**Spatial Regression Models**

- The topic of this chapter is spatial regression models. Initially we assume a model with a linear mean function

$$\mathbf{Z}(\mathbf{s}) = \mathbf{X}(\mathbf{s})\boldsymbol{\beta} + \mathbf{e}(\mathbf{s})$$

where $\mathbf{e}(\mathbf{s}) \sim (\mathbf{0}, \boldsymbol{\Sigma}(\boldsymbol{\theta}))$. $\mathbf{Z}(\mathbf{s})$ is an $n \times 1$ vector of responses, $\mathbf{X}(\mathbf{s})$ is an $n \times p$ matrix of predictors (typically including a leading column of 1's for the intercept term), $\boldsymbol{\beta}$ is a $p \times 1$ vector of regression coefficients, and $\mathbf{e}(\mathbf{s})$ is an $n \times 1$ vector of error terms. Inference will require an assumption about the distribution of the error terms which will often be $\mathbf{e}(\mathbf{s}) \sim N(\mathbf{0}, \boldsymbol{\Sigma}(\boldsymbol{\theta}))$. We start with linear regression models with uncorrelated errors, i.e. we assume that $\boldsymbol{\Sigma}(\boldsymbol{\theta}) = \sigma^2\mathbf{I}$. We thus assume that all meaningful spatial structure can be captured via the mean function leaving only random noise for the error process.

- *Ordinary Least Squares*: We wish to estimate $\boldsymbol{\beta}$ and $\sigma^2$. The method of least squares, which requires no distributional assumptions, leads to the well-known estimators:

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

and

$$\widehat{\sigma}^2 = \frac{1}{n-p}\left(\mathbf{Z}(\mathbf{s}) - \mathbf{X}(\mathbf{s})\widehat{\boldsymbol{\beta}}_{ols}\right)' \left(\mathbf{Z}(\mathbf{s}) - \mathbf{X}(\mathbf{s})\widehat{\boldsymbol{\beta}}_{ols}\right).$$

Note that $\widehat{\sigma}^2$ is typically referred to as the Mean Squared Error ($MSE$) in most books on linear models. Under an assumption of independent and identically distributed normal error terms the maximum likelihood estimators are

$$\widehat{\boldsymbol{\beta}}_{ml} = \widehat{\boldsymbol{\beta}}_{ols}$$

and

$$\widehat{\sigma}^2_{ml} = \frac{n-p}{n}\widehat{\sigma}^2.$$

- The estimated mean response vector at a given covariate pattern is

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

$\mathbf{H}$ is called the *hat* matrix.

- The error terms are not observable. We observe instead the raw residuals

$$\widehat{\mathbf{e}}_{ols}(\mathbf{s}) = \left(\mathbf{Z}(\mathbf{s}) - \mathbf{X}(\mathbf{s})\widehat{\boldsymbol{\beta}}_{ols}\right) = (\mathbf{I} - \mathbf{H})\mathbf{Z}(\mathbf{s}) = \mathbf{M}\mathbf{Z}(\mathbf{s}).$$

- The following results can be established for the linear model with uncorrelated errors.

1. $E\left[\widehat{\boldsymbol{\beta}}_{ols}\right] = \boldsymbol{\beta}$
2. $E\left[\widehat{\mathbf{e}}_{ols}(\mathbf{s})\right] = \mathbf{0}$
3. $\mathrm{Var}\left[\widehat{\boldsymbol{\beta}}_{ols}\right] = \sigma^2\left(\mathbf{X}(\mathbf{s})' \mathbf{X}(\mathbf{s})\right)^{-1}$
4. $\mathrm{Var}[\widehat{\mathbf{e}}_{ols}(\mathbf{s})] = \sigma^2\mathbf{M}$
5. $\mathbf{1}'\widehat{\mathbf{e}}_{ols}(\mathbf{s}) = 0$
6. $\mathbf{X}(\mathbf{s})' \widehat{\mathbf{e}}_{ols}(\mathbf{s}) = \mathbf{0}$

- *Testing Hypotheses About* $\boldsymbol{\beta}$: We write the model as

$$Z(\mathbf{s}_i) = \beta_0 + \beta_1 X_1(\mathbf{s}_i) + \beta_2 X_2(\mathbf{s}_i) + \cdots + \beta_{p-1} X_{p-1}(\mathbf{s}_i) + e(\mathbf{s}_i), \quad i = 1, \cdots n.$$

This is called the *full* model. Denote the residual sum of squares for this model by $SSR_f$. There are $n - p$ degrees of freedom associated with the full model. We wish to test whether various combinations of the $\beta_j$'s are

equal to 0. The null model under these hypotheses will be restricted or reduced forms of the full model and we denote the residual sum of squares associated with these reduced models by $SSR_r$. There are $n-p+k$ degrees of freedom associated with the reduced model where $k$ is the number of coefficients specified to be equal to 0. The test statistic is

$$F = \frac{SSR_r - SSR_f}{SSR_f}\left(\frac{n-p}{k}\right).$$

If the null hypothesis is true then $F \sim F_{k,n-p}$.

If the null hypothesis involves a single coefficient $H_0 : \beta_j = 0$ versus $H_a : \beta_j \neq 0$ then the test statistic has the familiar form

$$F = \left[\frac{\widehat{\beta}_j}{SE\left(\widehat{\beta}_j\right)}\right]^2$$

where $SE\left(\widehat{\beta}_j\right)$ is the standard error. Under the null hypothesis $F \sim F_{1,n-p}$. It is customary to use the test statistic

$$t = \text{sign}\left(\widehat{\beta}_j\right)\left|\frac{\widehat{\beta}_j}{SE\left(\widehat{\beta}_j\right)}\right|$$

which has a null distribution of $t_{n-p}$.

- *Residuals and Diagnostics*: The raw residuals are $\widehat{\mathbf{e}}\left(\mathbf{s}\right)_{ols} = \left(\mathbf{I}-\mathbf{H}\right)\mathbf{Z}\left(\mathbf{s}\right)$. It is customary to standardize the raw residuals and there are a number of ways to do this.

  1. *Semistudentized residual*:
  $$r_i^* = \frac{\widehat{e}_{ols}\left(\mathbf{s}_i\right)}{\widehat{\sigma}}.$$

  2. *Studentized residual*: The problem with $r_i^*$ is that $\widehat{\sigma}^2$ is an estimate of the variance of the error terms in the model not the residuals. In fact the residuals have different variances:
  $$\text{Var}\left(\widehat{e}_{ols}\left(\mathbf{s}_i\right)\right) = \sigma^2\left(1-h_{ii}\right)$$
  where $h_{ii}$ denotes the $i$th diagonal element of $\mathbf{H}$. The studentized residual is
  $$r_i = \frac{\widehat{e}_{ols}\left(\mathbf{s}_i\right)}{\widehat{\sigma}\sqrt{1-h_{ii}}}.$$

  3. *Externally studentized residual*: This standardized residual is also referred to as the *deleted* residual or *deleted studentized* residual. Let $\widehat{\sigma}^2_{-i}$ be the estimator of $\sigma^2$ when the $i$th data point is deleted. The externally studentized residual is defined to be
  $$t_i = \frac{\widehat{e}_{ols}\left(\mathbf{s}_i\right)}{\widehat{\sigma}_{-i}\sqrt{1-h_{ii}}}.$$

  The externally studentized residual is better at identifying outlying $Z$ observations than $r_i$ and both $r_i$ and $t_i$ are better than $r_i^*$.

- *Leverage values and other diagnostics*: The diagonal elements $h_{ii}$ of $\mathbf{H}$ are called leverage values. Their value is interpreted as the weight an observation has in determining its predicted value. It can be shown that $1/n \leq h_{ii} \leq 1$ and $\sum h_{ii} = p$. It can also be shown that $h_{ii}$ measures the distance between the covariates associated with the $i$th observation and the means of the covariate values for all $n$ observations (i.e. the centroid of the $\mathbf{X}$ space. Thus, high leverage values identify *potential* influential points in the $\mathbf{X}$ space. A standard rule of thumb is that a leverage value $h_{ii}$ is large if it exceeds $2\overline{h} = 2p/n$.

  Leverage values and standardized residuals are both important for the role they play in various diagnostic tools including:

1. *Leave-one-out OLS estimator*: The estimate of the regression coefficients when the $i$th observation is deleted is

$$\widehat{\boldsymbol{\beta}}_{ols,-i} = \widehat{\boldsymbol{\beta}}_{ols} - \left(\mathbf{X}(\mathbf{s})^{'}\mathbf{X}(\mathbf{s})\right)^{-1}\mathbf{x}(\mathbf{s}_i)\,\widehat{e}_{ols}(\mathbf{s}_i)_{-i}$$

where

$$\widehat{e}_{ols}(\mathbf{s}_i)_{-i} = Z(\mathbf{s}_i) - \widehat{Z}(\mathbf{s}_i)_{-i} = \frac{\widehat{e}_{ols}(\mathbf{s}_i)}{1-h_{ii}}.$$

Note that $\sum_{i=1}^{n}\left(\widehat{e}_{ols}(\mathbf{s}_i)_{-i}\right)^2 = \left[\widehat{\mathbf{e}}_{ols}(\mathbf{s})_{-i}\right]^{'}\left[\widehat{\mathbf{e}}_{ols}(\mathbf{s})_{-i}\right]$ is the $PRESS$ statistic which is often used as a model comparison tool; small values indicating good candidate models.

2. *Cook's Distance*: This is a measure of the influence of the $i$th observation on all fitted values. It is given by

$$D_i = \frac{\left[\widehat{\mathbf{e}}_{ols}(\mathbf{s})_{-i}\right]^{'}\left[\widehat{\mathbf{e}}_{ols}(\mathbf{s})_{-i}\right]}{p\widehat{\sigma}^2} = \frac{r_i^2 h_{ii}}{p(1-h_{ii})}.$$

Interpretation of $D_i$ is based on treating it as a percentile from an $F_{p,n-p}$ distribution. If the percentile is less than 20% the $i$th case is generally considered to have little influence. If the percentile value starts getting close to 50% then the $i$th case is generally interpreted as having influence in that the the fitted values obtained with and without the $i$th observation differ a good deal.

3. *DFFITS*: This is a measure of the influence the $i$th observation has on the $i$ fitted value. It is

$$DFFITS_i = t_i\sqrt{\frac{h_{ii}}{1-h_i i}}.$$

A standard rule of thumb for "large" values is $|DFFITS_i| > 1$ for small to medium data sets and $|DFFITS_i| > 2\sqrt{p/n}$ for larger data sets.

- *OLS residuals*: Residual analysis is a mainstay of diagnostic evaluation of fitted regression models. Plots of residuals are scrutinized to assess departures from linearity, homoscedasticity, and normality. In the spatial context in which we are working one common use of residuals is to assess the assumption

$$\boldsymbol{\Sigma} = \sigma^2\mathbf{I}.$$

In particular, the residuals are often subjected to a semivariogram analysis to determine the presence of residual spatial correlation which would be interpreted as an indication that the mean structure of the model was not adequate to explain the spatial variation.

Schabenberger and Gotway point out that although the error terms in the model $\mathbf{e}(\mathbf{s})$ are uncorrelated, and homoscedastic the residuals $\widehat{\mathbf{e}}_{ols}(\mathbf{s})$ are not. This point is also made in most linear regression courses (including our STAT 410) but it is not emphasized.

The result is that the residuals are

1. *correlated*
2. *heteroscedastic*

Schabenberger and Gotway (page 308) point out that the residuals can also be rank deficient if $\mathbf{X}(\mathbf{s})$ is rank deficient. The fact that the residuals are correlated and heteroscedastic follows from the fact that the variance-covariance matrix of the residuals is

$$\mathrm{Var}\left(\hat{\mathbf{e}}(\mathbf{s})\right) = \sigma^2\mathbf{M}$$

and $\mathbf{M} = \mathbf{I} - \mathbf{H}$ is not a diagonal matrix. It is also a singular matrix so this is one example of a non-invertible covariance matrix, i.e. it is positive semi-definite instead of positive definite. The residuals are correlated because of the linear constraints they must obey:

- $\mathbf{1}^{'}\widehat{\mathbf{e}}_{ols}(\mathbf{s}) = 0$
- $\mathbf{X}(\mathbf{s})^{'}\widehat{\mathbf{e}}_{ols}(\mathbf{s}) = \mathbf{0}$

The residuals are heteroscedastic because the diagonal elements of the hat matrix are not equal to one another. Further,

$$\text{Var}\left(\widehat{e}_{ols}\left(\mathbf{s}_i\right)\right) = \sigma^2\left(1 - h_{ii}\right) < \sigma^2.$$

Thus, using the residuals to estimate a covariance function (semivariogram) of the error process $\mathbf{e}$ is potentially problematical because we end up estimating the covariance function (semivariogram) of the $\widehat{\mathbf{e}}_{ols}$ instead of the semivariogram of the error process itself.

Schabenberger and Gotway discuss several approaches, all of which are less than ideal.

- *Error Recovery*: Based on the above listed problems there are only $n - p$ bits of independent information in the residuals about the error process. The goal is to try to recover information about the error process $\mathbf{e}$ from the residuals $\widehat{\mathbf{e}}_{ols}$ taking the redundant information into account. Two methods are discussed.

  1. *Recursive or Sequential Residuals*: These are residuals determined as observations are sequentially added to the computations. The model is initially fit to $p$ observations. The remaining $n - p$ observations are added one at a time and a prediction $\widehat{Z}\left(\mathbf{s}_j\right)$ is made using the $j-1$ observations previously entered in the model. A recursive residual $w_j$ is the scaled difference between $Z\left(\mathbf{s}_j\right)$ and $\widehat{Z}\left(\mathbf{s}_j\right)$. Let $\mathbf{X}_j\left(\mathbf{s}\right)$ be the $j \times p$ matrix containing the first $j$ rows of observations. The $j$th recursive residual can be calculated as

$$w_j = \frac{\widehat{e}_{ols}\left(\mathbf{s}_j\right)}{\left\{1 - \mathbf{x}_j\left(\mathbf{s}_j\right)^{'}\left(\mathbf{X}_j\left(\mathbf{s}_j\right)^{'}\mathbf{X}_j\left(\mathbf{s}_j\right)\right)^{-1}\mathbf{x}_j\left(\mathbf{s}_j\right)\right\}^{1/2}}.$$

     None of the observations are used in predicting themselves.

     There is a function `recresid` in the `strucchange` package in R that will caculate recursive residuals.

  2. *Error Recovery*: The idea is to find an $n \times n$ transformation matrix $\mathbf{Q}$ such that

$$Var\left[\mathbf{Q}^{'}\widehat{\mathbf{e}}_{ols}\left(\mathbf{s}\right)\right] = \begin{bmatrix} \mathbf{I}_{n-p} & \mathbf{0}_{(n-p)\times p} \\ \mathbf{0}_{p\times(n-p)} & \mathbf{0}_{p\times p} \end{bmatrix}.$$

     The first $n-p$ elements of the $n\times 1$ "residual" vector $\mathbf{Q}^{'}\widehat{\mathbf{e}}_{ols}\left(\mathbf{s}\right)$ are the Linearly Unbiased Scaled residuals, also called the Linearly Recovered Errors.

     We know that the variance-covariance matrix of the residuals is

$$\text{Var}\left(\widehat{\mathbf{e}}_{ols}\left(\mathbf{s}\right)\right) = \sigma^2\mathbf{M}.$$

     It can be shown that $\mathbf{M}$ can be written as a product

$$\mathbf{M} = \mathbf{P}\boldsymbol{\Delta}\mathbf{P}^{'}$$

     where $\boldsymbol{\Delta}$ is a diagonal matrix and $\mathbf{P}$ is orthogonal (implying that $\mathbf{P}^{'}\mathbf{P} = \mathbf{I}$). If $\mathbf{Y}$ is a random vector with variance-covariance matrix $\boldsymbol{\Sigma}_Y$ and $\mathbf{A}$ is a matrix of constants then

$$\text{Var}\left(\mathbf{AY}\right) = \mathbf{A}\boldsymbol{\Sigma}_y\mathbf{A}^{'}.$$

     Letting $\mathbf{Q} = \mathbf{P}/\sigma$ leads to

$$\text{Var}\left\{\left(\mathbf{P}^{'}/\sigma\right)\left[\widehat{\mathbf{e}}_{ols}\left(\mathbf{s}\right)\right]\right\} = \mathbf{P}^{'}\mathbf{P}\ \boldsymbol{\Delta}\mathbf{P}^{'}\mathbf{P} = \ \boldsymbol{\Delta}$$

     where

$$\boldsymbol{\Delta} = \begin{bmatrix} \mathbf{I}_{n-p} & \mathbf{0}_{(n-p)\times p} \\ \mathbf{0}_{p\times(n-p)} & \mathbf{0}_{p\times p} \end{bmatrix}.$$

     The linearly recovered error terms are linear combinations of the residuals and thus cannot be matched up with a corresponding $\widehat{e}$. As such they are not useful for identifying unusual observations. They are also not unique because the decomposition

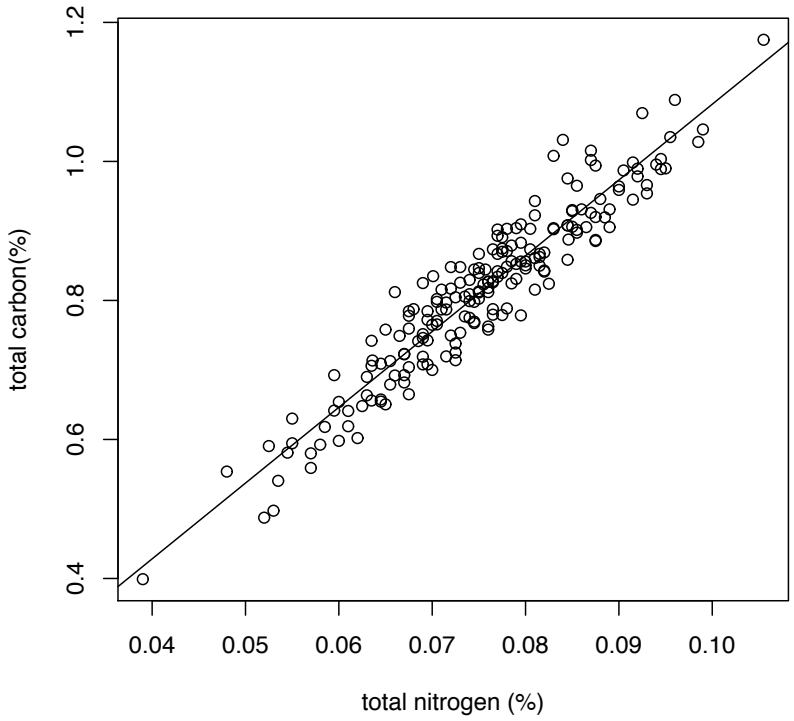
$$\mathbf{M} = \mathbf{P}\boldsymbol{\Delta}\mathbf{P}^{'}$$

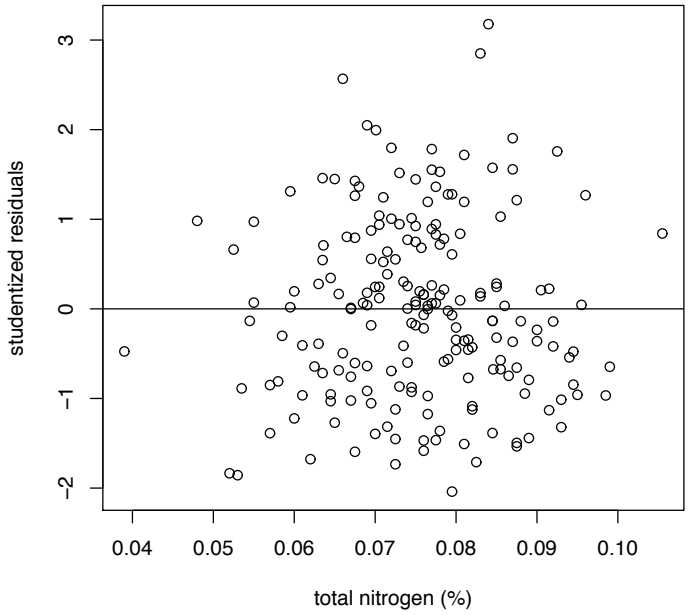
     is not unique. Their usefulness in the spatial context is apparently still an open question. I could not find any R functions that will automatically do this for us.

