in all counties in a state), as opposed to geostatistical data in which one has what could be considered a partial realization of the spatial process.

- Point Pattern or Point Process: Both geostatistical data and lattice data share a fixed, non-random spatial domain D. Point patterns are distinguished by the fact that the domain D is random. There are many scientific examples in which it makes sense to model spatial locations themselves as random events. An unmarked point process is one in which the spatial locations themselves are of primary interest. It will sometimes be the case that auxiliary information Z is observed at each point; a marked point process.
- Examples of all of the different spatial processes will be given in class.

Spatial Autocorrelation: Concept and Elementary Measures

- The material below relies a lot on the material in the text *Statistical Methods for Spatial Data Analysis* by Schabenberger and Gotway.

- Autocorrelation refers to correlation of a variable with itself, i.e. in the spatial context it refers to the correlation between $Z(\mathbf{s}_i)$ and $Z(\mathbf{s}_j)$. In general, spatially distributed variables tend to be positively correlated in that values of a variable measured at locations close in space should tend to be similar. If the spatial variable is not autocorrelated then knowing 2 locations are close to one another provides no information about their values. Positive correlation will manifest itself as systematic spatial variation in the values of the variable. It is possible (at least theoretically) to have negative correlation but this is somewhat rare.

- The General Cross-Product Statistic (M_2): The statistic is defined as:

$$M_2 = \sum_{i=1}^{n}\sum_{j=1}^{n} w_{ij}u_{ij}$$

  where w_{ij} is a measure of spatial proximity of locations $\mathbf{s}_i$ and $\mathbf{s}_j$ and u_{ij} is a measure of proximity of $Z(\mathbf{s}_i)$ and $Z(\mathbf{s}_j)$. Measures of proximity are generally determined by the goals of the study. A common measure of spatial proximity is Euclidean distance between 2 points:

$$w_{ij} = \|\mathbf{s}_i - \mathbf{s}_j\|.$$

  Another common measure is binary:

$$w_{ij} = \begin{cases} 1 & \text{if sites } i \text{ and } j \text{ are connected} \\ 0 & \text{if sites } i \text{ and } j \text{ are not connected} \end{cases}$$

  There are a number of these depending on how one defines “connected”. Three commonly used spatial contiguity definitions are the rook, bishop, and queen measures. These work well for lattice data. For non-lattice data other measures are possible. See the figure below (this one is from Schabenberger and Gotway).

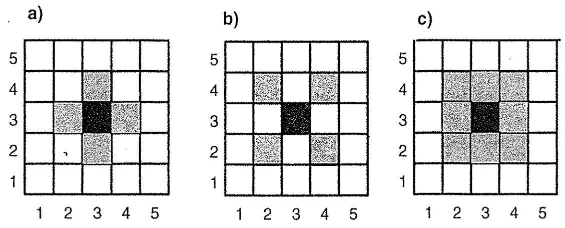


Figure 1.9 *Possible definitions of spatial connectedness (contiguity) for a regular lattice. Edges abut in the rook definition (a), corners touch in the bishop definition (b), edges and corners touch in the queen definition (c). Adapted from Figure 9.26 in Schabenberger and Pierce (2002).*

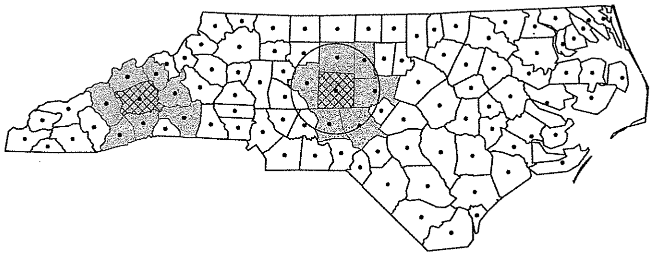


Figure 1.10 *Two definitions of spatial connectedness (contiguity) for an irregular lattice. The figure shows counties of North Carolina and two possible neighborhood definitions. The left neighborhood shows counties with borders adjoining that of the central, cross-hatched county. The neighborhood in the center of the map depicts counties whose centroid is within 35 miles of the central cross-hatched county.*

Two often used measures of distance between 2 Z values is the absolute value of the difference:

$$u_{ij} = |Z(\mathbf{s}_i) - Z(\mathbf{s}_j)|$$

or the squared deviation

$$u_{ij} = (Z(\mathbf{s}_i) - Z(\mathbf{s}_j))^2 .$$

A reminder: In 2 dimensions the Euclidean distance between 2 points $\mathbf{s}_i = (x_i, y_i)$ *and* $\mathbf{s}_j = (x_j, y_j)$ *is*

$$w_{ij} = \|\mathbf{s}_i - \mathbf{s}_j\| = \sqrt{(x_i - x_j)^2 + (y_i - y_j)^2}.$$

- The proximity measures can be summarized in matrices $\mathbf{W}$ and $\mathbf{U}$ in which case

$$M_2 = \sum_{i=1}^{n} \mathbf{u}_i^{'} \mathbf{w}_i$$

where $\mathbf{u}_i^{'}$ and $\mathbf{w}_i^{'}$ are the ith rows of $\mathbf{W}$ and $\mathbf{U}$, respectively. This can be done easily in Matlab, R, or Splus by summing the diagonal elements of $\mathbf{UW}^{'}$.

- Example: The table below shows a hypothetical data set. There are 9 spatial locations in lattice form that can take on values of 0 or 1. Note that intuitively it would appear that there is systematic variation in the Z-values.