- *Carbon-Nitrogen Example: I fit a linear model with total carbon as the response and total nitrogen as the predictor. A plot and summary of the results is given below.* $R^2 = 0.897$.

```
> summary(CN.lm)
Coefficients:
```

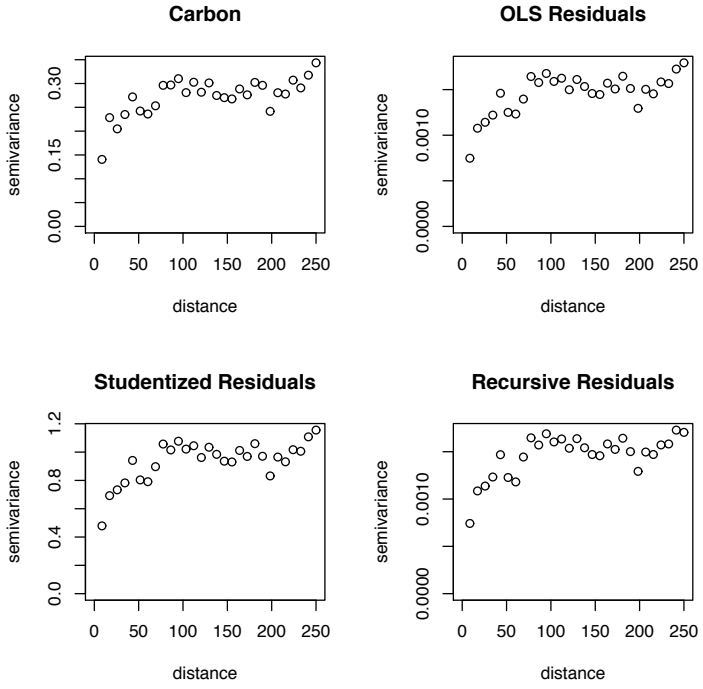
| | Estimate | Std. Error | t value | Pr(>\|t\|) |
|---|---|---|---|---|
| (Intercept) | $-0.007799$ | 0.020193 | $-0.386$ | 0.7 |
| tn | 10.900827 | 0.265224 | 41.100 | $< 2e-16$ |






*Outside of a couple of standardized residuals around 3 in size there is nothing in this plot that would cause much concern.*

*Empirical semivarigrams of the total carbon data, the OLS or raw residuals, the studentized residuals and the recursive residuals are shown below.*




*The commands that produced these are*

```
par(mfrow=c(2,2))
plot(variog(coords=CN.dat[,1:2],data=CN.dat$cn,uvec=seq(0,250,l=30)))
title(main="Carbon")
plot(variog(coords=CN.dat[,1:2],data=resid(CN.lm),uvec=seq(0,250,l=30)))
title(main="OLS Residuals")
plot(variog(coords=CN.dat[,1:2],data=studres(CN.lm),uvec=seq(0,250,l=30)))
title(main="Studentized Residuals")
recresid.CN=recresid(CN.lm)
plot(variog(coords=CN.dat[-(1:2),1:2],data=recresid.CN,uvec=seq(0,250,l=30)))
title(main="Recursive Residuals")
```

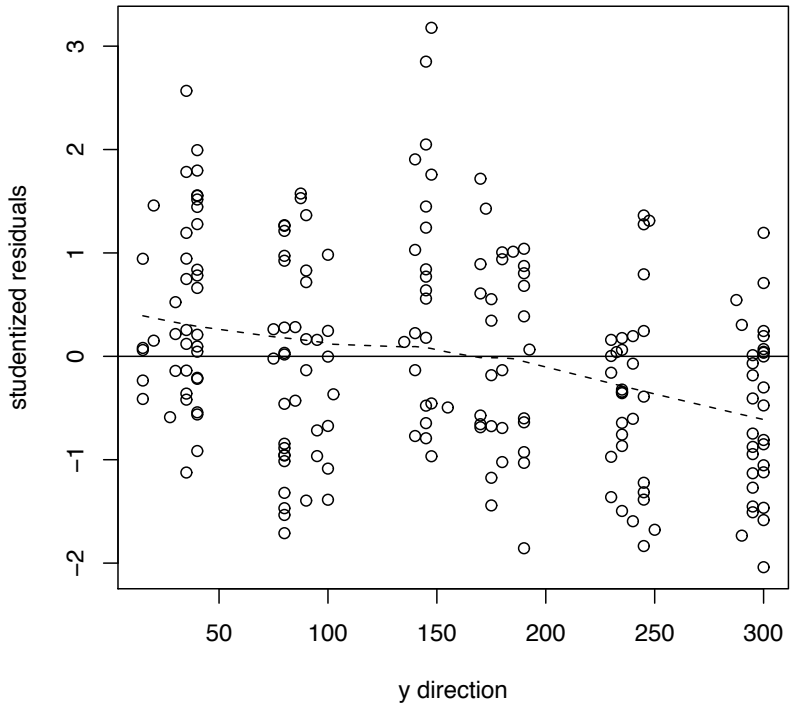
*The* `studres` *function is in the* `MASS` *library and the* `recresid` *is in the* `strucchange` *library.*

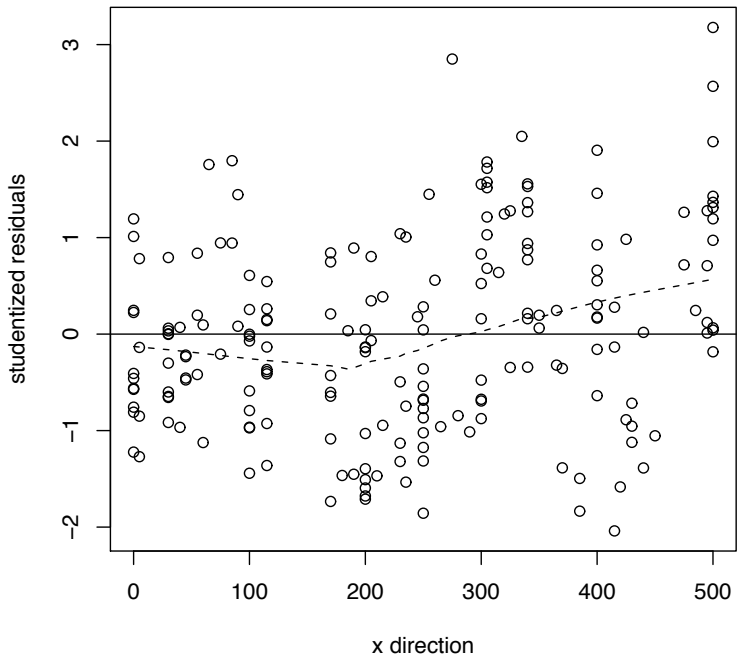
*A few things to note:*

1. *Note the clear indication of spatial correlation remaining after fitting the model.*
2. *Note that the sill in the OLS and recursive residual semivariograms is less than the sill in the carbon semivariogram indicating that regressing nitrogen on carbon explained some of the variation.*
3. *Note the sill for the studentized residuals is close to 1 as it must be because they have been standardized to have unit variance. Thus, this semivarigram could not be used to model the covariance structure by itself.*
4. *The recursive residual function fits the initial model to the first 2 observations so there are no residuals for those observations.*

- We have looked at trend models before (universal kriging). But there the emphasis was on prediction and the coefficients of the trend model were of secondary importance. Here the coefficients themselves are of interest.

  *Example: We have seen that the residuals from regressing total carbon on total nitrogen exhibit spatial correlation. Below are plots of the residuals from the simple linear regression versus the $x$ and $y$ coordinates.*






*There is a hint of a quadratic trend in the $x$ direction and a linear trend in the $y$ direction. I added a quadratic trend model to the model with total nitrogen in it. At a location $\mathbf{s} = (x, y)$ the model is*

$$Z(\mathbf{s}) = \beta_0 + \beta_{01}x + \beta_{10}y + \beta_{11}xy + \beta_{20}x^2 + \beta_{02}y^2 + \beta_2 X(\mathbf{s}) + e(\mathbf{s}).$$

*We are primarily interested in the relationship between total nitrogen and total carbon and the trend components have been entered into the model in the hopes of controlling for large scale spatial variation manifested as spatially autocorrelated residuals from the simple linear regression model. I centered the $x$ and $y$ coordinates in an effort to control for multicollinearity as the pairwise correlations between $x$ and $x^2$ and $y$ and $y^2$ are $0.96$ and $0.97$, respectively. Many people are surprised to find that quadratic terms can be highly correlated with linear terms. Interactions can also result in predictors with strong linear dependencies. Centering reduced these correlations to $0.09$ and $0.18$, respectively.*

*The results of fitting that model are given in the table below.*

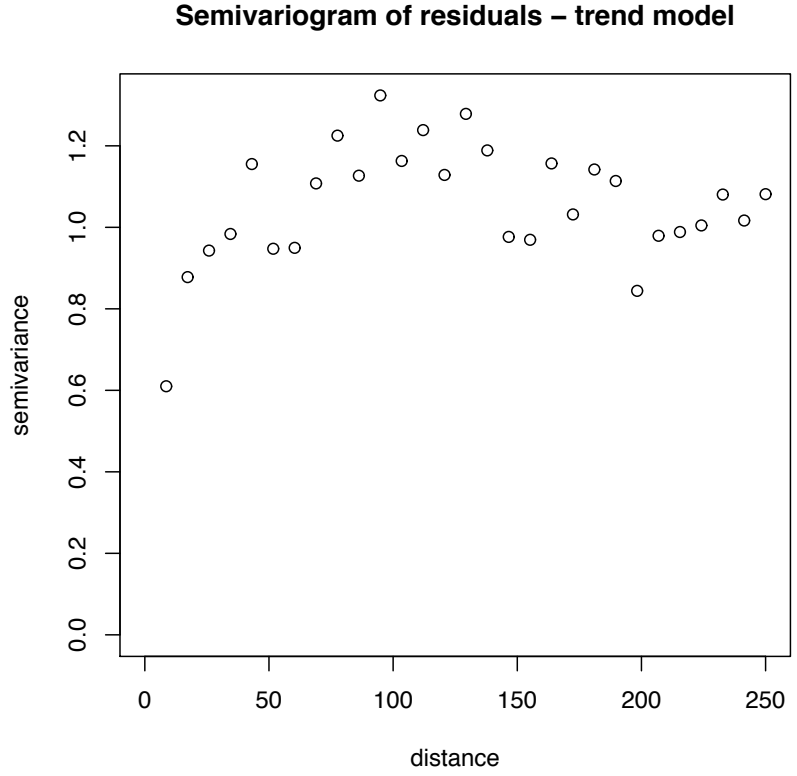
`Coefficients:`

| | Estimate | Std. Error | t value | Pr($>\|t\|$) |
|---|---|---|---|---|
| (Intercept) | $-2.241e-02$ | $2.234e-02$ | $-1.003$ | $0.317102$ |
| CN.x | $6.776e-05$ | $1.746e-05$ | $3.880$ | $0.000144$ |
| CN.y | $-1.350e-04$ | $2.938e-05$ | $-4.595$ | $7.9e-06$ |
| CN.xsq | $3.382e-07$ | $1.253e-07$ | $2.700$ | $0.007576$ |
| CN.ysq | $-3.556e-07$ | $3.768e-07$ | $-0.944$ | $0.346479$ |
| tn | $1.104e+01$ | $2.761e-01$ | $39.965$ | $< 2e-16$ |
| CN.x:CN.y | $-2.549e-07$ | $1.835e-07$ | $-1.389$ | $0.166471$ |

*($R^2 = 0.92$).*

*We do see significant trend effects, although the $R^2$ value has not increased all that much. Note that the effect of total nitrogen remains pretty much unchanged after controlling for large scale trend. The plot below shows the empirical semivariogram of the residuals from this model. It appears that we do still have some evidence*

*of spatial correlation but it is less now (i.e. the range has decreased). The trend model seems to have been a little successful in controlling for larger scale correlations but there is still evidence of spatial correlation over smaller scales.*




- *Linear Models with Correlated Errors:*
  - *Generalized Least Squares:* We have looked at the Generalized Least Squares ($GLS$) approach earlier. Given a model

$$\mathbf{Z}(\mathbf{s}) = \mathbf{X}(\mathbf{s})\boldsymbol{\beta} + \mathbf{e}(\mathbf{s})$$

with $\mathbf{e}(\mathbf{s}) \sim (\mathbf{0}, \boldsymbol{\Sigma}(\boldsymbol{\theta}))$ the $GLS$ estimator of $\boldsymbol{\beta}$ is

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

We do not know $\boldsymbol{\theta}$ and must estimate it yielding an *estimated generalized least squares estimator*

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

Asymptotically (assuming suitable *regularity conditions including normality*),

$$\widehat{\boldsymbol{\beta}}_{egls} \sim N\left(\boldsymbol{\beta}, \left(\mathbf{X}(\mathbf{s})' \boldsymbol{\Sigma}(\boldsymbol{\theta})^{-1} \mathbf{X}(\mathbf{s})\right)^{-1}\right).$$

If we are willing to assume $\mathbf{e}(\mathbf{s}) \sim N(\mathbf{0}, \boldsymbol{\Sigma}(\boldsymbol{\theta}))$ then we can find $ML$ or $REML$ estimators and the regularity conditions are automatically satisfied.

*Example: We now know that we have spatially correlated error terms in the carbon-nitrogen regression example. The semivariogram of the residuals provides us with reasonable starting values for estimation*

*of the covariance parameters in a GLS approach. Actually, we will estimate these parameters using both ML and REML. We could do that with* `likfit` *but* `geoR` *does not have the capability of fitting spatial regression models (at least easily) so we will use the* `gls` *function in the* `nlme` *library.*

*We assume an exponential covariance function. We have already seen the results from using* `lm` *to fit the uncorrelated errors model but we will reproduce them here using* `gls`*.*

```
CNgls.ols<-gls(TC∼TN,data=CN.dat)
summary(CNgls.ols)
Generalized least squares fit by REML
   AIC       BIC      logLik
 −689.021  −679.233  347.5105

Coefficients:
                Value     Std.Error   t-value   p-value
 (Intercept)  −0.007799   0.0201928  −0.38620    0.6998
     TN       10.900827   0.2652242   41.10043    0.0000

Residual standard error:  0.03982155
```

*Note that the default method of estimation in* `gls` *is REML and that the* `gls` *REML results match exactly the results from* `lm` *when fitting an uncorrelated errors model. We are assuming a pure nugget effect model with sill estimated to be* $(0.03982)^2 = 0.00159$*.*

*The R commands below generate REML estimates for the exponential model with no nugget effect.*

```
CNglsexp.reml<-gls(TC∼TN,data=CN.dat,corExp(c(15),form=∼x+y))
summary(CNglsexp.reml)
Generalized least squares fit by REML
    AIC       BIC      logLik
 −729.9188  −716.868  368.9594

Correlation Structure:  Exponential spatial correlation
Formula:  ∼ x + y
Parameter estimate(s):
range
14.12821

Coefficients:
                Value     Std.Error   t-value   p-value
 (Intercept)  0.006378   0.02282106   0.27946    0.7802
     TN       10.732504   0.29604290  36.25320    0.0000

Residual standard error:  0.04138313
```

*We estimate the practical range to be* $3 * 14.128 = 42.38$ *and the sill to be* $(0.0414)^2 = 0.00171$*.*

*We do not get standard errors for the covariance parameters but we can construct confidence intervals easily using the* `intervals` *command. The default confidence level is 95%.*

```
intervals(CNglsexp.reml)
Approximate 95% confidence intervals

Coefficients:
                lower         est.         upper
 (Intercept)  −0.03863314   0.006377568   0.05138827
     tn       10.14860874  10.732503588  11.31639843

Correlation structure:
        lower     est.     upper
 range  9.25926  14.12823  21.55752
```

```
Residual standard error:
   lower       est.       upper
 0.03635096  0.04138314  0.04711194
```

*Maximum likelihood estimates can be generated by including* `method=''ML''` *in the argument list.*

*Fitting the nugget effect in* `gls` *is a bit different. This function estimates the (effective) range and nugget effect of a correlation function. The sill is estimated separately. Thus, the estimated nugget effect from a* `gls` *fit is actually an estimate of the proportion of the sill that is due to the nugget. The semivariogram of the studentized residual plot can be helpful in providing an initial estimate of the nugget effect for a* `gls` *fit. Looking back at that plot suggests an initial value of around* 0.4.

```
CNglsexp.reml2<-gls(TC∼TN,data=CN.dat,corExp(c(15,.4),form=∼x+y,nugget=T))
summary(CNglsexp.reml2)
Generalized least squares fit by REML
    AIC        BIC      logLik
 −742.2505  −725.9371  376.1253

Correlation Structure:  Exponential spatial correlation
Formula:  ∼x + y
Parameter estimate(s):
    range      nugget
 58.36308847  0.3562304

Coefficients:
                Value     Std.Error   t-value   p-value
 (Intercept)  −0.010622  0.02424491  −0.43810    0.6618
      TN      10.995825  0.29968656  36.69109    0.0000

Residual standard error:  0.04211086
intervals(CNglsexp.reml2)
Approximate 95% confidence intervals


Correlation structure:
           lower       est.        upper
 range   19.4965418  58.3630884  174.7104757
 nugget   0.1944255   0.3562304   0.5592159

Residual standard error:
   lower       est.       upper
 0.03454806  0.04211086  0.05132920
```

*The sill is estimated to be* $(0.04211086)^2 = 0.00177$ *and* $0.3562304$ *of that is due to the nugget effect, i.e. the estimated nugget is* $(0.04211086)^2(0.3562304) = 0.000621$.

| *Model* | $\widehat{\beta}_1$ | *Nugget* | *Sill* | *(Practical) Range* |
|---|---|---|---|---|
| *Ind.* | 10.900 | – | 0.00159 | 0 |
| *Errors* | (0.265) | – | (0.00131, 0.00196) | – |
| *Exponential* | 10.733 | – | 0.00171 | 42.385 |
| *REML* | (0.296) | – | (0.00132, 0.00222) | (27.778, 64.673) |
| *Exponential* | 10.738 | – | 0.00167 | 40.985 |
| *ML* | (0.295) | – | (0.00130, 0.00215) | (26.969, 62.284) |
| *Exponential* | 10.996 | 0.356 | 0.00177 | 175.089 |
| *REML* | (0.300) | (0.194, 0.559) | (0.00119, 0.00263) | (58.490, 524.131) |
| *Exponential* | 10.976 | 0.363 | 0.00167 | 144.929 |
| *ML* | (0.299) | (0.196, 0.571) | (0.00121, 0.00229) | (58.588, 358.515) |

*Estimates of* $\beta_1$ *are comparable in all cases but the standard errors are higher in the correlated errors models. Estimates of the sill are also comparable with wider confidence intervals in the correlated errors models. Whether or not we include a nugget effect clearly matters. The estimates of the range are quite*

*different in the nugget effect models. Note also the increased uncertainty in our estimates of the sill. Based on the semivariogram plots we looked at earlier it would seem prudent to include a nugget effect. We will look at model comparisons more formally a bit later.*

– *Residual Analysis*: The generalized least squares raw residuals are

$$\widehat{\mathbf{e}(\mathbf{s})}_{egls}(\mathbf{s}) = \mathbf{Z}(\mathbf{s}) - \widehat{\mathbf{Z}}(\mathbf{s}) = \left(\mathbf{I} - \mathbf{H}(\widehat{\boldsymbol{\theta}})\right)\mathbf{Z}(\mathbf{s}) = \mathbf{M}(\widehat{\boldsymbol{\theta}})\mathbf{Z}(\mathbf{s})$$

where

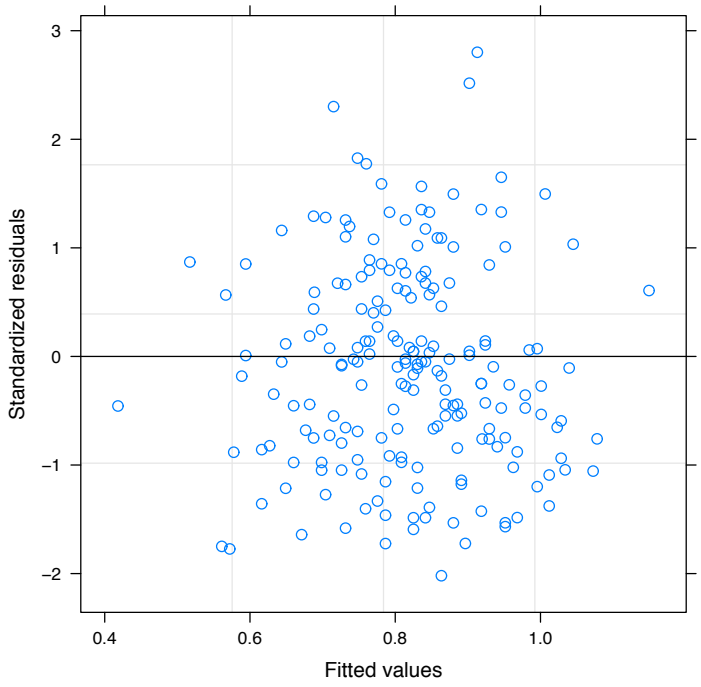
$$\mathbf{H}(\widehat{\boldsymbol{\theta}}) = \mathbf{X}(\mathbf{s})\left(\mathbf{X}(\mathbf{s})'\boldsymbol{\Sigma}(\widehat{\boldsymbol{\theta}})^{-1}\mathbf{X}(\mathbf{s})\right)^{-1}\mathbf{X}(\mathbf{s})'\boldsymbol{\Sigma}(\widehat{\boldsymbol{\theta}})^{-1}.$$

These residuals suffer from the same problems as the $\widehat{\mathbf{e}(\mathbf{s})}_{ols}$ do. We will make use of some of the diagnostic tools in `nlme`.

*Example: Carbon-Nitrogen regression example:*

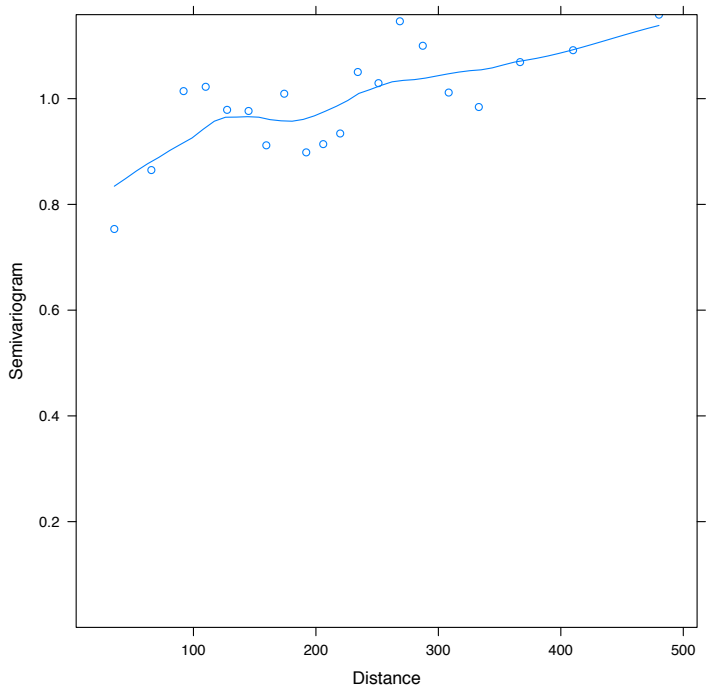
1. `qqnorm`*: This function produces a normal probability plot of the residuals. For example, the normal probability plot of normalized residuals (more on this normalization below) from the REML fit with a nugget effect is shown below and looks pretty good.*
   `qqnorm(CNglsexp.reml2,~resid(.,type="n"))`

2. `plot.gls`*:*
   *A plot of standardized residuals versus the fitted values in the REML model with a nugget effect is seen below and shows nothing of real concern.*
   `plot(CNglsexp.reml2)`



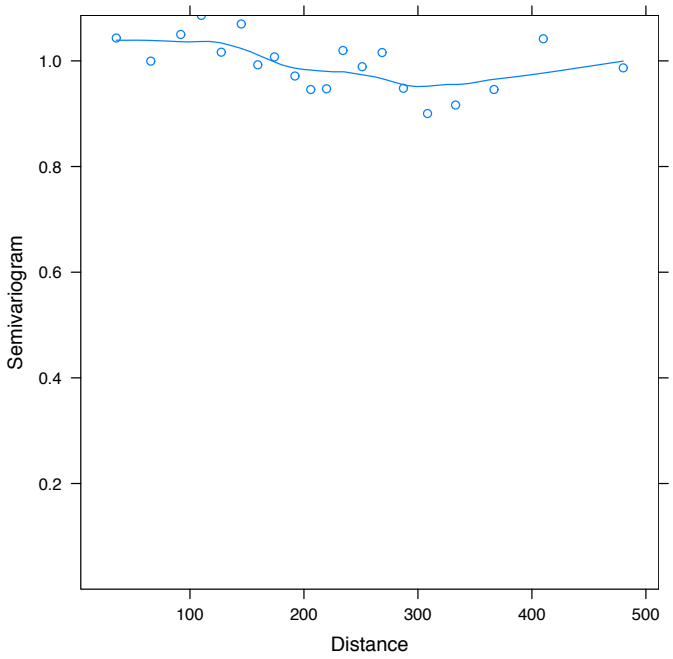

3. `plot.Variogram`:
*This is quite useful especially when applied to normalized residuals. Normalized residuals have been transformed in such a way that they will be uncorrelated if the assumed covariance model is reasonable. A plot of the empirical semivariogram of the normalized residuals should show a pure nugget effect. We know the OLS model is inadequate. Below is the* `plot.Variogram` *plot of the normalized residuals from that fitted model. You can see a clear deviation from a pure nugget effect.*

```
plot(Variogram(CNgls.ols,resType="n",form=~x+y))
```



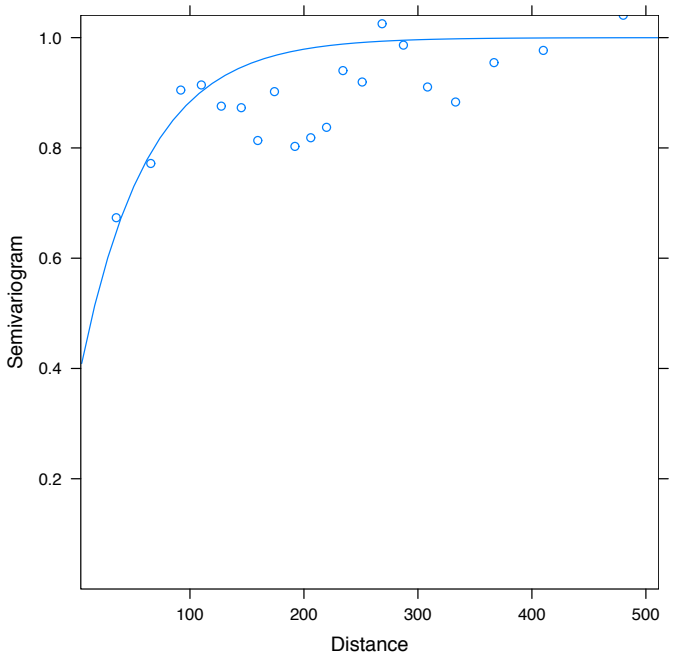
*A similar plot of the REML fit with an exponential model with nugget effect is shown next and it is much better. Clearly the exponential covariance model works better than an assumption of independent errors.*

```
plot(Variogram(CNglsexp.reml2,resType="n"))
```



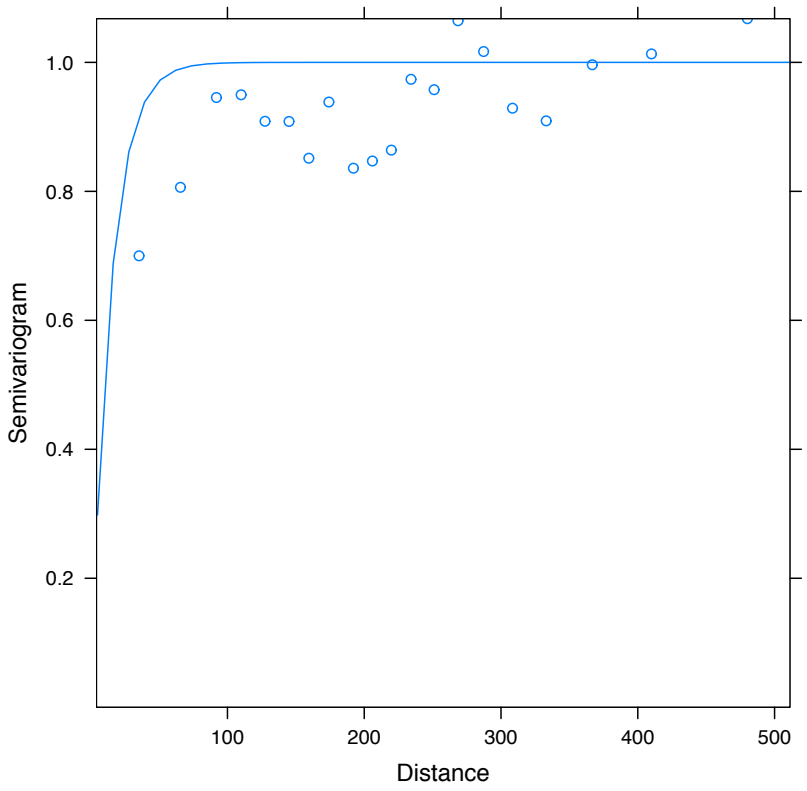

*The plot is also useful in providing a visual assessment of how well the estimated covariance function fits the empirical semivariogram.*

```
plot(Variogram(CNglsexp.reml2))
```




*The nugget effect model seems to fit a bit better than the exponential covariance model without a nugget effect.*

```
plot(Variogram(CNglsexp.reml))
```




To see how the normalization works consider a random vector $\mathbf{Y}$ with covariance matrix $\mathbf{\Sigma}$. We can find write this as a product of 2 *square root* matrices $\mathbf{\Sigma}^{1/2}$, i.e.

$$\mathbf{\Sigma} = \mathbf{\Sigma}^{1/2}\mathbf{\Sigma}^{1/2}.$$

It is important to note that the elements of $\mathbf{\Sigma}^{1/2}$ are not equal to the square root of the elements of $\mathbf{\Sigma}$ unless $\mathbf{\Sigma}$ is a diagonal matrix. Also, $\mathbf{\Sigma}^{1/2}$ is not unique. If $\mathbf{\Sigma}$ is positive definite then $\mathbf{\Sigma}^{1/2}$ is positive definite and its inverse $\mathbf{\Sigma}^{-1/2}$ exists. So the transformed vector $\mathbf{\Sigma}^{-1/2}\mathbf{Y}$ has covariance matrix

$$\begin{aligned} \text{Var}\left(\mathbf{\Sigma}^{-1/2}\mathbf{Y}\right) &= \mathbf{\Sigma}^{-1/2}\mathbf{\Sigma}\mathbf{\Sigma}^{-1/2} \\ &= \mathbf{\Sigma}^{-1/2}\mathbf{\Sigma}^{1/2}\mathbf{\Sigma}^{1/2}\mathbf{\Sigma}^{-1/2} \\ &= \mathbf{I} \end{aligned}$$

This normalization is applied to the residuals using the estimated covariance matrix. If the estimated matrix is reasonable then a semivariogram of the normalized residuals should be a pure nugget effect. This is what we saw in the results of the

```
plot(Variogram(CNglsexp.reml2,resType="n"))
```

– *Inference in GLS*: Schabenberger and Gotway argue on pages 341-343 of their text that testing hypotheses about $\boldsymbol{\beta}$ should require an adjustment to the standard $F$ (or $t$) procedures. They discuss one method but it is, so far as I can tell, not implemented in any of the R packages we use. Technically they are correct but the bottom line is that once we move away from the uncorrelated errors least squares model (and the associated normal errors assumption) all test procedures essentially are based on approximations (and additional assumptions) of some sort. Tests and confidence intervals for $\boldsymbol{\beta}$ and $\boldsymbol{\theta}$ are based on asymptotic results and we will use those here.

We have already seen how to use the `intervals` formula to generate confidence intervals for model parameters. As just indicated these are approximate intervals; further they are correct under an assumption that the covariance model is correct, in fact that the fitted model itself is the *true* model. We can use information theoretic methods (AIC) to assess models with different covariance structures.