Locations			Z-values		
1	2	3	0	0	0
4	5	6	0	1	1
7	8	9	0	1	1

The proximity matrices are

$$\mathbf{W} = \begin{pmatrix} 0 & 1 & 0 & 1 & 0 & 0 & 0 & 0 & 0 \\ 1 & 0 & 1 & 0 & 1 & 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 & 1 & 0 & 0 & 0 \\ 1 & 0 & 0 & 0 & 1 & 0 & 1 & 0 & 0 \\ 0 & 1 & 0 & 1 & 0 & 1 & 0 & 1 & 0 \\ 0 & 0 & 1 & 0 & 1 & 0 & 0 & 0 & 1 \\ 0 & 0 & 0 & 1 & 0 & 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 & 1 & 0 & 1 & 0 & 1 \\ 0 & 0 & 0 & 0 & 0 & 1 & 0 & 1 & 0 \end{pmatrix}$$

and, using either the absolute deviation or squared deviation measures above

$$\mathbf{U} = \begin{pmatrix} 0 & 0 & 0 & 0 & 1 & 1 & 0 & 1 & 1 \\ 0 & 0 & 0 & 0 & 1 & 1 & 0 & 1 & 1 \\ 0 & 0 & 0 & 0 & 1 & 1 & 0 & 1 & 1 \\ 0 & 0 & 0 & 0 & 1 & 1 & 0 & 1 & 1 \\ 1 & 1 & 1 & 1 & 0 & 0 & 1 & 0 & 0 \\ 1 & 1 & 1 & 1 & 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 & 1 & 0 & 1 & 1 \\ 1 & 1 & 1 & 1 & 0 & 0 & 1 & 0 & 0 \\ 1 & 1 & 1 & 1 & 0 & 0 & 1 & 0 & 0 \end{pmatrix}$$

The measure of spatial proximity is the binary contiguity measure based on the rook measure. Note for example that the "distance" between lattice members 1 and 5 is $w_{15} = 0$. We have $M_2 = 8$. One way to assess whether or not this value is unusual is with a Monte Carlo (or permutation) test. If the Z values are not spatially correlated then the assignment of 0's and 1's to the 9 lattice elements would be random. How unusual is a random assignment or arrangement of 5 0's and 4 1's seen in the table? We can carry out a Monte Carlo procedure by randomly rearranging (permuting) the rows and columns of $\mathbf{U}$, computing the value of M_2 for each possible permutation, and checking to see how unusual our observed value of $M_2 = 8$ is by comparing it to the null distribution (distribution under an assumption of no spatial correlation) of the M_2 values. There are $9! = 362880$ possible $\mathbf{U}$ matrices and we do not have time to list them all. But we can sample from the possibilities, compute M_2 for each sample, and generate an approximation to the null distribution of M_2. Below is some simple R code to do that. I entered the Z values as a vector z. I entered the $\mathbf{W}$ matrix given above (this was the most tedious part). I used the `dist` function to create the distance matrix $\mathbf{U}$ for 999 random permutations of z. I then computed 999 M_2 values for the resulting distance matrices. I added in the observed value of $M_2 = 8$ to give me 1000 M_2 values.

```
z=c(0,0,0,0,1,1,0,1,1)
M2=rep(0,1000)
for(i in 1:999){
+ U1=as.matrix(dist(sample(z),diag=T,upper=T))
+ M2[i]=sum(diag(U1%*%t(W)))}
table(M2)
M2
  8  10  12  14  16  18  20  24
100  84 291 296  94  92  35   8
```

Even though the observed spatial arrangement of Z values appears decidedly nonrandom, arrangements yielding a value of $M_2 = 8$ would occur by chance about 1 out of every 10 times. We could look at a histogram of the values, also. We will have more to say about Monte Carlo procedures later as they are a useful approach to inference in spatial data analysis.

- M_2 itself is not used much to summarize spatial autocorrelation, but it forms the basis for the development of other statistics that are commonly used. We will look at a couple of those below.

- Useful Measures - We will briefly examine 3 summary measures: Join-Count Statistics, Moran's I, and Geary's c.

  - Join-Count Statistics: Typically these statistics are used to analyze lattice (regional) data but their applicability extends further than that. We define $Z(\mathbf{s}_i)$ to be either 1 (if some specified condition exists at $\mathbf{s}_i$) or 0 (if the specified condition does not exist). We define

    $$w_{ij} = \begin{cases} 1 & \text{if sites } i \text{ and } j \text{ are connected} \\ 0 & \text{if sites } i \text{ and } j \text{ are not connected} \end{cases}$$

    Traditionally the Z's are actually color-coded with black (B) specifying $Z(\mathbf{s}_i) = 1$ and white (W) specifying $Z(\mathbf{s}_i) = 0$. A map of the values will present as a mosaic of black and white regions. The definition of *connected* can impact the results and some thought should be given to this matter rather than just blindly taking a standard measure off the shelf. The 3 common definitions (rook, bishop, and queen) we saw above are commonly used. Interest centers around whether or not the colors tend to cluster and this is determined by counting the number of black-black (BB) joins, black-white (BW) joins, or white-white (WW) joins, and comparing the observed counts to those expected under an assumption of no spatial correlation. The number of BB joins is

    $$BB = \frac{1}{2}\sum_{i=1}^{n}\sum_{j=1}^{n} w_{ij} Z(\mathbf{s}_i) Z(\mathbf{s}_j).$$

    This is $(1/2)M_2$ with $u_{ij} = Z(\mathbf{s}_i) Z(\mathbf{s}_j)$. The number of BW joins is

    $$BW = \frac{1}{2}\sum_{i=1}^{n}\sum_{j=1}^{n} w_{ij} \left(Z(\mathbf{s}_i) - Z(\mathbf{s}_j)\right)^2.$$

    This is $1/2(M_2)$ with $u_{ij} = \left(Z(\mathbf{s}_i) - Z(\mathbf{s}_j)\right)^2$. Upton and Fingleton (1985) comment that there is "little to choose between the 3 statistics" but note that BW is "marginally more informative" whatever that means. Most sources in the literature and most practitioners appear to use BB. We will discuss BB and BW only.

    The expected value and variances depend on the sampling plan. Under binomial sampling it is assumed that $Z(\mathbf{s}_i) \sim Bin(n, \pi)$. Formulas for $E[BB]$, $V(BB)$, $E[BW]$, and $V(BW)$ are given on page 20 of Schabenberger and Gotway. However, their use obviously requires knowledge of π. Letting n_1 equal to the number of black lattice elements we can use n_1/n as an estimate of π. It appears that most practitioners (and software packages) proceed conditional on n_1, so-called non-free sampling. Under this assumption and letting

    $$P_{(k)} = \frac{(n_1)(n_1 - 1)\cdots(n_1 - k + 1)}{(n)(n-1)\cdots(n-k+1)}$$

it can be shown that

$$E\left[BB\right]=\frac{1}{2}P_{(2)}\sum_{i=1}^{n}\sum_{j=1}^{n}w_{ij}$$

and

$$V\left(BB\right)=\frac{S_1}{4}\left[P_{(2)}-2P_{(3)}+P_{(4)}\right]+\frac{S_2}{4}\left[P_{(3)}-P_{(4)}\right]+\frac{w_{..}^2}{4}P_{(4)}-\frac{1}{4}\left[w_{..}P_{(2)}\right]^2$$

where

$$S_1=\frac{1}{2}\sum_{i=1}^{n}\sum_{j=1}^{n}\left(w_{ij}+w_{ji}\right)^2$$

and

$$S_2=\sum_{i=1}^{n}\left\{\sum_{j=1}^{n}w_{ij}+\sum_{j=1}^{n}w_{ji}\right\}^2$$

and $w_{..}=\sum_i\sum_j w_{ij}$. The `spdep` package in R has the capability of estimating these statistics.

A randomization (actually permutation) test similar to that discussed above for M_2 can be used for inference. A large sample test is also available based on assumed normality of BB and BW in which case the test statistics are

$$BB: \qquad \frac{BB-E\left[BB\right]}{\sqrt{V\left(BB\right)}}$$

or

$$BW: \qquad \frac{BW-E\left[BW\right]}{\sqrt{V\left(BW\right)}}.$$

A null of no spatial autocorrelation can be tested against the alternative of spatial autocorrelation using a two-sided test. If the investigator has an a priori idea of direction (positive versus negative spatial correlation) then appropriate one-sided tests may be conducted. For example, if we see more BB joins than expected under a null hypothesis of no spatial autocorrelation then that suggests black lattice areas tend to cluster together (positive spatial autocorrelation). If we see more BW joins than expected under the null then that suggests a checkerboard pattern (negative correlation).

Example:

The data are counts of Atriplex hymenelytra in 256 2.5m × 2.5m *quadrats arranged in a* 16 × 16 *grid in Death Valley. For this example we code any quadrat containing 1 or more shrubs as a* 1 *(Black) and quadrats containing no shrubs as a* 0 *(White). We use the* `joincount.mc` *function in the* `spdep` *package spdep to carry out a Monte Carlo*

I created a 256 × 3 *matrix* `atriplx.dat` *containing the data. The first* 2 *columns contain the spatial coordinates and the third column contain the BW coding. Here are the first* 5 *rows of the data matrix.*

```
atriplx.dat[1:5,]
   x  y  z
1  1  1  0
2  1  2  0
3  1  3  0
4  1  4  0
5  1  5  1
```

The spatial coordinates (i, j) refer to the ith row and jth column of the lattice. We can suppose that the question of interest is whether or not there is any evidence that the plants tend to cluster together.

There are two functions in `spdep` *that are needed to get the data into the correct form for analysis. The first is* `dnearneigh` *which*

```
identifies neighbours of region points by Euclidean distance between lower
(greater than) and upper (less than or equal to) bounds, or with longlat = TRUE,
by Great Circle distance in kilometers.
```

We will work with the rook's measure.

```
require(spdep)
atriplex<-read.table("c:../mydoc.D/Courses/STAT534/atrplx.dat",header=F)
atriplex.nb<-dnearneigh(as.matrix(atriplex[,1:2]),d1=0,d2=1)
atriplex.nb
Neighbour list object:
Number of regions: 256
Number of nonzero links: 960
Percentage nonzero weights: 1.464844
Average number of links: 3.75
```

Note the use of the `as.matrix` *function;* `atriplex` *is a data frame and* `atriplex[,1:2]` *is not recognized as a numeric object by* `dnearneigh`*. The* `as.matrix` *function corrects that problem.*

The output is a bit cryptic but the function is computing the number of neighbors associated with each grid cell. We have a square region and for these data `d1=0` *and* `d2=1` *define neighbors in terms of the rook's measure. We can see the number of neighbors associated with each grid cell.*

```
card(atriplex.nb)
  [1] 2 3 3 3 3 3 3 3 3 3 3 3 3 3 3 2 3 4 4 4 4 4 4 4 4 4 4 4 4 3 3 4 4 4 4
 [38] 4 4 4 4 4 4 4 4 4 4 3 3 4 4 4 4 4 4 4 4 4 4 4 4 4 4 3 3 4 4 4 4 4 4 4 4 4
 [75] 4 4 4 4 4 3 3 4 4 4 4 4 4 4 4 4 4 4 4 4 4 3 3 4 4 4 4 4 4 4 4 4 4 4 4 4 4
[112] 3 3 4 4 4 4 4 4 4 4 4 4 4 4 4 4 3 3 4 4 4 4 4 4 4 4 4 4 4 4 4 4 3 3 4 4 4
[149] 4 4 4 4 4 4 4 4 4 4 4 3 3 4 4 4 4 4 4 4 4 4 4 4 4 4 4 3 3 4 4 4 4 4 4 4 4
[186] 4 4 4 4 4 4 3 3 4 4 4 4 4 4 4 4 4 4 4 4 4 4 3 3 4 4 4 4 4 4 4 4 4 4 4 4 4
[223] 4 3 3 4 4 4 4 4 4 4 4 4 4 4 4 4 4 3 2 3 3 3 3 3 3 3 3 3 3 3 3 3 3 2
```