*Example: We will continue to look at the carbon-nitrogen example. We will compare the following models:*

1. *Independent Errors Model*
2. *Exponential Covariance Model*
3. *Spherical Covariance Model*
4. *Gaussian Model*

*In each case we will assume that that a nugget effect is included in the model. All models were fit using the REML method. AIC comparisons are only valid when the same method is used to fit all models.*

| *Model* | $\widehat{\beta}_1$ | *Nugget* | *Sill* | *(Practical) Range* | *AIC* | $\Delta AIC$ |
|---|---|---|---|---|---|---|
| *Ind. Errors* | 10.901 | 1 | 0.00156 | 0 | −689.021 | 53.23 |
| *Exponential* | 10.996 | 0.356 | 0.00177 | 175.08 | −742.251 | 0 |
| *Spherical* | 11.038 | 0.402 | 0.00183 | 143.93 | −739.204 | 3.05 |
| *Gaussian* | 10.939 | 0.513 | 0.00163 | 144.30 | −735.444 | 6.81 |

*The exponential covariance model comes out on top in this set of 4 candidate models.*

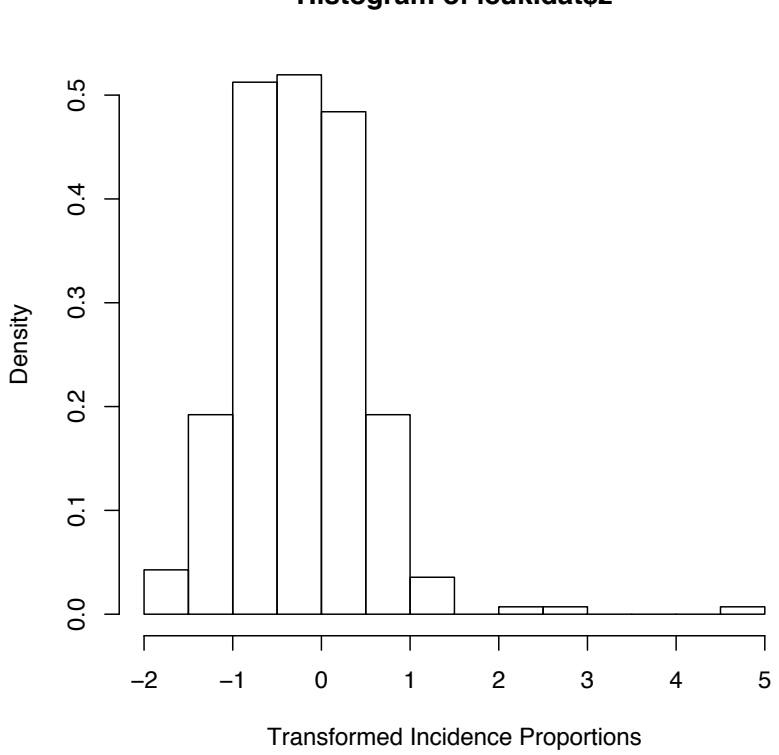
– *Multiple Regression - NY Leukemia Data: The data are the number of incident leukemia cases recorded over a 4 year period (1978-1982) in 281 census tracts in 8 counties in New York. There were 592 cases recorded among approximately 1 million people at risk. Data manipulation resulted in the number of cases per census tract not being necessarily an integer. The "counts" ranged in value from 0 to 9.29 in the census tracts. A number of covariates were also measured (see below).*

*Ordinary and generalized least squares are not ideal for count data because a number of the standard assumptions are commonly violated for such data. In particular, the important assumption of constant variance may not hold.*

*Let $Y_i$ be the number of cases in tract $i$ and $n_i$ be the population size in tract $i$. We could assume that $Y_i \sim Bin(n_i, p_i)$. Then with $n_i$ large and $p_i$ small (not unreasonable for rare diseases) the counts would be approximately Poisson distributed. You may recall that a standard variance stabilizing transformation for count data is the square root transformation. The binomial assumption is problematic here however, because the implied assumption that the $Y_i$'s are a sum of independent Bernoulli counts is likely a bit of a stretch. Some suggested transformations for counts that are more appropriately thought of as sums of dependent Bernoulli counts include*

$$\begin{aligned} W_i &= \frac{1000\,(Y_i+1)}{n_i} \\ V_i &= \left(\frac{1000 Y_i}{n_i}\right)^{1/2} + \left(\frac{1000\,(Y_i+1)}{n_i}\right)^{1/2} \\ Z_i &= \log W_i \end{aligned}$$

*A histogram of the transformed incidence proportions is shown below. We see some indication of outliers. Be careful with this plot though. It cannot be used to assess assumptions of normality for the error process.*




*We will use $Z_i$ as our response for this example. You will explore the impact of the outliers in a homework problem later, but it is worth noting now that those outliers are the census tracts with the 3 lowest population sizes,* 9, 99, *and* 143.

*Given our starting assumption we know that the mean and variance will be related so we expect heteroscedasticity. We will need to consider this possibility below.*

*The goal of the analysis was to assess the relationship between leukemia incidence and the proximity to waste sites containing TriChloroEthylene (TCE). Three covariates were of interest:*

1. *$X_1$ the inverse distance between the centroid of a census tract and the nearest TCE waste site*
2. *$X_2$ the percent of population in a tract over 65*
3. *$X_3$ the percent of population who own their own home*

*$X_1$ is the covariate of primary interest and the other 2 were included in an effort to control for possible confounding: cancer risk rises with increasing age and cancer risk is associated with socio-economic status.*

*We obviously want a linear relationship between each of the covariates and our response. There is no problem with the age covariate ($X_2$) and the homeownership covariate ($X_3$) but the relationship between $Z$ and the inverse distance to a waste site is not as nice. The researchers transformed this variable as follows:* $X_1 = \log(100 InvDist)$.

*We will fit 6 different multiple regression models:*

1. *Homoscedastic, independent errors*
2. *Heteroscedastic, independent errors*
3. *Correlated errors, exponential model*
4. *Correlated errors, spherical model*
5. *Weighted correlated errors, exponential model*
6. *Weighted correlated errors, spherical model*

*Model 1 yielded the following results.*

```
leukgls.ols<-gls(z ~ pexp+age65+home,data=leuk.dat)
summary(leukgls.ols)
Coefficients:
```

| | Value | Std.Error | t-value | p-value |
|---|---|---|---|---|
| (Intercept) | −0.517276 | 0.1585572 | −3.262396 | 0.0012 |
| pexp | 0.048836 | 0.0350635 | 1.392795 | 0.1648 |
| age65 | 3.950890 | 0.6054983 | 6.525022 | 0.0000 |
| home | −0.560041 | 0.1703080 | −3.288403 | 0.0011 |

*The residual standard error is* 0.657 *and* $AIC = 577.894$.

*The* `gls` *function can carry out a weighted least squares analysis using a* `weights` *argument. This is actually a generalized least squares model with a diagonal covariance matrix. The ith diagonal element contains the variance of the ith error term,* $\sigma_i^2$. *The model is*

$$Z_i = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + \beta_3 X_{3i} + \epsilon_i$$

*with* $\epsilon_i \sim N\left(0, \sigma_i^2\right)$. *The variances in this example are believed to depend on the census tract population sizes,* $n_i$. *In particular the assumption was made that*

$$\sigma_i^2 = \frac{\sigma^2}{n_i}.$$

*Thus, the covariance matrix is assumed to be*

$$\mathbf{\Sigma} = \sigma^2 \mathbf{D}$$

*where* $\mathbf{D}$ *is a diagonal matrix whose ith diagonal element is* $d_{ii} = 1/n_i$. *The model was fit in* `gls` *by the following command (*`pop` *is the variable containing the* $n_i$ *values).*

```
leukgls.wls<-gls(z ~ pexp+age65+home,data=leuk.dat,weights=varFixed(~1/pop))
summary(leukgls.wls)

Coefficients:
```

| | Value | Std.Error | t-value | p-value |
|---|---|---|---|---|
| (Intercept) | −0.778374 | 0.1411604 | −5.514113 | 0.0000 |
| pexp | 0.076256 | 0.0273122 | 2.792021 | 0.0056 |
| age65 | 3.856560 | 0.5712645 | 6.750918 | 0.0000 |
| home | −0.398695 | 0.1530514 | −2.604973 | 0.0097 |

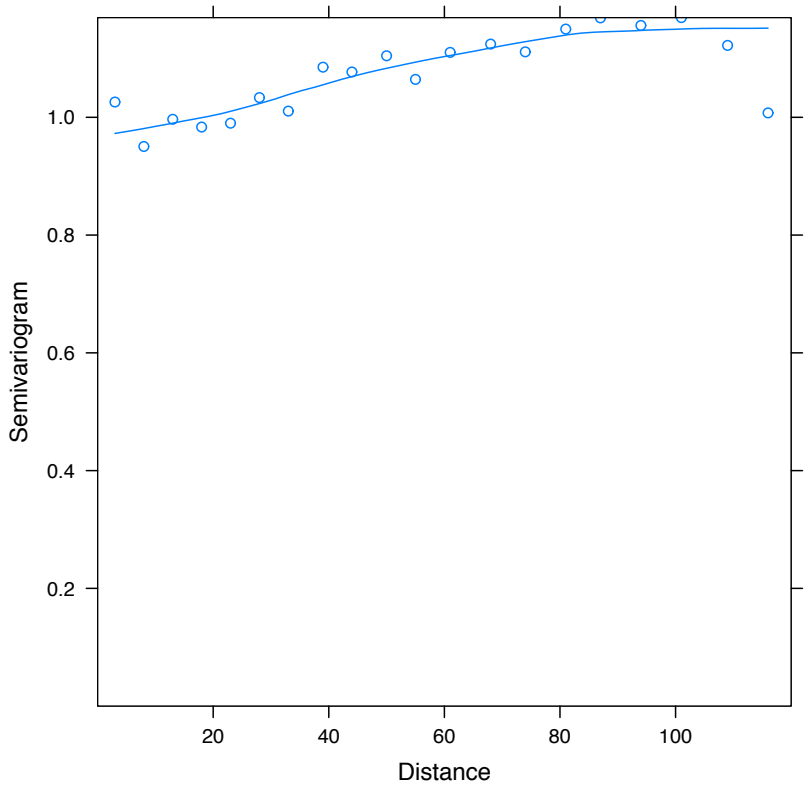
*The residual standard error (the estimate of* $\sigma$*) is* 33.495 *and* $AIC = 525.155$.

*Note that, after controlling for the 2 possibly confounding variables, the predictor of interest* `pexp` *is not significant in the OLS fit but is in the WLS fit. The reduction in AIC suggests that weighting produced a much better model. The weighting factor (*$1/n_i$*) essentially downweights the 3 outliers reducing their impact on the analysis. The question arises as to whether or not a weighted analysis would be needed if we removed those 3 observations from the data set which you will explore further in a homework problem.*

*Fitting generalized least squares models with correlated error structures requires an initial estimate of covariance model parameters. The plot below shows the empirical semivariogram of the residuals from the OLS fit.*

```
plot(Variogram(leukgls.ols,maxDist=120),xlim=c(0,120))
```

*We have a very large nugget effect and in fact it is not clear that modeling spatial correlation will help all that much in the end but it is worth checking. We will choose an initial "nugget" of around* 0.9 *and a range of* 18 *(or, assuming an exponential covariance model, a range parameter initial value of* 6*).*




```
leukgls.exp<-gls(z ~ pexp+age65+home,data=leuk.dat,
+ corExp(c(6,.9),form=~x+y,nugget=T)) # the + is a continuation symbol
summary(leukgls.exp)
```

```
Correlation Structure:  Exponential spatial correlation
Formula:  ~x + y
Parameter estimate(s):
   range     nugget
 1.9320742  0.8228562
```

```
Coefficients:
                Value     Std.Error    t-value    p-value
 (Intercept) −0.702142    0.1958089  −3.585855     0.0004
    pexp      0.080122    0.0431079   1.858639     0.0641
   age65      3.701832    0.6200298   5.970411     0.0000
    home     −0.344637    0.2035023  −1.693530     0.0915
```

*The residual standard error is* 0.656 *and* $AIC = 576.130$.

*The results from the fit with a spherical semivariogram are shown below.*

```
leukgls.spher<-gls(z ~ pexp+age65+home,data=leuk.dat,
+ corSpher(c(6,.9),form=~ x+y,nugget=T))
summary(leukgls.spher)
```

```
Correlation Structure:  Spherical spatial correlation
Formula:  ~ x + y
Parameter estimate(s):
   range     nugget
 6.9267081  0.8701206
```

```
Coefficients:
```

| | Value | Std.Error | t-value | p-value |
|---|---|---|---|---|
| (Intercept) | −0.722173 | 0.1972280 | −3.661615 | 0.0003 |
| pexp | 0.082657 | 0.0433564 | 1.906446 | 0.0576 |
| age65 | 3.709290 | 0.6188037 | 5.994292 | 0.0000 |
| home | −0.324511 | 0.2044359 | −1.587350 | 0.1136 |

*The residual standard error is* 0.656 *and* $AIC = 574.886$.

*The weighted results for the exponential and spherical models are:*

```
leukgls.spherw<-gls(z pexp+age65+home,data=leuk.dat,
+ weights=varFixed(∼ 1/pop),
+ corr=corSpher(c(6,.9),form=∼ x+y,nugget=T))
leukgls.expw<-gls(z pexp+age65+home,data=leuk.dat,
+ weights=varFixed(∼ 1/pop),
+ corr=corExp(c(6,.9),form=∼ x+y,nugget=T))
summary(leukgls.expw)

Correlation Structure:  Exponential spatial correlation
Formula:  ∼ x + y
Parameter estimate(s):
   range     nugget
 2.7686895  0.8915268
Coefficients:
```

| | Value | Std.Error | t-value | p-value |
|---|---|---|---|---|
| (Intercept) | −0.911776 | 0.1635553 | −5.574723 | 0.0000 |
| pexp | 0.095031 | 0.0325153 | 2.922654 | 0.0038 |
| age65 | 3.571075 | 0.5913111 | 6.039248 | 0.0000 |
| home | −0.237195 | 0.1734149 | −1.367787 | 0.1725 |

```
summary(leukgls.spherw)

Correlation Structure:  Spherical spatial correlation
Formula:  ∼ x + y
Parameter estimate(s):
   range     nugget
 6.8565321  0.8869828

Coefficients:
```

| | Value | Std.Error | t-value | p-value |
|---|---|---|---|---|
| (Intercept) | −0.916105 | 0.1648450 | −5.557371 | 0.0000 |
| pexp | 0.095627 | 0.0321494 | 2.974469 | 0.0032 |
| age65 | 3.576348 | 0.5920387 | 6.040734 | 0.0000 |
| home | −0.228457 | 0.1761173 | −1.297184 | 0.1956 |

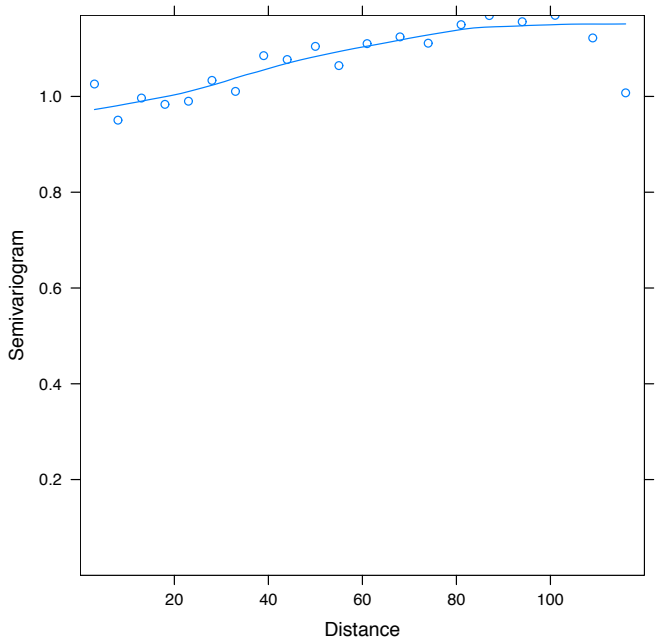
*The residual standard error for the exponential model is* 33.541 *and* $AIC = 526.218$. *The residual standard error for the spherical model is* 33.538 *and* $AIC = 525.383$.

*Summarizing the AIC results yields:*

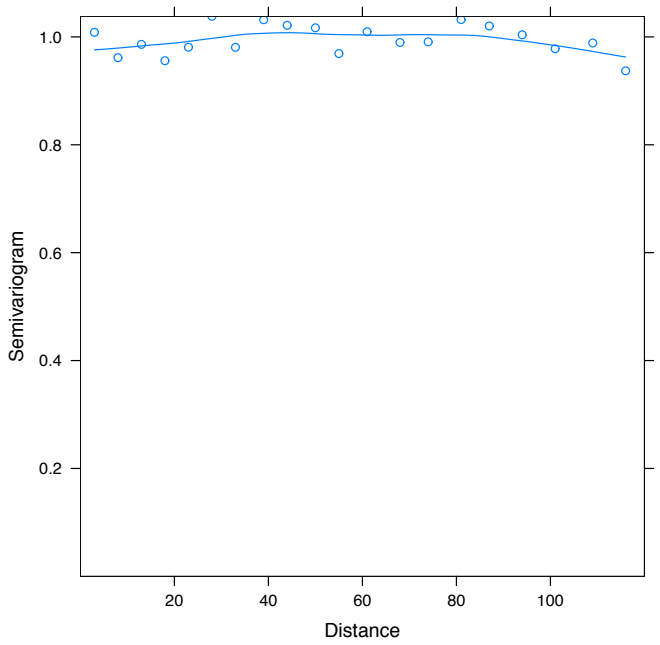
| *Model* | $\widehat{\beta_1}$ | $SE\left(\widehat{\beta_1}\right)$ | $AIC$ | $\Delta AIC$ |
|---|---|---|---|---|
| $OLS$ | 0.0488 | 0.0351 | 577.894 | 52.739 |
| $WLS$ | 0.0763 | 0.0273 | 525.155 | 0 |
| $GLS - Exp$ | 0.0801 | 0.0431 | 576.130 | 50.975 |
| $GLS - Spher$ | 0.0827 | 0.0434 | 574.886 | 49.731 |
| $GLS - ExpWt$ | 0.0950 | 0.0325 | 526.218 | 1.031 |
| $GLS - SpherWt$ | 0.0956 | 0.0321 | 525.383 | 0.228 |

*The $WLS$ fit is as good as any of them as judged by $AIC$. We can evaluate how good a job the different functions did in controlling for spatial dependency by looking at the semivariograms of the normalized residuals. We will look at the $OLS$, $WLS$, and spherical model results. The exponential and spherical model results are almost indistinguishable.*

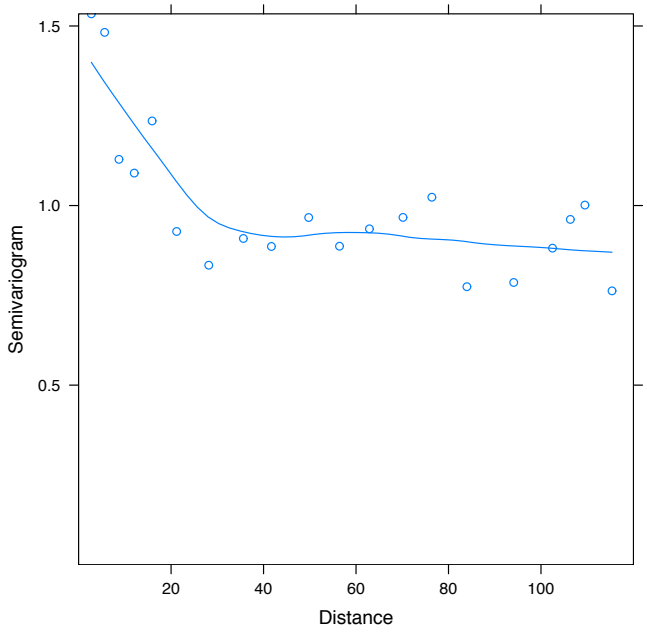
*$OLS$ fit*



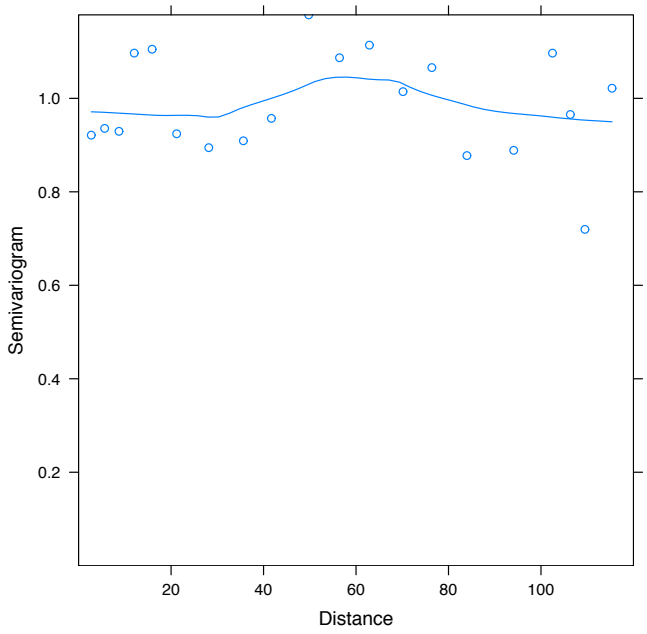

*$WLS$ fit*



*Spherical Model - Unweighted*




*Weighted Spherical Model*




*The weighted models clearly have performed the best but it is not apparent that we need the more complicated covariance structure of the weighted spherical model. Other checks would be needed to decide. It would be helpful to look at fitted values in the spatial context. A handout in class will show the differences between the fitted values from the weighted least squares independent errors model and the weighted generalized least squares fit with the exponential covariance model. Although both of these models appear to be equivalent in some sense they do provide somewhat different assessments based on the fitted values.*

- *Spatial Autoregression Models*: These models are typically applied to lattice or regional data. Autoregression refers to the practice of using (some) neighboring sites as predictors in a regression model. Unlike the spatial regression models we have fit so far we will not use the parametric covariance functions we have been working with to model the spatial correlation structure. There are at least 2 reasons for this:

  1. We are working on a lattice and the spatial coordinates are generally centroids of some region. It may not even make sense to estimate a mean response at some other coordinate that is not a centroid.
  2. A lot of times lattice data do not have a lot of observations. It is hard to model spatial covariance when one is working with a small sample size (i.e. empirical semivariograms are not particularly informative).

  There are 2 types of autoregressive models - Simultaneous Autoregressive ($SAR$) Models and Conditionally Autoregressive (CAR) Models.

  – $SAR$ Models: We start with our basic model:

  $$\mathbf{Z}(\mathbf{s}) = \mathbf{X}(\mathbf{s})\boldsymbol{\beta} + \mathbf{e}(\mathbf{s}).$$

  We then autoregress the error terms on other error terms:

  $$\mathbf{e}(\mathbf{s}) = \mathbf{B}\mathbf{e}(\mathbf{s}) + \mathbf{v}.$$

  It is typical to assume that

  $$\mathbf{v} \sim N\left(\mathbf{0}, \sigma^2\mathbf{I}\right).$$

  The role of the elements in $\mathbf{B}$ may be made clearer by noting that the second matrix equation implies that the $i$th element of the error vector can be written

  $$e(\mathbf{s}_i) = \sum_{j=1}^{n} b_{ij} e(\mathbf{s}_j) + v(\mathbf{s}_i).$$

  The *spatial dependence parameters* $\{b_{ij}\}$'s measure the contributions of $e(\mathbf{s}_j),\ j \neq i$ to the variation in $e(\mathbf{s}_i)$. We assume that $b_{ii} = 0$ as it would not make much sense to regress $e(\mathbf{s}_i)$ on itself. We can rewrite the model,

  $$\begin{aligned} Z(\mathbf{s}_i) &= \mathbf{x}(\mathbf{s}_i)'\boldsymbol{\beta} + \sum_{j=1}^{n} b_{ij} e(\mathbf{s}_j) + v(\mathbf{s}_i) \\ &= \mathbf{x}(\mathbf{s}_i)'\boldsymbol{\beta} + \sum_{j=1}^{n} b_{ij}\left[Z(\mathbf{s}_j) - \mathbf{x}(\mathbf{s}_j)'\boldsymbol{\beta}\right] + v(\mathbf{s}_i) \end{aligned}$$

  with the second term following from the fact that $e(\mathbf{s}_j) = Z(\mathbf{s}_j) - \mathbf{x}(\mathbf{s}_j)'\boldsymbol{\beta}$. Note that if $b_{ij} = 0$ for all $i$ and $j$ we just have the standard multiple regression model.
  Rewrite the above as

  $$v(\mathbf{s}_i) = Z(\mathbf{s}_i) - \mathbf{x}(\mathbf{s}_i)'\boldsymbol{\beta} + \sum_{j=1}^{n} b_{ij}\left[Z(\mathbf{s}_j) - \mathbf{x}(\mathbf{s}_j)'\boldsymbol{\beta}\right]$$

  or in matrix notation as

  $$(\mathbf{I} - \mathbf{B})(\mathbf{Z}(\mathbf{s}) - \mathbf{X}(\mathbf{s})\boldsymbol{\beta}) = \mathbf{v}.$$

  If $(\mathbf{I} - \mathbf{B})$ is invertible then

  $$\mathbf{Z}(\mathbf{s}) - \mathbf{X}(\mathbf{s})\boldsymbol{\beta} = (\mathbf{I} - \mathbf{B})^{-1}\mathbf{v}$$

  which makes it easier to see that

  $$\mathrm{Var}(\mathbf{Z}(\mathbf{s})) = \boldsymbol{\Sigma}_{SAR} = (\mathbf{I} - \mathbf{B})^{-1}\boldsymbol{\Sigma}_{\mathbf{v}}(\mathbf{I} - \mathbf{B})^{-1}.$$

The choice of the spatial dependence parameters in $\mathbf{B}$ induces a spatial covariance model for $\mathbf{Z}(\mathbf{s})$. This is called a simultaneous autoregressive model because it is applied simultaneously to each data location.

Clearly estimation of $\mathbf{B}$ is important. A common simplifying assumption is $\mathbf{B} = \rho\mathbf{W}$ where $\mathbf{W}$ is known and $\rho$ is to be estimated. $\mathbf{W}$ is commonly chosen to be one of the spatial proximity matrices we discussed earlier in the course (recall the rook, bishop, etc. definitions of distance). More specifically we could choose

1.
$$w_{ij} = \begin{cases} 1 & \text{regions } i \text{ and } j \text{ share a boundary} \\ 0 & \text{otherwise} \end{cases}$$
for $i \neq j$ with $w_{ii} = 0$.

2.
$$w_{ij} = \begin{cases} 1 & d_{ij} < \delta \\ 0 & \text{otherwise} \end{cases}$$
for $i \neq j$ with $w_{ii} = 0$.

3. $w_{ij} = d_{ij}^{-\alpha}$ for some $\alpha > 0$ with $w_{ii}$ defined to be 0.

These all yield symmetric $\mathbf{W}$ matrices but other schemes producing nonsymmetric matrices are allowed. There are some constraints, however (more on that below).

The $SAR$ model can be rewritten as

$$Z(\mathbf{s}_i) = \mathbf{x}(\mathbf{s}_i)'\boldsymbol{\beta} + \rho \sum_{j \in \mathcal{N}_i} w_{ij} \left[ Z(\mathbf{s}_j) - \mathbf{x}(\mathbf{s}_j)'\boldsymbol{\beta} \right] + v(\mathbf{s}_i)$$

for $i = 1, \cdots, n$ and $\mathcal{N}_i = \{j : j \text{ is a neighbor of } i\}$. This equation shows more clearly how the model uses information surrounding regions of a lattice. Our text also notes that we can express the model in matrix form as follows:

$$\begin{aligned} \mathbf{Z}(\mathbf{s}) &= \mathbf{X}(\mathbf{s})\boldsymbol{\beta} + (\mathbf{I} - \rho\mathbf{W})^{-1}\mathbf{v} \\ &= \mathbf{X}(\mathbf{s})\boldsymbol{\beta} - \rho\mathbf{W}\mathbf{X}(\mathbf{s})\boldsymbol{\beta} + \rho\mathbf{W}\mathbf{Z}(\mathbf{s}) + \mathbf{v}. \end{aligned}$$

The first equation shows how the $SAR$ model induces a spatial correlation structure in the linear model. The second makes clear that the $SAR$ model is composed of a linear model with uncorrelated errors (independent under normality) ($\mathbf{Z}(\mathbf{s}) = \mathbf{X}(\mathbf{s})\boldsymbol{\beta} + \mathbf{v}$) plus 2 additional spatially lagged variables:

1. the spatially lagged covariates: $\rho\mathbf{W}\mathbf{X}(\mathbf{s})\boldsymbol{\beta}$
2. the spatially lagged responses: $\rho\mathbf{W}\mathbf{Z}(\mathbf{s})$.

This model reduces to the uncorrelated errors linear model in the absence of spatial correlation ($\rho = 0$).

Constraints are needed on $\rho$ to insure that $(\mathbf{I} - \rho\mathbf{W})$ is nonsingular. Let $\lambda_1 < \lambda_2 < \cdots < \lambda_n$ be the ordered eigenvalues of $\mathbf{W}$ with $\lambda_1 < 0$ and $\lambda_n > 0$. Then it can be shown that $1/\lambda_1 < \rho < 1/\lambda_n$. Asymptotically on square lattices ($n \to \infty$) we have $-0.25 < \rho < 0.25$. The actual constraints could be checked by finding the eigenvalues of $\mathbf{W}$. If the rows of $\mathbf{W}$ are standardized so that they all sum to 1 then $\rho < 1$ (although it can be less than $-1$ so it is not correct to think of $\rho$ as a correlation).

The above model is called a single parameter $SAR$ model. Multi-parameter versions also exist, i.e. we could have

$$\mathbf{B} = \sum_{i=1}^{k} \rho_i \mathbf{W}_i$$

where $\mathbf{W}_i, i = 1, \cdots, k$ specifies neighbors at different distances. We will only consider the single parameter version here.

*Estimation and Inference*: We assume that $\mathbf{B} = \rho\mathbf{W}$ and $\boldsymbol{\Sigma}_v = \sigma^2\mathbf{I}$ from which it follows that

$$\boldsymbol{\Sigma}_{SAR} = \sigma^2 (\mathbf{I} - \rho\mathbf{W})^{-1} \left( \mathbf{I} - \rho\mathbf{W}' \right)^{-1}.$$

Assuming that $\mathbf{\Sigma}_{SAR}$ is known the generalized least squares estimators of $\boldsymbol{\beta}$ and $\sigma^2$ are

$$\begin{aligned}\widehat{\boldsymbol{\beta}}_{gls} &= \left(\mathbf{X}(\mathbf{s})' \mathbf{\Sigma}_{SAR}^{-1} \mathbf{X}(\mathbf{s})\right)^{-1} \mathbf{X}(\mathbf{s})' \mathbf{\Sigma}_{SAR}^{-1} \mathbf{Z}(\mathbf{s}) \\ \widehat{\sigma}^2 &= \frac{\left(\mathbf{Z}(\mathbf{s}) - \mathbf{X}(\mathbf{s}) \widehat{\boldsymbol{\beta}}_{gls}\right)' \mathbf{\Sigma}_{SAR}^{-1} \left(\mathbf{Z}(\mathbf{s}) - \mathbf{X}(\mathbf{s}) \widehat{\boldsymbol{\beta}}_{gls}\right)}{n-p}\end{aligned}$$

Estimation of $\rho$ would yield estimated generalized least squares estimators. A "least squares" estimator of $\rho$ is

$$\widehat{\rho} = \frac{\mathbf{Z}(\mathbf{s})' \mathbf{W}'\mathbf{W}\mathbf{Z}(\mathbf{s})}{\mathbf{Z}(\mathbf{s})' \mathbf{W}'\mathbf{W}^2\mathbf{Z}(\mathbf{s})}.$$

Maximum likelihood estimators for the normal case:

$$\mathbf{Z}(\mathbf{s}) \sim N\left(\mathbf{X}(\mathbf{s})\boldsymbol{\beta}, (\mathbf{I}-\mathbf{B})^{-1} \mathbf{\Sigma}_{\mathbf{v}} (\mathbf{I}-\mathbf{B})^{-1}\right)$$

are

$$\begin{aligned}\widehat{\boldsymbol{\beta}}_{ml} &= \left(\mathbf{X}(\mathbf{s})' \mathbf{\Sigma}_{SAR}\left(\widehat{\boldsymbol{\theta}}_{ml}\right)^{-1} \mathbf{X}(\mathbf{s})\right)^{-1} \mathbf{X}(\mathbf{s})' \mathbf{\Sigma}_{SAR}\left(\widehat{\boldsymbol{\theta}}_{ml}\right)^{-1} \mathbf{Z}(\mathbf{s}) \\ \widehat{\sigma}^2_{ml} &= (1/n)\left(\mathbf{Z}(\mathbf{s}) - \mathbf{X}(\mathbf{s})\widehat{\boldsymbol{\beta}}_{ml}\right)' \mathbf{\Sigma}_{SAR}\left(\widehat{\boldsymbol{\theta}}_{ml}\right)^{-1} \left(\mathbf{Z}(\mathbf{s}) - \mathbf{X}(\mathbf{s})\widehat{\boldsymbol{\beta}}_{ml}\right)\end{aligned}$$

We can assume a more general form for $\mathbf{\Sigma}_{SAR}$:

$$\mathbf{\Sigma}_{SAR}(\boldsymbol{\theta}) = \sigma^2 \mathbf{V}(\boldsymbol{\theta})_{SAR}$$

where $\boldsymbol{\theta}$ contains the $b_{ij}$'s, and parameters in $\mathbf{\Sigma}_{\mathbf{v}}$. Of course, estimation of all these parameters would not be feasible in general so the practical assumptions of $\mathbf{B} = \rho\mathbf{W}$ with $\mathbf{W}$ known and $\mathbf{V}$ being known are often made resulting in $\boldsymbol{\theta} = (\sigma^2, \rho)'$. $REML$ estimators also exist.

– $CAR$ Models: We assume that $Z(\mathbf{s}_i)$ is dependent on a set of defined neighbors, i.e. $Z(\mathbf{s}_i)$ depends on $Z(\mathbf{s}_j)$ only if $\mathbf{s}_j \in \mathcal{N}_i$. However, unlike the $SAR$ model, which is a model of the joint probability distribution of $Z(\mathbf{s}_i), i = 1, \cdots, n$, the $CAR$ model is a model of the conditional distribution of $Z(\mathbf{s}_i)$ given the observation of $Z$'s in the set of neighbors. A $CAR$ model specifies the conditional mean and variance:

$$\begin{aligned}E\left[Z(\mathbf{s}_i) | \mathbf{Z}(\mathbf{s}_{-i})\right] &= \mathbf{x}(\mathbf{s}_i)' \boldsymbol{\beta} + \sum_{j=1}^{n} c_{ij}\left[Z(\mathbf{s}_j) - \mathbf{x}(\mathbf{s}_j)' \boldsymbol{\beta}\right] \\ \mathrm{Var}\left[Z(\mathbf{s}_i) | \mathbf{Z}(\mathbf{s}_{-i})\right] &= \sigma_i^2\end{aligned}$$

with $c_{ij} \neq 0$ for $j \in \mathcal{N}_i$ and $c_{ii} = 0$. (Restricted) maximum likelihood estimation, the usual approach(es), requires the joint distribution and care is needed in defining the conditional distributions in such a way that a valid joint distribution exists. The conditions required for a valid joint Gaussian distribution given conditional Gaussian distributions are fairly mild and this is the standard assumption. The resulting joint distribution is Gaussian with mean $\mathbf{X}(\mathbf{s})\boldsymbol{\beta}$ and variance

$$\mathbf{\Sigma}_{CAR} = (\mathbf{I}-\mathbf{C})^{-1} \mathbf{\Sigma}_c$$

with $\mathbf{\Sigma}_c = \sigma_i^2\mathbf{I}, i = 1, \cdots, n$. An added complication here is that constraints are needed to insure symmetry of the covariance matrix (otherwise it is not a covariance matrix). Those constraints are $\sigma_j^2 c_{ij} = \sigma_i^2 c_{ji}$. Two common simplifying assumptions about the structure of the covariance matrix is $\mathbf{C} = \rho\mathbf{W}$ and $\mathbf{\Sigma}_c = \sigma^2\mathbf{I}$. Generalized least squares estimators of $\boldsymbol{\beta}$ and $\sigma^2$ exist and have the same form as the $gls$ estimators under the $SAR$ model with $\Sigma_{CAR}$ replacing $\Sigma_{SAR}$. The estimator of $\rho$ is

$$\widehat{\rho} = \frac{\widehat{\mathbf{e}}_{ols}' \mathbf{W} \widehat{\mathbf{e}}_{ols}}{\widehat{\mathbf{e}}_{ols}' \mathbf{W}^2 \widehat{\mathbf{e}}_{ols}}$$

where $\widehat{\mathbf{e}}_{ols}$ is the residual vector from *ordinary* least squares estimation.