Now we need to create a special list object using another function `nb2listw` *which*

```
supplements a neighbours list with spatial weights for the chosen coding scheme.
```

The results are

```
atriplex.listw<-nb2listw(atriplex.nb,style="B")
atriplex.listw
Characteristics of weights list object:
Neighbour list object:
Number of regions: 256
Number of nonzero links: 960
Percentage nonzero weights: 1.464844
Average number of links: 3.75

Weights style: B
Weights constants summary:
    n    nn  S0   S1    S2
B 256 65536 960 1920 14624
```

The style of B specifies basic binary coding which is what we want for the lattice coordinates we have here. $nn = 65536 = 256(256)$ *and* S_0, S_1, S_2 *are as defined above.*

We can use the large sample approach from above if we want and in this case, given the sample size, that is probably fine but it makes more sense to use a Monte Carlo approach. The `joincount.mc` *function will carry out such a test for us on the BB data. It works by randomly scrambling the black and white colors and computing the number of BB joins each time. The resulting permutation distribution, computed under an assumption of no spatial autocorrelation, of the test statistic is used to asses how unusual the observed number of BB joins is.*

```
joincount.mc(as.factor(atriplex[,3]),atriplex.listw,nsim=999)

## WW joins

        Monte-Carlo simulation of join-count statistic

data:  as.factor(atriplex[, 3])
weights: atriplex.listw
number of simulations + 1: 1000

Join-count statistic for 0 = 268, rank of observed statistic = 583.5,
p-value = 0.4165
alternative hypothesis: greater
sample estimates:
    mean of simulation variance of simulation
             266.89790               24.22403
```

```
## BB joins


        Monte-Carlo simulation of join-count statistic

data:  as.factor(atriplex[, 3])
weights: atriplex.listw
number of simulations + 1: 1000

Join-count statistic for 1 = 39, rank of observed statistic = 973,
p-value = 0.027
alternative hypothesis: greater
sample estimates:
    mean of simulation variance of simulation
              30.77477               18.25283
```

By default 999 different permutations are assessed and the inclusion of the observed value gives us 1000 values of the test statistic. We see same color results; WW and BB joins. The alternative hypothesis is `greater` *because we are interested in the evidence for postiive correlation. The result are equivocal. Based on the WW joins we see no evidence of any clustering of W grid cells ($p = 0.42$) but the BB results do suggest some evidence of clustering ($p = 0.03$). It is important to keep in mind that the alternative is greater than, i.e. positive correlation. For the WW results the lack of evidence against positive correlation could result from a random distribution of plants or because the plants were negatively correlated with one another.*

The `joincount.test` *function will provide us the large sample test statistics for tests based on asymptotic normality. The output is in the same format and the results are very similar to the Monte Carlo results. We can also get large sample BW results using* `joincount.multi`

```
joincount.multi(as.factor(atriplex[,4]),atriplex.listw)
     Joincount Expected Variance z-value
0:0    268.000  266.838   22.971  0.2424
1:1     39.000   30.588   17.688  2.0001
1:0    173.000  182.574   70.668 -1.1388
Jtot   173.000  182.574   70.668 -1.1388
```

If we have an upper tail alternative for BB we need the lower tail alternative for BW and the approximate p value is

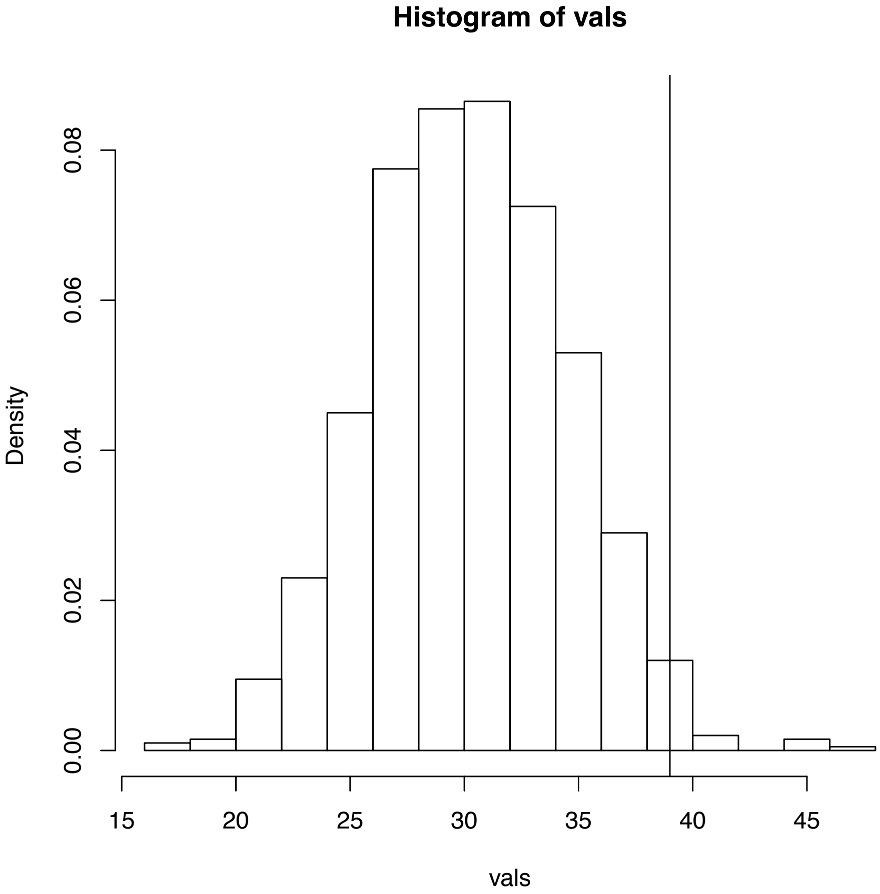
```
pnorm(-1.1388)
0.1273933
```

which is not much evidence against the null. Technically a continuity correction would help with the large sample statistics, but we will not worry about that here.