Maximum likelihood estimation, under an assumption of a Gaussian joint distribution, proceeds as seen with $SAR$ models except that $\mathbf{\Sigma}_{CAR}(\boldsymbol{\theta})$ replaces $\mathbf{\Sigma}_{SAR}(\theta)$ in the objective functions.

– $SAR$ vs. $CAR$: $SAR$ and $CAR$ models differ in how they specify the interactions among data in a lattice process. $SAR$ models imply that all data interact simulataneoulsy and $CAR$ models imply that a particular response is conditioned on the immediate neighbors. The $CAR$ models are spatial analogs of autogressive or $AR$ models in time series. The 2 approaches lead to different joint distributions for the responses. Assuming normality, and other assumptions specified above, the joint distribution in a $SAR$ model is

$$\mathbf{Z}(\mathbf{s}) \sim N\left(\mathbf{X}(\mathbf{s})\boldsymbol{\beta}, (\mathbf{I}-\mathbf{B})^{-1}\mathbf{\Sigma}_{\mathbf{v}}\left(\mathbf{I}-\mathbf{B}^{'}\right)^{-1}\right)$$

and the joint distribution under a $CAR$ model is

$$\mathbf{Z}(\mathbf{s}) \sim N\left(\mathbf{X}(\mathbf{s})\boldsymbol{\beta}, (\mathbf{I}-\mathbf{C})^{-1}\mathbf{\Sigma}_{c}\right).$$

The 2 models look very similar and indeed it can be shown that any $SAR$ model has a representation as a $CAR$ model. The converse is not true in general, however. Which is better? There is not a single answer for that. The error terms in the $SAR$ model are not independent of $\mathbf{Z}(\mathbf{s})$ whereas they are in the $CAR$ model making the parameters in $CAR$ models easier to estimate and more interpretable. Some have argued that it is more reasonable to assume that spatial lattice data arose from a Markov type process than from the spatial process implied by the $SAR$ model. Such an assumption also links $CAR$ models more closely to models typically applied in time series analysis.

– *New York Leukemia Example: The same* $\mathbf{W}$ *matrix will be used in both the* $SAR$ *and* $CAR$ *models with the queen's definition of contiguity. Thus, the elements of* $\mathbf{W}$ *are*

$$w_{ij} = \begin{cases} 1 & \textit{regions } i \textit{ and } j \textit{ share any portion of a border} \\ 0 & \textit{otherwise} \end{cases}$$

*SAR and CAR models can be fit using functions available in the* `spdep` *library. We will assume that* $\Sigma_{\mathbf{v}} = \sigma_s^2\mathbf{I}$ *for the SAR model and* $\Sigma_c = \sigma_c^2\mathbf{I}$ *for the CAR model. Thus, we have*

$$\begin{aligned} \mathbf{\Sigma}_{SAR} &= \sigma_s^2(\mathbf{I}-\rho_s\mathbf{W})^{-1}(\mathbf{I}-\rho_s\mathbf{W})^{-1} \\ \mathbf{\Sigma}_{CAR} &= \sigma_c^2(\mathbf{I}-\rho_c\mathbf{W})^{-1} \end{aligned}$$

*We will fit weighted and unweighted models using the* `spautolm` *function in* `spdep`*. The New York leukemia data set is included with the* `spdep` *documentation which is nice because it is not easy to set up* $\mathbf{W}$ *when the data are on an irregular lattice. The 4 models were fit with the following commands.*

```
require(spdep)
# access the data
data(NY_data) # unpacks the data
listw.NY<-listw_NY # done for convenience
leukSAR <- spautolm(z ~ pexp + age65 + home, data=leuk.dat,
+ listw=listw.NY, family="SAR", method="full")
leukCAR <- spautolm(z ~ pexp + age65 + home, data=leuk.dat,
+ listw=listw.NY, family="CAR", method="full")
leukSAR.wt<-spautolm(z ~ pexp + age65 + home,
+ data=leuk.dat,weights=leuk.dat$pop,
+ listw=listw.NY, family="SAR", method="full")
leukCAR.wt<-spautolm(z ~ pexp + age65 + home,
+ data=leuk.dat,weights=leuk.dat$pop,
+ listw=listw.NY, family="CAR", method="full")
# method specifies use of the full weights matrix
# listw.NY contains the weights that induce the correlation structure
```

*Specifying the matrix of neighbor weights is clearly crucial. We take the default here but in practice a good deal of thought should be given as to how to do this. The object* `listw.NY` *contains the information. It is a list with 3 components*

1. `style`*: Set to B in this case indicating that the weights in this matrix will be* 0 *or* 1*.*
2. `neighbours`*: Summary information on the neighbors.*

```
listw.NY$neighbours
Neighbour list object:
Number of regions: 281
Number of nonzero links: 1522
Percentage nonzero weights: 1.927534
Average number of links: 5.41637
```

3. `weights`*: This is a large file but it provides information on the number of neighbors.*

```
 listw.NY$weights
[[1]]
[1] 1 1 1 1 1 1 1 1

[[2]]
[1] 1 1 1 1 1 1

[[3]]
[1] 1 1 1

[[4]]
[1] 1 1 1

[[5]]
[1] 1 1 1 1
```

*So, for example, lattice unit 5 has 4 neighbors.*

*We can use the function* `listw2mat` *to get the actual weight matrix.*

```
a<-listw2mat(listw.NY)
a[1:5,1:5]
            [,1] [,2] [,3] [,4] [,5]
36007000100    0    1    0    0    0
36007000200    1    0    1    0    0
36007000300    0    1    0    0    0
36007000400    0    0    0    0    1
36007000500    0    0    0    1    0
```

*Above we saw that unit 5 (*36007000500*) had 4 neighbors and the sum of the numbers in row 5 of this matrix is indeed 4.*

```
apply(a,1,sum)[5]
36007000500
          4
```

*The results are shown in the tables below along with an AIC comparison. The AIC comparison also includes the ordinary and weighted least squares solutions from earlier. All AIC's are based on maximum likelihood fits. The* `spautolm` *function does not allow for REML fits.*

1. *Unweighted SAR Model:*

```
summary(leukSAR,adj.se=T)

Coefficients:
```

| | Estimate | Std. Error | t value | Pr(> \|z\|) |
|---|---|---|---|---|
| (Intercept) | −0.618193 | 0.178055 | −3.4719 | 0.0005168 |
| pexp | 0.071014 | 0.042353 | 1.6767 | 0.0935997 |
| age65 | 3.754200 | 0.629216 | 5.9665 | $2.424e-09$ |
| home | −0.419890 | 0.192706 | −2.1789 | 0.0293380 |

Lambda: 0.040487
LR test value: 5.2438
p-value: 0.022026

Log likelihood: −276.1069
Residual variance (sigma squared): 0.41985, (sigma: 0.64796)
Number of observations: 281
Number of parameters estimated: 6
AIC: 564.21

*The table of coefficient estimates and standard errors is (hopefully) self-explanatory. The estimate of* $\rho$ *is* 0.041*. The results of a likelihood ratio test of the null hypothesis* $\rho = 0$ *is presented, also. The estimate of* $\sigma^2$ *is* 0.420 *and* $AIC = 564.21$*.*

2. *Unweighted CAR Model:*

```
summary(leukCAR,adj.se=T)
```

Coefficients:

| | Estimate | Std. Error | t value | Pr(> \|z\|) |
|---|---|---|---|---|
| (Intercept) | −0.648362 | 0.182432 | −3.5540 | 0.0003794 |
| pexp | 0.077899 | 0.044006 | 1.7702 | 0.0766957 |
| age65 | 3.703830 | 0.631697 | 5.8633 | $4.538e-09$ |
| home | −0.382789 | 0.196971 | −1.9434 | 0.0519706 |

Lambda: 0.084123
LR test value: 5.8009
p-value: 0.016018

Log likelihood: −275.8283
Residual variance (sigma squared): 0.41346, (sigma: 0.64301)
Number of observations: 281
Number of parameters estimated: 6
AIC: 563.66

3. *Weighted SAR Model:*

```
summary(leukCAR.w,adj.se=T)
```

Coefficients:

| | Estimate | Std. Error | t value | Pr(> \|z\|) |
|---|---|---|---|---|
| (Intercept) | −0.797063 | 0.145090 | −5.4936 | $3.939e-08$ |
| pexp | 0.080545 | 0.028537 | 2.8224 | 0.004766 |
| age65 | 3.816731 | 0.580181 | 6.5785 | $4.752e-11$ |
| home | −0.380778 | 0.157633 | −2.4156 | 0.015709 |

Lambda: 0.0095636
LR test value: 0.32665
p-value: 0.56764

Log likelihood: −251.6017
Residual variance (sigma squared): 1120.1, (sigma: 33.468)
Number of observations: 281
Number of parameters estimated: 6
AIC: 515.2

*4. Weighted CAR Model:*

```
summary(leukCAR.wt,adj.se=T)

Coefficients:
```

| | Estimate | Std. Error | t value | Pr(> \|z\|) |
|---|---|---|---|---|
| (Intercept) | −0.790154 | 0.145904 | −5.4156 | $6.109e-08$ |
| pexp | 0.081922 | 0.028799 | 2.8446 | 0.004446 |
| age65 | 3.825858 | 0.581876 | 6.5750 | $4.864e-11$ |
| home | −0.386820 | 0.158568 | −2.4395 | 0.014710 |

```
Lambda:  0.022419
LR test value:  0.38785
p-value:  0.53343

Log likelihood:  −251.5711
Residual variance (sigma squared):  1118.8, (sigma:  33.449)
Number of observations:  281
Number of parameters estimated:  6
AIC: 515.14
```

*We can compare the results from these spatial models with the $OLS$ and $WLS$ results. The key parameter is $\beta_1$. The table below shows $\widehat{\beta}_1$, $SE\left(\widehat{\beta}_1\right)$, and the $AIC$ values from the models.*

| *Model* | $\widehat{\beta}_1$ | $SE\left(\widehat{\beta}_1\right)$ | $AIC$ | $\Delta AIC$ |
|---|---|---|---|---|
| $OLS$ | 0.049 | 0.0352 | 567.46 | 49.93 |
| $WLS$ | 0.076 | 0.0273 | 513.53 | 0 |
| $SAR$ | 0.071 | 0.0423 | 564.21 | 50.68 |
| $CAR$ | 0.078 | 0.044 | 563.66 | 50.13 |
| $SAR - Wt$ | 0.081 | 0.0285 | 515.20 | 1.67 |
| $CAR - Wt$ | 0.082 | 0.0288 | 515.14 | 1.61 |

*The results for the parameter $\rho$ in the autoregression models are shown below.*

| *Model* | $\hat{\rho}$ | *p-value* |
|---|---|---|
| $SAR$ | 0.0405 | 0.022 |
| $CAR$ | 0.0841 | 0.016 |
| $SAR$ - $Wt$ | 0.0096 | 0.568 |
| $CAR$ - $Wt$ | 0.0224 | 0.533 |

*We note that the test of $\rho = 0$ yields low p-values in the unweighted models and high p-values in the weighted models. Thus we see evidence of spatial autocorrelation among the residuals in the unweighted autogregression models but none in the weighted models. This suggests that the spatial autocorrelation evident in the unweighted fits is actually due to heteroscedasticity (and the 3 outliers would certainly contribute to that). These results are also consistent with the generalized least squares approach when we compared semivariograms from the unweighted OLS model with the weighted WLS model.*

*There are other tests available for the null hypothesis of $\rho = 0$ including a large sample test (Wald test) and a test based on Moran's $I$. I have been unable to find any implementation of these in R although the test based on Moran's I would be easy to get (anyone still looking for a short presentation topic?).*

– *Conclusion*: Waller and Gotway have some nice summary comments on spatial regression and autoregression models. $SAR$ and $CAR$ models are typically fit to lattice data. Generalized least squares models can also be fit to such data with spatial correlation stuctures modeled using geostatistical ideas, or at least covariance functions borrowed from geostatistics. Thus, spatial autoregression models differ from generalized least squares models in how the spatial correlation is modeled. Spatial dependence in autoregression is handled via the spatial proximity matrix $\mathbf{W}$.

Waller and Gotway provide some general rules of thumb borrowed from Griffith (1996):

1. Any reasonable method of modelling spatial correlation is better than ignoring it completely.
2. Different spatial dependence models can give different results, however. Exploratory analysis is crucial and different approaches should be tried and compared.
3. Inference is often based on large sample results. In the presence of spatial correlation "large" means larger than we might normally think (recall the effective sample size discussions we had earlier in the semester). Waller and Gotway suggest that "spatial correlation reduces the information contained in a sample of independent data by a factor of 2".
4. It is important to account for heterogeneity in lattice data. Such heterogeneity can arise as a result of different regions having different population sizes.
5. Keep it simple, i.e. as a general rule choose the simplest model that explains variability in the data adequately. And remember - interpretation is a part of the principle of parsimony.

As a final caveat Bivand, Pebesma, and Gomez-Rubio caution; "Although there may be computing environments within which it seems easier to fit spatial regression models, arguably few give the analyst both reasonable defaults and the oppurtunity to examine in as much detail as is needed the internal workings of the methods used, and of their implementations in software."

- This is a large family of models that extend the concepts of linear regression models to those settings in which the standard assumptions are not met. Logistic regression and Poisson regression are 2 of the more common examples of generalized linear models.

  A *GLM* has 3 components:

  1. A systematic component: a linear combination of covariates or predictor variables.
  2. A random component: a proposed distribution for the responses.
  3. A link function: a function that specifies the *link* between the systematic and random components. In particular, it specifies how $\mu = E(Z)$ is related to the linear predictor (systematic component):

$$g(\mu) = \beta_0 + \beta_1 X_1 + \cdots + \beta_{p-1} X_{p-1}.$$

  We will look briefly at these in a bit more detail. For convenience I am going to suppress the spatial notation by writing $Z$ and $\mathbf{X}$ instead of $Z(\mathbf{s})$ and $\mathbf{X}(\mathbf{s})$. That makes the notation a bit cleaner.

  - *Random Component*: We have independent responses $Z_i, i = 1, \cdots, n$. The distribution of these can be anything but a common assumption, and one which much of the theory is based on, is that the distribution is a member of the *exponential family*. If $Z$ is a random variable whose probability distribution is a member of the exponential family of distributions then its pdf can be written

$$f(z, \boldsymbol{\theta}) = \exp\left[a(\boldsymbol{\theta}) + b(z) + zQ(\boldsymbol{\theta})\right]$$

    The joint distribution of a random sample is then

$$f(z_1, \cdots, z_n, \boldsymbol{\theta}) = \exp\left[\sum_{i=1}^{n} a(\boldsymbol{\theta}) + \sum_{i=1}^{n} b(z_i) + \sum_{i=1}^{n} z_i Q(\boldsymbol{\theta})\right].$$

    The binomial, poisson, and normal families of distributions are all members of the exponential family of distributions.

    *Example: Suppose $Z \sim Bin(n, p)$ with*

$$f(z, p) = \binom{n}{z} p^z (1-p)^{n-z}, z = 0, \cdots, n.$$

    *We can rewrite this as*

$$f(z, p) = \exp\left[n \log(1-p) + \log\binom{n}{z} + z \log\left(\frac{p}{1-p}\right)\right]$$

    *which is in exponential family form with*

$$\begin{aligned} a(p) &= n \log(1-p) \\ b(z) &= \log\binom{n}{z} \\ Q(p) &= \log\left(\frac{p}{1-p}\right) \end{aligned}$$

  - *Systematic Component*: A set of covariates or predictor variables is of interest and a linear combination of those, $\beta_0 + \beta_1 X_1 + \cdots + \beta_{p-1} X_{p-1}$, makes up the systematic component.
  - *Link Function*: The link function $g$ describes the relationship between the mean $\mu = E[Z]$ and the systematic component. In particular

$$g(\mu) = \beta_0 + \beta_1 X_1 + \cdots + \beta_{p-1} X_{p-1}.$$

The simplest link function is the identity link

$$g(\mu) = \mu = \beta_0 + \beta_1 X_1 + \cdots + \beta_{p-1} X_{p-1}$$

which is the link function used in the regression models in STAT 410/412 and STAT 505/506. Every member of the exponential family of distributions has a special link function called the *natural* or *canonical* link function.

- *Logistic Regression*: We have a binary response variable $Z$:

$$Z = \begin{cases} 1 & \text{success} \\ 0 & \text{failure} \end{cases}$$

We assume that the probability of a success is a function of covariates or predictors $[X_1, \cdots, X_{p-1}]'$,

$$P[Z|X_1, \cdots, X_{p-1}] = \mu = \frac{\exp(\beta_0 + \beta_1 X_1 + \cdots + \beta_{p-1} X_{p-1})}{1 + \exp(\beta_0 + \beta_1 X_1 + \cdots + \beta_{p-1} X_{p-1})}.$$

We want to express this as a linear model and it is not hard to show that

$$g(\mu) = \log\left(\frac{\mu}{1-\mu}\right) = \beta_0 + \beta_1 X_1 + \cdots + \beta_{p-1} X_{p-1}.$$

This is called the logit link and binary regression with a logit link is called logistic regression. The logit link is the canonical link. Note that the odds

$$\frac{\mu}{1-\mu} = \exp(\beta_0 + \beta_1 X_1 + \cdots + \beta_{p-1} X_{p-1}).$$

Parameter interpretation is not as easy as in linear regression. In terms of the coefficient $\beta_k$ we can see that the ratio of the odds of a success at $X_k + 1$ to the odds at $X_k$ is $\exp(\beta_k)$ because

$$\frac{\left(\frac{\mu(X_k+1)}{1-\mu(X_k+1)}\right)}{\left(\frac{\mu(X_k)}{1-\mu(X_k)}\right)} = \frac{\exp(\beta_0 + \beta_1 X_1 + \cdots + \beta_k(X_k+1) + \cdots + \beta_{p-1} X_{p-1})}{\exp(\beta_0 + \beta_1 X_1 + \cdots + \beta_k X_k + \cdots + \beta_{p-1} X_{p-1})} = \exp(\beta_k).$$

Thus, the odds of a success at $X_k + 1$ is $\exp(\beta_k)$ times the odds of a success at $X_k$ if all other variables are held constant.