One disadvantage of the `joincount.mc` *function is that it only allows for one-sided testing. As long as the permutation distribution is symmetric an approximate two-*

sided p-value can be easily computed. But a symmetric distribution is not guaranteed. However, it is possible to access the 1000 *values of the test statistic.*

```
results<-joincount.mc(as.factor(atriplex[,4]),atriplex.listw,nsim=999)
# get values of BB test statistic
vals<-results[[2]]$res
hist(vals,prob=T)
abline(v=39)
```



In this case we do have symmetry so an approximate two-sided p-value of 0.054 *would seem reasonable. We will look at permutation tests in more detail later in the course. The* `joincount.test` *function allows for two-sided testing. And of course it is easy to generate two-sided p-values based on the output from* `joincount.multi`*.*

The texts that I have looked at suggest you only need to look at one of these measures but it is obvious from this example that there may not be consistency in the results. It might be a good idea to look at all 3 *and if there is no consistency among the results then that would suggest a good deal of care before drawing conclusions. I would not be willing to conclude that the plants in the above example are positively spatially correlated with one another.*

- Moran's I and Geary's c: These measures are more commonly used on continuous Z values. A key assumption is that the mean is constant, i.e. $E\left[Z\left(\mathbf{s}\right)\right] = \mu$. Moran's I is

$$I = \frac{n}{(n-1)S^2 w_{\cdot\cdot}} \sum_{i=1}^{n} \sum_{j=1}^{n} w_{ij} \left(Z\left(\mathbf{s}_i\right) - \overline{Z}\right)\left(Z\left(\mathbf{s}_j\right) - \overline{Z}\right)$$

where S^2 is the sample variance of the $Z\left(\mathbf{s}_i\right)$ values. Geary's c is defined to be

$$c = \frac{1}{2S^2 w_{\cdot\cdot}} \sum_{i=1}^{n} \sum_{j=1}^{n} w_{ij} \left(Z\left(\mathbf{s}_i\right) - Z\left(\mathbf{s}_j\right)\right)^2.$$

I is more sensitive to extreme values of $Z\left(\mathbf{s}_i\right)$ whereas c is more sensitive to differences between $Z\left(\mathbf{s}_i\right)$ and $Z\left(\mathbf{s}_j\right)$. Generally, they will provide much the same information about spatial autocorrelation. Upton and Fingleton (1985) note that tests based on Moran's I are "consistently more powerful than those based on c" citing the work of Cliff and Ord (1981).

The mean and variance of I based on a randomization approach to inference are

$$E\left[I\right] = -\frac{1}{n-1}$$

and

$$V\left(I\right) = \frac{n\left\{\left(n^2 - 3n + 3\right)S_1 - nS_2 + 3w_{\cdot\cdot}^2\right\} - k\left\{n(n-1)S_1 - 2nS_2 + 6w_{\cdot\cdot}^2\right\}}{(n-1)(n-2)(n-3)w_{\cdot\cdot}^2} - \frac{1}{(n-1)^2}$$

where

$$k = \frac{\sum_{i=1}^{n}\left(Z\left(\mathbf{s}_i\right) - \overline{Z}\right)^4}{n\sum_{i=1}^{n}\left(Z\left(\mathbf{s}_i\right) - \overline{Z}\right)^2}.$$

The mean for c is $E[c] = 1$.

Suitable assumptions yield large sample approximate normal distributions for both statistics. The mean for I and c are the same as that given above but the variance formulas are different. For example, assuming $Z\left(\mathbf{s}_i\right) \sim N\left(\mu, \sigma^2\right)$ and no spatial correlation we have

$$V\left(I\right) = \frac{n^2(n-1)S_1 - n(n-1)S_2 - 2w_{\cdot\cdot}^2}{(n+1)(n-1)^2 w_{\cdot\cdot}^2}.$$

If sites tend to be connected to other sites that have similar values of Z then we we expect $I > E[I]$ (positive correlation) and sites connected to other sites with dissimilar values of Z will tend to yield $I < E[I]$ (negative correlation). The direction of the inequalities are reversed for Geary's c.

Example: The drongo is an Asiatic bird. There are several species of drongos in fact. One study looked at spatial gradients (clines) in morphological characteristics, e.g. wing lengths. Analysis using the `spdep` *function* `moran.mc` *requires us to treat the data as if it were on a 1-dimensional lattice. Here are the data from 9 locations for one species.*

145.7	152.25	156.5	169.3	175.0	181.25	168.5	160.2	147.6

The following spatial coordinates were specified:

```
drongo.xy
      [,1] [,2]
 [1,]    1    1
 [2,]    2    1
 [3,]    3    1
 [4,]    4    1
 [5,]    5    1
 [6,]    6    1
 [7,]    7    1
 [8,]    8    1
 [9,]    9    1
```

The cline data were also entered:

```
cline
145.7 152.25 156.5 169.3 175.0 181.25 168.5 160.2 147.6
```

We need the neighbours and list objects again.

```
drongo.nb<-dnearneigh(drongo.xy,d1=0,d2=1)
drongo.list<-nb2listw(drongo.nb,style="B")
## Moran
moran.mc(cline,drongo.list,nsim=99)

        Monte-Carlo simulation of Moran's I

data:  cline
weights: drongo.list
number of simulations + 1: 100

statistic = 0.6027, observed rank = 98, p-value = 0.02
alternative hypothesis: greater
## Geary
geary.mc(cline,drongo.list,nsim=99)

        Monte-Carlo simulation of Geary C

data:  cline
weights: drongo.list
number of simulations + 1: 100

statistic = 0.27765, observed rank = 2, p-value = 0.02
alternative hypothesis: greater
```

It appears that we have evidence for a nonrandom spatial pattern in wing length in this species of drongos.

A couple of notes:

- *The style is again specified as B. If we had actual spatial coordinates we would have used a different style (more on that later).*
- *Above we saw that large values of I and small values of c were expected under positive correlation. But both default alternative directions are given as* `greater`*. However, note that under Moran's I the rank is* 98 *yielding a p-value of* 0.02 *(upper tail) and under Geary's c the rank is* 2 *which also yields a p-value of* 0.02 *and that looks like a lower-tail result. The function* `geary.mc` *is reversing direction. The same thing happens in the large sample testing versions where the numerator of the test statistic in Moran's I is $I - E(I)$ and in Geary's c it is $E(c) - c$.*

As just indicated there is a large sample approximation and a function `moran.test` *and* `geary.test` *but the sample size is not large enough to justify that approach here.*

- Spatial Trend or Spatial Correlation? We mentioned above that one key assumption for Moran and Geary is that of a constant spatial mean. In order for these statistics to be considered measures of spatial *correlation* we must assume that the observed values of Z are all realizations drawn from a distribution with constant mean and variance. That is, the $E\left[Z\left(\mathbf{s}_i\right)\right]$ does not depend on $\mathbf{s}$. Schabenberger and Gotway have an example where this assumption is violated, i.e. Moran's I detects spurious correlation. We have a 10×10 grid. Independent draws are made from a $N(\mu(x, uy), 1)$ distribution with

  $$\mu(x, y) = 1.4 + 0.1x + 0.2y + 0.002x^2$$

  These observations are stochastically independent of one another but the mean function results in a spatial trend surface over the grid, i.e. we do not have a constant mean. The results are

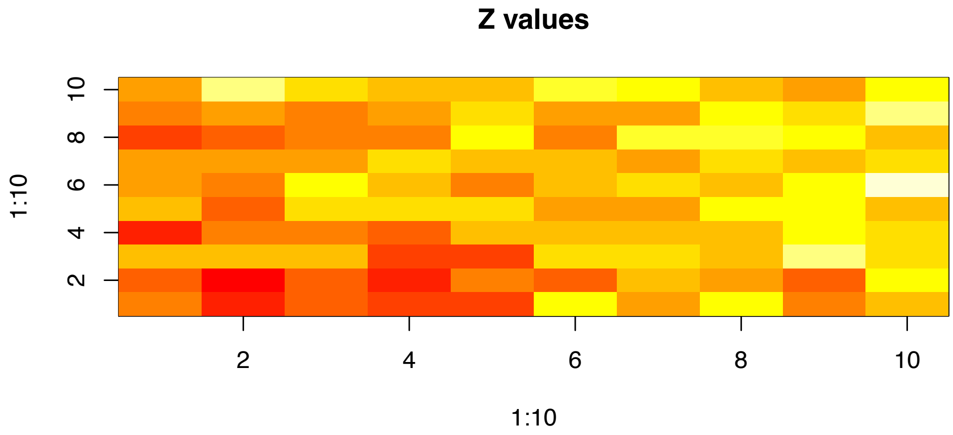
```
xy<-expand.grid(c(1:10),c(1:10))
xy<-cbind(xy[,2],xy[,1])
z<-rep(0,100)
for(i in 1:100){
z[i]<-rnorm(1,1.4+0.1*xy[i,1]+0.2*xy[i,2]+0.002*xy[i,1]^2,1)}
nb<-dnearneigh(as.matrix(xy),d1=0,d2=1)
list.nb<-nb2listw(nb,style="B")
moran.mc(z,list.nb,nsim=999)

        Monte-Carlo simulation of Moran I

data:  z
weights: list.nb
number of simulations + 1: 1000
```

```
statistic = 0.20722, observed rank = 999, p-value = 0.001
alternative hypothesis: greater

       Monte-Carlo simulation of Geary C

data:  z
weights: list.nb
number of simulations + 1: 1000

statistic = 0.75824, observed rank = 2, p-value = 0.002
alternative hypothesis: greater
```

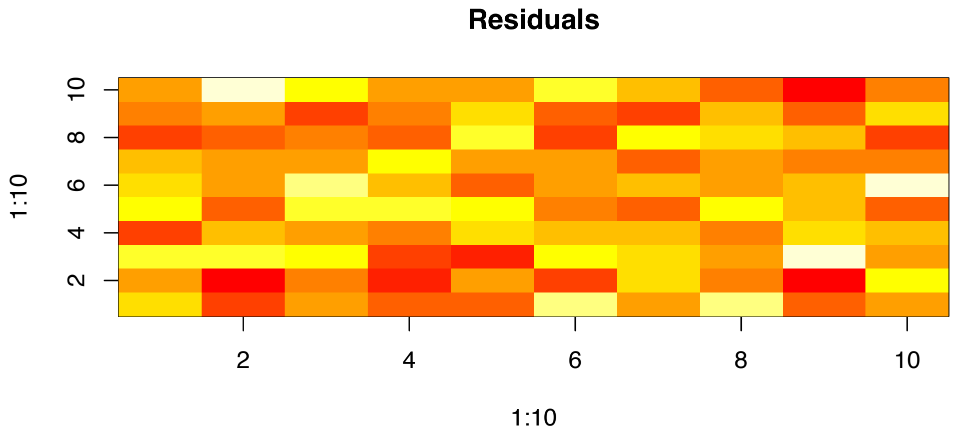
In both cases we see strong evidence of correlation. But the observations are known to be independent of one another. We detrend and check the residuals.

```
fit<-lm(z~xy[,1]+xy[,2]+I(xy[,1]^2))
resids<-residuals(fit)
moran.mc(resids,list.nb,nsim=999)
      Monte-Carlo simulation of Moran I

data:  resids
weights: list.nb
number of simulations + 1: 1000

statistic = -0.064753, observed rank = 229, p-value = 0.771
alternative hypothesis: greater
```

Now we see no evidence. Note that the assumption of a constant mean will be satisfied by the residuals from the least squares fit. A heat image plot of the Z values and the residuals is shown below. Note the trend in the top plot.