Logistic regression models can be fit to binary responses or to binomial counts. We will take a quick look at both.

- *Poisson Regression*: The response $Z$ is a Poisson count. We assume that the response is a function of covariates or predictors with a log link:

$$\log(E[Z|X_1, \cdots, X_{p-1}]) = \log(\mu) = \beta_0 + \beta_1 X_1 + \cdots + \beta_{p-1} X_{p-1}.$$

The log link is the canonical link for the Poisson family.

Noting that

$$\mu = \exp(\beta_0 + \beta_1 X_1 + \cdots + \beta_{p-1} X_{p-1})$$

we can see that the mean count at $X_k + 1$ is $\exp(\beta_k)$ times the mean count at $X_k$ if all other variables are held constant.

- Maximum likelihood estimation is generally used to estimate parameters in $GLM$'s. Software for fitting these models is readily available. The `glm` function in R will fit both logistic and Poisson regression models. Inference for regression coefficients typically proceeds under an assumption of asymptotic normality. Likelihood ratio tests or information theoretic based comparisons are typically used in model building.

- *Overdispersion*: It is not uncommon to find that the variability of the response in logistic and Poisson regression models is greater than expected under the assumption of binomial or Poisson count data. This problem is referred to as *overdispersion*. Heterogeneity among observations is the most common cause of overdispersion, although outliers can also cause the problem. Overdispersion cannot be assessed with ungrouped binary data in logistic regression, something which not many people seem to understand.

- *Rate Models:* When event counts are recorded over time or space it may be more appropriate to model the rate at which events occur. For example, $Z$ may be the number of diseased individuals recorded over a time period $t$. A loglinear model for the mean rate is

$$\log\left(\frac{\mu}{t}\right) = \log\mu - \log t = \beta_0 + \beta_1 X_1 + \cdots + \beta_{p-1}X_{p-1}.$$

  The adjustment $\log t$ is called the *offset*. Note that the model being fit is a Poisson regression model with a predictor $\log t$ that has a known coefficient of 1. Most software can handle offsets including R and SAS.

- We will be restricting ourselves to logistic and Poisson regression models. We will be working in a spatial context so let $\mathbf{Z} = \left[Z\left(\mathbf{s}_1\right), \cdots, Z\left(\mathbf{s}_n\right)\right]'$. In many applications, $Z_i = Z\left(\mathbf{s}_i\right)$ will be a regional response, i.e. lattice data. We assume that a function of the mean vector $\boldsymbol{\mu} = \left[\mu\left(\mathbf{s}_1\right), \cdots, \mu\left(\mathbf{s}_n\right)\right]'$ is linearly related to a set of covariates:

$$g\left(\boldsymbol{\mu}\right) = g\left(E\left[\mathbf{Z}|\mathbf{X}\right]\right) = \mathbf{X}\boldsymbol{\beta}.$$

  The variance is often a function of the mean in generalized linear models and this relationship is quantified in the variance function $v\left(\mu\right)$. For Poisson data with a log link we have $v(\mu) = \mu$ and for binary data with a logit link we have $v\left(\mu\right) = \mu\left(1-\mu\right)$. Under an assumption of independent observations we model

$$\mathrm{Var}\left(\mathbf{Z}\right) = \boldsymbol{\Sigma} = \sigma^2\mathbf{V}_{\boldsymbol{\mu}}$$

  where $\mathbf{V}_{\boldsymbol{\mu}}$ is a diagonal matrix with $v_{ii} = v\left(\mu\left(\mathbf{s}_i\right)\right)$. The parameter $\sigma^2$ is interpreted as a dispersion parameter which "allows for inexactness in the variance to mean relationships" as Waller and Gotway put it.

  In the presence of spatial autocorrelation we need to make adjustments to $\boldsymbol{\Sigma}$ similar to what we did in the ordinary regression models we have already seen. We model the covariance matrix as

$$\boldsymbol{\Sigma}\left(\boldsymbol{\theta}\right) = \sigma^2\mathbf{V}_{\boldsymbol{\mu}}^{1/2}\mathbf{R}\left(\boldsymbol{\theta}\right)\mathbf{V}_{\boldsymbol{\mu}}^{1/2}$$

  where $\mathbf{R}\left(\boldsymbol{\theta}\right)$ is the spatial correlation matrix with elements determined by an appropriately chosen spatial correlation function (or correlogram), $\rho\left(\mathbf{s}_i - \mathbf{s}_j; \boldsymbol{\theta}\right)$. The sill of a correlogram is necessarily equal to 1 so we will typically only need to estimate a range parameter and possibly a nugget. A nugget effect is incorporated as

$$\begin{aligned}\boldsymbol{\Sigma}\left(\boldsymbol{\theta}\right) &= c_0\mathbf{V}_{\boldsymbol{\mu}} + \sigma_1^2\mathbf{V}_{\boldsymbol{\mu}}^{1/2}\mathbf{R}\left(\boldsymbol{\theta}\right)\mathbf{V}_{\boldsymbol{\mu}}^{1/2} \\ &= \mathbf{V}_{\boldsymbol{\mu}}^{1/2}\left(c_0\mathbf{I} + \sigma_1^2\mathbf{R}\left(\boldsymbol{\theta}\right)\right)\mathbf{V}_{\boldsymbol{\mu}}^{1/2}\end{aligned}$$

  Thus, we have

$$\mathrm{Var}\left[Z\left(\mathbf{s}_i\right)\right] = \left(c_0 + \sigma_1^2\right)v\left(\mu\left(\mathbf{s}_i\right)\right)$$

  and

$$\mathrm{Cov}\left[Z\left(\mathbf{s}_i\right), Z\left(\mathbf{s}_j\right)\right] = \sigma_1^2\sqrt{v\left(\mu\left(\mathbf{s}_i\right)\right)v\left(\mu\left(\mathbf{s}_j\right)\right)}\left[\rho\left(\mathbf{s}_i - \mathbf{s}_j; \boldsymbol{\theta}\right)\right].$$

  *Example: Suppose we have 5 observations collected at 5 locations* $\mathbf{s}_i, i = 1, \cdots, 5$,

$$\mathbf{X} = \begin{bmatrix} X\left(\mathbf{s}_1\right) \\ X\left(\mathbf{s}_2\right) \\ X\left(\mathbf{s}_3\right) \\ X\left(\mathbf{s}_4\right) \\ X\left(\mathbf{s}_5\right) \end{bmatrix} = \begin{bmatrix} 2 \\ -3 \\ 3 \\ -4 \\ 2 \end{bmatrix}.$$

*We assume that the responses are counts to be modeled with a Poisson regression model with log link, thus:*

$$\log\left(E\left[\mathbf{Z}|\mathbf{X}\right]\right) = \begin{bmatrix} \log\left(\mu\left(\mathbf{s}_1\right)\right) \\ \log\left(\mu\left(\mathbf{s}_2\right)\right) \\ \log\left(\mu\left(\mathbf{s}_3\right)\right) \\ \log\left(\mu\left(\mathbf{s}_4\right)\right) \\ \log\left(\mu\left(\mathbf{s}_5\right)\right) \end{bmatrix} = \begin{bmatrix} 1 & 2 \\ 1 & -3 \\ 1 & 3 \\ 1 & -4 \\ 1 & 2 \end{bmatrix} \begin{bmatrix} \beta_0 \\ \beta_1 \end{bmatrix}.$$

*The variance function is* $v\left(\mu\right) = \mu$ *so*

$$\mathbf{V}_{\boldsymbol{\mu}}^{1/2} = \begin{bmatrix} \sqrt{\mu\left(\mathbf{s}_1\right)} & 0 & 0 & 0 & 0 \\ 0 & \sqrt{\mu\left(\mathbf{s}_2\right)} & 0 & 0 & 0 \\ 0 & 0 & \sqrt{\mu\left(\mathbf{s}_3\right)} & 0 & 0 \\ 0 & 0 & 0 & \sqrt{\mu\left(\mathbf{s}_4\right)} & 0 \\ 0 & 0 & 0 & 0 & \sqrt{\mu\left(\mathbf{s}_5\right)} \end{bmatrix}.$$

*The log link of* $\log\left(\mu\left(\mathbf{s}_i\right)\right) = \beta_0 + \beta_1 X\left(\mathbf{s}_i\right)$ *implies*

$$\mu\left(\mathbf{s}_i\right) = \exp\left(\beta_0 + \beta_1 X\left(\mathbf{s}_i\right)\right)$$

*so*

$$\mathbf{V}_{\boldsymbol{\mu}}^{1/2} = \begin{bmatrix} \sqrt{e^{(\beta_0+2\beta_1)}} & 0 & 0 & 00 & \\ 0 & \sqrt{e^{(\beta_0-3\beta_1)}} & 0 & 0 & 0 \\ 0 & 0 & \sqrt{e^{(\beta_0+3\beta_1)}} & 0 & 0 \\ 0 & 0 & 0 & \sqrt{e^{(\beta_0-4\beta_1)}} & 0 \\ 0 & 0 & 0 & 0 & \sqrt{e^{(\beta_0+2\beta_1)}} \end{bmatrix}.$$

*Suppose that the coordinates are* $\mathbf{s}_1 = (0,1), \mathbf{s}_2 = (1,1), \mathbf{s}_3 = (1,0), \mathbf{s}_4 = (0.5,-0.5), \mathbf{s}_5 = (-1,0)$ *and an exponential correlation function* $\rho\left(\mathbf{s}_i - \mathbf{s}_j; \boldsymbol{\theta}\right) = \exp\left(-\|\mathbf{s}_i - \mathbf{s}_j\|\right)$ *is specified, i.e. we set the range parameter equal to 1. The resulting correlation matrix is*

$$\mathbf{R} = \begin{bmatrix} 1 & 0.37 & 0.24 & 0.21 & 0.24 \\ 0.37 & 1 & 0.37 & 0.21 & 0.11 \\ 0.24 & 0.37 & 1 & 0.49 & 0.14 \\ 0.21 & 0.21 & 0.49 & 1 & 0.21 \\ 0.24 & 0.11 & 0.14 & 0.21 & 1 \end{bmatrix}$$

*Estimation*

- Maximum likelihood estimation of parameters requires specification of the joint distribution. This is easy in the normal errors models. Schabenberger and Gotway give a simple example (page 355) with a simple model

$$\mathbf{Z} = \boldsymbol{\mu} + \boldsymbol{\epsilon}$$

where $\boldsymbol{\epsilon} \sim N(\mathbf{0}, \sigma^2\mathbf{I})$. The marginal distributions of the responses are also normal. If we have a more general covariance structure $\boldsymbol{\epsilon} \sim N(\mathbf{0}, \boldsymbol{\Sigma})$ the marginal distributions are still normal. A valid model for the multivariate distribution leads to valid models for the marginal distribution.

In generalized linear models an assumption of independent observations means that the joint distribution is the product of the marginals and maximum likelihood estimation is possible. This is how `glm` in R estimates parameters in logistic and Poisson regression models. But we can no longer generalize in the presence of spatial correlation as we did in the normal case because there is (as Schabenberger and Gotway put it) "no claim made at this point that the underlying joint distributions may be a 'multivariate Binomial' distribution, or some such thing." According to Schabenberger and Gotway it is not clear that a joint distribution with the desired properties even exists.

There are ways around the problem all with their own particular set of baggage. We can specify so-called conditional or mixed models (Generalized Linear Mixed Models) or use Bayesian methods. The use of maximum likelihood methods for mixed models is still problematical and there does not appear to be a consensus in the literature on the best way to proceed. One way is via a pseudo-likelihood approach which can also be used to fit the marginal (fixed effects models) under discussion here. Bayesian methods, despite being the current "flavor of the month" in the ecological sciences, is a non-trivial exercise in modelling with considerable potential for problems.

Another approach is to forego distributional assumptions and rely only on the first and second moment properties of the responses - a method known as quasi-likelihood estimation. We will consider all of these below. We will look first at quasi-likelihood estimation and then consider pseudo-likelihood estimation for fixed effects models. Then we wil look at an example involving the use of these two methods to fit a fixed-effects or marginal model. After that we will consider the mixed model approach and finish up with a brief look at the Bayesian approach.

- *Quasi-Likelihood Estimation*: Quasi-likelihood estimation for marginally specified (i.e. fixed effect) models is based on information in the first 2 moments (mean and variance) bypassing any additional distributional assumptions. Let $\boldsymbol{\mu} = [\mu_1, \cdots, \mu_n]^{'}$ be the vector of means for the (regional) responses $Z_1, \cdots, Z_n$ with link $g(\boldsymbol{\mu}) = \mathbf{X}\boldsymbol{\beta}$. We define the quasi-likelihood function $Q$ to be a function such that

$$\frac{\partial Q(\boldsymbol{\mu}; \mathbf{Z})}{\partial \boldsymbol{\mu}} = \mathbf{V}^{-1}(\mathbf{Z} - \boldsymbol{\mu}).$$

  In a spatial context $\mathbf{V}$ captures the spatial correlation structure. Note that $\boldsymbol{\mu}$ and $\mathbf{V}$ are both functions of $\boldsymbol{\beta}$. It can be shown that, under suitable and fairly mild regularity conditions, $Q$ acts a lot like the derivative of a log-likelihood function. A set of quasi-likelihood score equations is generated by differentiating $Q$ with respect to each element of $\boldsymbol{\beta}$:

$$\frac{\partial Q}{\partial \boldsymbol{\beta}} = \frac{\partial Q}{\partial \boldsymbol{\mu}} \frac{\partial \boldsymbol{\mu}}{\partial \boldsymbol{\beta}} = \boldsymbol{\Delta}^{'} \mathbf{V}^{-1}(\mathbf{Z} - \boldsymbol{\mu}) = \mathbf{0}$$

  where $\boldsymbol{\Delta}$ is a matrix with elements $\delta_{ij} = \partial \mu_i / \partial \beta_j$. The quasi-likelihood estimator of $\boldsymbol{\beta}$ is the solution to this system of equations. The existence of a solution requires $\mathbf{V}$ satisfy certain regularity conditions which can be guaranteed by setting

$$\mathbf{V} = \sigma^2 \mathbf{V}_{\boldsymbol{\mu}}^{1/2} \mathbf{R}(\boldsymbol{\theta}) \mathbf{V}_{\boldsymbol{\mu}}^{1/2}$$

  where $\sigma^2$ is the dispersion parameter, $\mathbf{R}$ contains the correlations and $\mathbf{V}_{\boldsymbol{\mu}}$ is diagonal with scale parameters on the diagonal. The result is a set of *Generalized Estimating Equations*. Details of the estimation procedure are omitted although it is very similar to the iterative procedure for fitting generalized least squares models in the ordinary regression setting. The important point here is that in many if not most applicable practical settings quasi-likelihood produces parameter estimates with nice statistical properties.

- A second method of handling spatial correlation and overdispersion is *pseudo-likelihood* estimation. The term pseudolikelihood is used in a confusingly large number of ways in statistics. In the context in which we are working it is introduced as a method of producing estimates of parameters in generalized linear models with spatial correlation structures. Pseudolikelihood differs from quasi-likelihood in that true joint likelihoods are being used to estimate parameters at each step in the iterative process.

  A first order Taylor series expansion of $g$ about $\mu$ is used to create pseudo-data:

$$Z_i^{(p)} = g(\widehat{\mu}_i) + g^{'}(\widehat{\mu}_i)(Z_i - \widehat{\mu}_i)$$

  where $\widehat{\mu}$ is a current estimate of $\mu$. Given some not unreasonable assumptions

$$\begin{aligned} E\left[\mathbf{Z}^{(p)} | \boldsymbol{\beta}\right] &= \mathbf{X}\boldsymbol{\beta} \\ \text{Var}\left(\mathbf{Z}^{(p)} | \boldsymbol{\beta}\right) &= \boldsymbol{\Sigma}_{\widehat{\boldsymbol{\mu}}} \end{aligned}$$

  where

$$\boldsymbol{\Sigma}_{\widehat{\boldsymbol{\mu}}} = \sigma^2 \boldsymbol{\Delta}_{\widehat{\boldsymbol{\mu}}} \mathbf{V}_{\widehat{\boldsymbol{\mu}}}^{1/2} \mathbf{R} \mathbf{V}_{\widehat{\boldsymbol{\mu}}}^{1/2}$$

with $\boldsymbol{\Delta}_{\widehat{\boldsymbol{\mu}}}$ a diagonal matrix with diagonal elements $\delta_{ii} = \partial g(\mu_i)/\partial \mu_i$ evaluated at $\widehat{\boldsymbol{\mu}}$. The marginal mean and variance of the pseudo-data are

$$\begin{aligned} E\left[\mathbf{Z}^{(p)}\right] &= \mathbf{X}\boldsymbol{\beta} \\ \text{Var}\left(\mathbf{Z}^{(p)}\right) &= \boldsymbol{\Sigma}_{\widehat{\boldsymbol{\mu}}}. \end{aligned}$$

In other words, we have an ordinary linear regression model with spatially correlated errors which can be fit using generalized least squares. Define $\boldsymbol{\Sigma}_{\mathbf{Z}^{(p)}} = \boldsymbol{\Sigma}_{\widehat{\boldsymbol{\mu}}}$ and proceed as follows:

1. Compute an initial estimate of $\widehat{\boldsymbol{\mu}}$ using a nonspatial $GLM$.
2. Compute the pseudodata.
3. Use maximum likelihood (or $REML$) with the pseudodata to get estimates of the spatial correlation parameters and $\sigma^2$.
4. Use these estimates to get generalized least squares estimates of $\boldsymbol{\beta}$ and $\sigma^2$:

$$\begin{aligned} \widehat{\boldsymbol{\beta}} &= \left(\mathbf{X}\boldsymbol{\Sigma}^{-1}_{\mathbf{Z}^{(p)}}\mathbf{X}\right)^{-1}\mathbf{X}'\boldsymbol{\Sigma}^{-1}_{\mathbf{Z}^{(p)}}\mathbf{Z}^{(p)} \\ \widehat{\sigma}^2 &= \frac{1}{n}\left(\mathbf{Z}^{(p)} - \mathbf{X}\boldsymbol{\beta}\right)'\boldsymbol{\Sigma}^{-1}_{\mathbf{Z}^{(p)}}\left(\mathbf{Z}^{(p)} - \mathbf{X}\boldsymbol{\beta}\right) \end{aligned}$$

5. Use the inverse link function to update $\widehat{\boldsymbol{\mu}}$:

$$\widehat{\boldsymbol{\mu}} = g^{-1}\left(\mathbf{X}\widehat{\boldsymbol{\beta}}\right).$$

6. Repeat until convergence.

The standard errors for the estimated regression coefficients can be approximated as

$$\widehat{\text{Var}}\left(\widehat{\boldsymbol{\beta}}\right) = \widehat{\sigma}^2\left(\mathbf{X}'\widehat{\boldsymbol{\Sigma}}^{-1}_{\mathbf{Z}^{(p)}}\mathbf{X}\right)^{-1}$$

where $\widehat{\boldsymbol{\Sigma}}^{-1}_{\mathbf{Z}^{(p)}}$ is defined to be $\boldsymbol{\Sigma}^{-1}_{\mathbf{Z}^{(p)}}$ using the final converged solution of the spatial correlation parameters $\widehat{\boldsymbol{\theta}}$.