Z values



Residuals

- Suppose
  $$\left\{Z(\mathbf{s}) : \mathbf{s} \in D \subset \mathcal{R}^d\right\}$$
  is a spatial process with $E\left[Z\left(\mathbf{s}\right)\right] = \mu(\mathbf{s})$. The covariance between 2 values $Z\left(\mathbf{s}\right)$ and $Z\left(\mathbf{s}+\mathbf{h}\right)$ is defined to be
  $$C\left[\mathbf{s},\mathbf{h}\right] = \text{Cov}\left[Z\left(\mathbf{s}\right), Z\left(\mathbf{s}+\mathbf{h}\right)\right] = E\left[\left\{Z\left(\mathbf{s}\right) - \mu(\mathbf{s})\right\}\left\{Z\left(\mathbf{s}+\mathbf{h}\right) - \mu(\mathbf{s})\right\}\right].$$

  The correlation function is
  $$R(\mathbf{s},\mathbf{h}) = Cor\left(Z(\mathbf{s}), Z(\mathbf{s}+\mathbf{h})\right) = \frac{C(\mathbf{s},\mathbf{h})}{\sqrt{V(Z(\mathbf{s}))V(Z(\mathbf{s}+\mathbf{h}))}}$$

- Simplifying assumptions will be needed to proceed much further. We will have more to say about this later but for now we assume *second-order stationarity*:
  $$E\left[Z\left(\mathbf{s}\right)\right] = \mu, \quad \text{for all } \mathbf{s}$$
  $$\text{Cov}\left[Z\left(\mathbf{s}\right), Z\left(\mathbf{s}+\mathbf{h}\right)\right] = C(\mathbf{h}).$$

  This assumption implies that
  $$V\left(Z\left(\mathbf{s}\right)\right) = \text{Cov}\left[Z\left(\mathbf{s}\right), Z\left(\mathbf{s}+\mathbf{0}\right)\right] = C(\mathbf{0})$$
  and the correlation function is
  $$R(\mathbf{h}) = \frac{C(\mathbf{h})}{C(\mathbf{0})}.$$

  The implication of this assumption is that the mean, variance, covariance, and correlation do not depend on $\mathbf{s}$ and that the covariance and correlation between 2 variables at 2 locations $\mathbf{s}$ and $\mathbf{s}+\mathbf{h}$ depend only $\mathbf{h}$

- Another measure of spatial relatedness that arose out of work in geostatistics is the *semivariogram* which is defined to be
  $$\gamma\left(\mathbf{s},\mathbf{h}\right) = \frac{1}{2}V\left[Z\left(\mathbf{s}\right) - Z\left(\mathbf{s}+\mathbf{h}\right)\right] = \frac{1}{2}\left\{V\left[Z\left(\mathbf{s}\right)\right] + V\left[Z\left(\mathbf{s}+\mathbf{h}\right)\right] - 2\text{Cov}\left[Z\left(\mathbf{s}\right), Z\left(\mathbf{s}+\mathbf{h}\right)\right]\right\}.$$

  Under second-order stationarity we have
  $$\gamma\left(\mathbf{h}\right) = \frac{1}{2}\left[C(\mathbf{0}) + C(\mathbf{0}) - 2C(\mathbf{h})\right] = C(\mathbf{0}) - C(\mathbf{h}).$$

- Assuming positive spatial correlation we expect the covariance and correlation between $Z\left(\mathbf{s}\right)$ and $Z\left(\mathbf{s}+\mathbf{h}\right)$ to decrease as the lag distance $\|\mathbf{h}\|$ increases. As a consequence the semivariogram should increase with increasing distance, at least initially.
  - Sill: If $C\left(\mathbf{h}\right) \to 0$ as $\|\mathbf{h}\| \to \infty$ then $\gamma\left(\mathbf{h}\right) \to C\left(\mathbf{0}\right)$. $C\left(\mathbf{0}\right)$ is referred to as the *sill*.

- Range: The semivariogram may attain the sill or it may approach it asymptotically. The smallest value of $\|\mathbf{h}\|$ at which $\gamma(\mathbf{h}) = C(\mathbf{0})$ is called the *range*. If the semivariogram approaches the sill asymptotically then the range (often referred to as the *practical range* in this case) is generally defined to be the smallest value of $\|\mathbf{h}\|$ at which the semivariogram is equal to 95% of the sill. The range is interpreted as the distance at which values of Z become uncorrelated, or in the case of the practical range the distance at which spatial correlation becomes unimportant.
- Nugget: It is true that $\gamma(\mathbf{0}) = 0$ but there are cases where the semivariogram is not continuous at the origin. If $\gamma(\mathbf{h}) \to \theta_0$ as $\mathbf{h} \to \mathbf{0}$ then θ_0 is called the *nugget*. The nugget effect can arise from microscale processes that are not observable because they occur at values of $\|\mathbf{h}\|$ smaller than the spacings between the observations. It can also arise from measurement error.

We will have much more to say on all of these measures later in the course.

Why does it matter?

- Suppose $Z(1), Z(2), \cdots, Z(n)$ are independent and identically distributed as normal random variables with unknown mean μ and known variance σ^2. The best estimator of the unknown mean is the sample mean

$$\overline{Z} = (1/n) \sum_{i=1}^{n} Z(i)$$

and

$$V\left(\overline{Z}\right) = \frac{\sigma^2}{n}.$$

A 95% confidence interval for μ is

$$\overline{Z} \pm 1.96 \frac{\sigma}{\sqrt{n}}.$$

- Now suppose that $Z(i), i = 1, \cdots, n$ are observations on a one-dimensional transect and

$$\mathrm{Cov}\left(Z(i), Z(j)\right) = \sigma^2 \rho^{|i-j|}$$

where $-1 < \rho < 1$. Note that

$$\mathrm{Cor}\left(Z(i), Z(j)\right) = \rho^{|i-j|}.$$

It can be shown that, under these assumptions,

$$Vjn\left(\overline{Z}\right) = \frac{\sigma^2}{n}\left[1 + 2\left(\frac{\rho}{1-\rho}\right)\left(1 - \frac{1}{n}\right) - 2\left(\frac{\rho}{1-\rho}\right)^2 \left(\frac{1-\rho^{n-1}}{n}\right)\right].$$

Suppose $n = 10$ and $\rho = 0.26$. Then a 95% confidence interval computed under an assumption of independence would be

$$\overline{Z} \pm 1.96 \frac{\sigma}{\sqrt{10}}$$

whereas the correct formula would be

$$\overline{Z} \pm 2.485 \frac{\sigma}{\sqrt{10}}.$$

The width of the interval computed assuming independence is only about 79% the true width. Note that these data might not be considered all that strongly correlated and the correlation drops off quickly.

- Effective Sample Size: The effective sample size is the equivalent number of independent observations. For the above example we find the number of observations of independent random variables that would yield the same confidence interval as the correlated random variables. That is, we solve

$$1.96/\sqrt{n'} = 2.485/\sqrt{10}$$

for $n^{'}$. That value is $n^{'} = 6.22$. Thus, for this example, 6 independent observations yield approximately the same precision as 10 correlated observations. More generally, again for this example, we have

$$n^{'} = \frac{n}{\left[1 + 2\left(\frac{\rho}{1-\rho}\right)\left(1 - \frac{1}{n}\right) - 2\left(\frac{\rho}{1-\rho}\right)^2\left(\frac{1-\rho^{n-1}}{n}\right)\right]}.$$

For large n we have

$$n^{'} \approx \frac{n\left(1-\rho\right)}{1+\rho} < n$$

for $\rho > 0$ showing that positive correlation has an effect even with large sample sizes and even with small correlations. For example; with $\rho = 0.05$ the effective sample size is about 90% of the actual sample size.

```
n<-50:75
cbind(n,n*0.95/1.05)
        n
 [1,]  50 45.23810
 [2,]  51 46.14286
 [3,]  52 47.04762
 [4,]  53 47.95238
 [5,]  54 48.85714
 [6,]  55 49.76190
 [7,]  56 50.66667
 [8,]  57 51.57143
 [9,]  58 52.47619
[10,]  59 53.38095
[11,]  60 54.28571
[12,]  61 55.19048
[13,]  62 56.09524
[14,]  63 57.00000
[15,]  64 57.90476
[16,]  65 58.80952
[17,]  66 59.71429
[18,]  67 60.61905
[19,]  68 61.52381
[20,]  69 62.42857
[21,]  70 63.33333
[22,]  71 64.23810
[23,]  72 65.14286
[24,]  73 66.04762
[25,]  74 66.95238
[26,]  75 67.85714
```

- Linear Models with Spatially Correlated Errors: Suppose we have spatial data $Z\left(\mathbf{s}_i\right)$, $i = 1, \cdots, n$. We model these as a linear combination of (possibly) spatial explanatory variables, $x_j\left(\mathbf{s}_i\right), j = 1, \cdots, p-1$,

$$Z\left(\mathbf{s}_i\right) = \beta_0 + \beta_1 x_1\left(\mathbf{s}_i\right) + \cdots + \beta_{p-1} x_{p-1}\left(\mathbf{s}_i\right) + \epsilon\left(\mathbf{s}_i\right), i = 1, \cdots, n.$$

You have seen this model before under an assumption that $\epsilon\left(\mathbf{s}_i\right)$ are independent and identically (and normally) distributed random error terms, i.e.

$$\epsilon\left(\mathbf{s}_i\right) \sim N\left(0, \sigma^2\right).$$

In matrix notation we have $\boldsymbol{Z} = \boldsymbol{X\beta} + \boldsymbol{\epsilon}$ and the ordinary least squares (OLS) estimator of the parameter vector $\boldsymbol{\beta}$ is

$$\widehat{\boldsymbol{\beta}}_{\boldsymbol{OLS}} = \left(\boldsymbol{X}'\boldsymbol{X}\right)^{\mathbf{-1}}\boldsymbol{X}'\boldsymbol{Z}$$

and under an assumption of normally distributed errors this is also the maximum likelihood estimator (MLE). (Note: Even if the errors are correlated with one another the OLS estimator is still valid, although it would no longer be the MLE.)

In a spatial context the assumption of uncorrelated errors may not be valid. We generalize the above model by assuming that

$$\boldsymbol{V}\left(\boldsymbol{\epsilon}\right) = \boldsymbol{\Sigma}$$

where $\boldsymbol{\Sigma}$ is a (spatial) covariance matrix. If $\boldsymbol{\Sigma}$ is known then the generalized least squares estimator (GLS) is

$$\widehat{\boldsymbol{\beta}}_{\boldsymbol{GLS}} = \left(\boldsymbol{X}'\boldsymbol{\Sigma}^{\mathbf{-1}}\boldsymbol{X}\right)^{\mathbf{-1}}\boldsymbol{X}'\boldsymbol{\Sigma}^{\mathbf{-1}}\boldsymbol{Z}$$

Typically, the spatial covariance matrix is unknown and must be estimated. There are any number of ways to do this and we will explore some of them when we discuss spatial regression models later in the course. The resulting estimtors will then be referred to as estimated generalized least squares estimators ($EGLS$).