- *Example: We continue with our earlier example incorporating an overdispersion parameter $\sigma^2$ and an unknown range parameter $a$. For the marginal model $\log(\boldsymbol{\mu})$ and $\mathbf{V}^{1/2}_{\boldsymbol{\mu}}$ are unchanged but we now have*

$$\mathbf{R}(a) = \begin{bmatrix} 1 & \exp(-1.0/a) & \exp(-1.4/a) & \exp(-1.6/a) & \exp(-1.4/a) \\ \exp(-1.0/a) & 1 & \exp(-1.0/a) & \exp(-1.6/a) & \exp(-2.2/a) \\ \exp(-1.4/a) & \exp(-1.0/a) & 1 & \exp(-0.7/a) & \exp(-2.0/a) \\ \exp(-1.6/a) & \exp(-1.6/a) & \exp(-0.7/a) & 1 & \exp(-1.6/a) \\ \exp(-1.4/a) & \exp(-2.2/a) & \exp(-2.0/a) & \exp(-1.6/a) & 1 \end{bmatrix}$$

*The link function is $g(\mu) = \log(\mu)$ so $g'(\mu) = 1/\mu$ and with $\mu = \exp(\beta_0 + \beta_1 X)$ we get*

$$\boldsymbol{\Delta}_{\boldsymbol{\mu}} = \begin{bmatrix} e^{-(\beta_0+2\beta_1)} & 0 & 0 & 0 & 0 \\ 0 & e^{-(\beta_0-3\beta_1)} & 0 & 0 & 0 \\ 0 & 0 & e^{-(\beta_0+3\beta_1)} & 0 & 0 \\ 0 & 0 & 0 & e^{-(\beta_0-4\beta_1)} & 0 \\ 0 & 0 & 0 & 0 & e^{-(\beta_0+2\beta_1)} \end{bmatrix}.$$

*The pseudodata are*

$$\mathbf{Z}^{(p)} = \begin{bmatrix} 1 & 2 \\ 1 & -3 \\ 1 & 3 \\ 1 & -4 \\ 1 & 2 \end{bmatrix}\begin{bmatrix} \beta_0 \\ \beta_1 \end{bmatrix} + \boldsymbol{\Delta}_{\boldsymbol{\mu}}\left(\begin{bmatrix} Z(\mathbf{s}_1) \\ Z(\mathbf{s}_2) \\ Z(\mathbf{s}_3) \\ Z(\mathbf{s}_4) \\ Z(\mathbf{s}_5) \end{bmatrix} - \begin{bmatrix} \exp(\beta_0 + 2\beta_1) \\ \exp(\beta_0 - 3\beta_1) \\ \exp(\beta_0 + 3\beta_1) \\ \exp(\beta_0 - 4\beta_1) \\ \exp(\beta_0 + 2\beta_1) \end{bmatrix}.\right)$$

*The pseudodata depend only on $\boldsymbol{\mu}$ (i.e. only on $\boldsymbol{\beta}$) so, given estimates of $\boldsymbol{\beta}$ the pseudodata become "responses" in a linear regression model with covariance matrix*

$$\boldsymbol{\Sigma}_{\mathbf{Z}^{(p)}} = \sigma^2 \boldsymbol{\Delta}_{\boldsymbol{\mu}} \mathbf{V}_{\boldsymbol{\mu}}^{1/2} \mathbf{R}(a) \mathbf{V}_{\boldsymbol{\mu}}^{1/2} \boldsymbol{\Delta}_{\boldsymbol{\mu}}.$$

*So we*

1. *Estimate $\beta_0$ and $\beta_1$ using a nonspatial GLM.*
2. *Compute the means $\mu_i = \exp(\beta_0 + \beta_1 Z_i)$*
3. *Evaluate $\boldsymbol{\Delta}_{\boldsymbol{\mu}}$, $\mathbf{V}_{\boldsymbol{\mu}}^{1/2}$, and $\mathbf{Z}^{(p)}$.*
4. *Use the pseudodata as data to get (restricted) maximum likelihood estimates of $\sigma^2$ and $a$.*
5. *Substitute $\widehat{\sigma}^2$ and $\widehat{a}$ into $\boldsymbol{\Sigma}_{\mathbf{Z}^{(p)}}$ to get generalized least squares estimates of $\beta_0$ and $\beta_1$*
6. *Iterate 2-5 until convergence.*

- *Example: Virginia Lead Example: We will look at fitting generalized linear models in R and SAS. The data were collected in 2000 as part of a Virginia program to monitor elevated lead levels in the blood of children under the age 6. The data were collected at the county level and are thus lattice (regional) data. Variables include*

  1. *$Z(\mathbf{s}_i)$: the number of children found to have elevated lead levels in county $i$.*
  2. *$n(\mathbf{s}_i)$: the number of children tested in county $i$.*
  3. *$X_1(\mathbf{s}_i)$: the median housing value (units of \$100,000).*
  4. *$X_2(\mathbf{s}_i)$: the number of children under 17 years of age living in poverty per 100,000 such children at risk.*

  *The proportion of children in each county with elevated lead levels is estimated to be $p(\mathbf{s}_i) = Z(\mathbf{s}_i)/n(\mathbf{s}_i)$.*

  *We can start modeling such data as either binomial counts or Poisson counts.*

  *Binomial regression: We assume*

  $$Z(\mathbf{s}_i) \sim Bin(n(\mathbf{s}_i), \mu(\mathbf{s}_i))$$

  *and model the binomial counts using a logit link*

  $$g(\mu(\mathbf{s}_i)) = \log\left\{\frac{\mu(\mathbf{s}_i)}{1-\mu(\mathbf{s}_i)}\right\} = \beta_0 + \beta_1 X_1(\mathbf{s}_i) + \beta_2 X_2(\mathbf{s}_i)$$

  *and variance function*

  $$v(\mu(\mathbf{s}_i)) = n(\mathbf{s}_i)\mu(\mathbf{s}_i)(1-\mu(\mathbf{s}_i)).$$

  *A dispersion parameter $\sigma^2$ can be incorporated as a multiplicative factor in the variance function, i.e.*

  $$v(\mu(\mathbf{s}_i)) = \sigma^2 n(\mathbf{s}_i)\mu(\mathbf{s}_i)(1-\mu(\mathbf{s}_i)).$$

  *If $\sigma^2 \neq 1$ then we would technically no longer have binomial counts. Note that $E[p(\mathbf{s}_i)] = \mu(\mathbf{s}_i)$.*

  *First we fit a standard logistic regression model. We are assuming that the dispersion parameter is equal to 1.*

```
> valead1.bin<-glm(cbind(elevated72,tested72-elevated72) ~ mval1000+pov1000,
+ data=va2000,family=binomial)
> summary(valead1.bin)
```

```
Coefficients:
```

| | Estimate | Std. Error | z value | Pr(>\|z\|) |
|---|---|---|---|---|
| (Intercept) | −2.18499 | 0.06342 | −34.454 | $< 2e-16$ |
| mval1000 | −0.65524 | 0.08303 | −7.891 | $2.99e-15$ |
| pov1000 | 0.65664 | 0.55889 | 1.175 | 0.24 |

```
(Dispersion parameter for binomial family taken to be 1)

Null deviance:  966.63 on 132 degrees of freedom
Residual deviance:  884.45 on 130 degrees of freedom
AIC: 1252.8
```

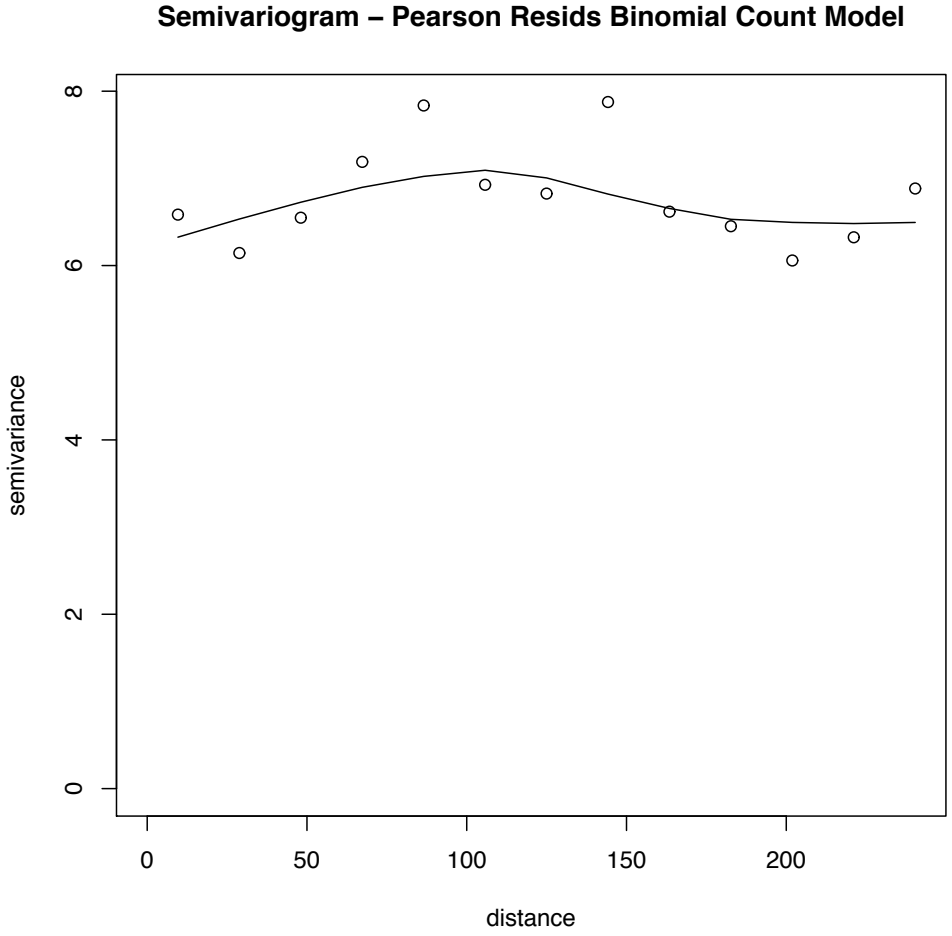
*If we carried out the typical goodness of fit test we would have assumed that* $G^2 = 884.45$ *is a realization from a chi-squared distribution with* 130 *degrees of freedom which would yield a p-value of* 0. *Small p-values can arise for 1 of 3 reasons;*

1. *The model is incorrect in that it is missing important covariates.*
2. *The binomial is not an adequate model for the distribution of responses.*
3. *There are outlying observations.*

*Overdispersion may result from either of (2) or (3) (or both).*

*However, the chi-squared approximation may not be very good. Even though we are working with grouped data (*133 *counties) a number of the counties have small sample sizes, i.e. the number of children tested is small. For example,* 15 *of them have* 10 *or fewer children tested for elevated lead levels and* 8 *have* 5 *or fewer.*

*I checked for spatial correlation. Based on this plot there does not appear to be all that much, but we will come back to this later. Spatial correlation is one cause of overdispersion and the cure for overdispersion in this case would be to model the spatial correlation.*



*We might also expect there to be overdispersion due to heterogeneity among and within the counties. Suppose that the risk of elevated lead levels is due to the 2 predictors we have in the model plus another predictor we have left out of the model. Then the responses would be binomially distributed at each fixed combination of the 3 potential predictors. We are focusing only on the 2 predictors in the model above which implies that at each fixed combination of those 2 we actually have a mixture of binomial populations. That is, binary responses in a county may have different probabilities. If one believes that this is a possible cause for overdispersion then the best cure is to find the missing covariate(s).*

*One can also fit generalized linear models that allow for overdispersion such as a negative binomial model for "Poisson" counts or a beta binomial model for "binomial" counts. Under an assumption of a multiplicative scale parameter the easiest way is to refit the model using quasi-likelihood to estimate all parameters. As described above the quasi-likelihood estimates of the regression coefficients will be identical to the maximum likelihood estimates, but we also get an estimate of the overdispersion parameter and adjusted standard errors.*

```
> valead2.bin=glm(cbind(elevated72, tested72-elevated72)~mval1000+pov1000,
+ data=va2000,family=quasibinomial)
```

```
Coefficients:
```

| | Estimate | Std. Error | t value | Pr(> \|t\|) |
|---|---|---|---|---|
| (Intercept) | −2.1850 | 0.1712 | −12.764 | $< 2e-16$ |
| mval1000 | −0.6552 | 0.2241 | −2.924 | 0.00408 |
| pov1000 | 0.6566 | 1.5086 | 0.435 | 0.66409 |

```
(Dispersion parameter for quasibinomial family taken to be 7.285833)

Null deviance:  966.63 on 132 degrees of freedom
Residual deviance:  884.45 on 130 degrees of freedom
AIC: NA
```

*Note that the estimated regression coeficients are identical to those in the first model but the standard errors are quite a bit higher. How are the standard errors adjusted? The standard error for* `mval1000` *is* 0.08303 *in the first model. The corresponding standard error in the second model is* $0.2241 = 0.08303\sqrt{7.286}$. *R does not return a standard error for the overdispersion parameter so we do not know if it is "statistically significantly" different from* 1, *but generally that is not of much concern anyway. One can also get the quasi-likelihood results from the original model;*

```
> sum(residuals(valead1.bin,type="pearson")^2)/130
[1] 7.285833
```

*and then make the standard error adjustments*

```
> sqrt(7.28533*diag(vcov(valead1.bin)))
(Intercept)    mval1000     pov1000
  0.1711737   0.2241199   1.5085297
```

*without ever fitting the* `glm` *quasibinomial model.*

*SAS code and fitting results are shown below. PROC GLIMMIX was used to fit the models. This is a procedure for fitting generalized linear mixed models and we are not fitting mixed models yet but unlike the mixed model functions in R, GLIMMIX can be used to fit fixed effects models (in the case of a fixed effects logistic model with no overdispersion PROC GLIMMIX defaults to PROC GENMOD) or provide a fit equivalent to the* `quasibinomial` *fit we got in R above. We can also trick GLIMMIX into providing a standard error for the overdispersion parameter if we want to see one.*

```
title1 'Logistic regression model'
proc glimmix data=SGdata.VA2000;
   logtested72 = log(tested72);
   model elevated72/Tested72 = mval1000 pov1000 / solution;
   run;

title1 'Logistic regression - overdispersion;
proc glimmix data=SGdata.VA2000;
   logtested72 = log(tested72);
   model elevated72/Tested72 = mval1000 pov1000 / solution;
   random _residual_;
   run;

title1 'Logistic regression model - overdispersion with SE'
proc glimmix data=SGdata.VA2000;
   logtested72 = log(tested72);
   model elevated72/Tested72 = mval1000 pov1000 / solution;
   random _residual_ / type=sp(exp)(easting northing);
   /* Hold range at 0, implying no correlations */
   parms (1) (0) / hold=2;
   run;
```

*Note the use of the* `random` *statement in the second and third models. We will have more to say about that later but for now we can say that the overdispersion parameter has been estimated by specifying a pure nugget effect model for one variance component in a mixed model. The output is shown below.*

*For the model without overdispersion we have*

Parameter Estimates

| Effect | Estimate | Standard Error | DF | t Value | Pr > \|t\| |
|---|---|---|---|---|---|
| Intercept | -2.1850 | 0.06342 | 130 | -34.45 | <.0001 |
| mval1000 | -0.6552 | 0.08303 | 130 | -7.89 | <.0001 |
| pov1000 | 0.6566 | 0.5589 | 130 | 1.17 | 0.2422 |

*which matches exactly the output from R. Fit statistics are also provided*

Fit Statistics

| | |
|---|---|
| -2 Log Likelihood | 1246.84 |
| AIC (smaller is better) | 1252.84 |
| AICC (smaller is better) | 1253.02 |
| BIC (smaller is better) | 1261.51 |
| CAIC (smaller is better) | 1264.51 |
| HQIC (smaller is better) | 1256.36 |
| Pearson Chi-Square | 947.16 |
| Pearson Chi-Square / DF | 7.29 |

*Note that SAS actually does give us an estimate of the overdispersion parameter but we would have to do the standard error modifications ourselves and, like R, we do not get a standard error.*

*For the model with overdispersion we have*

Fit Statistics

| | |
|---|---|
| -2 Log Likelihood | 1246.84 |
| AIC (smaller is better) | 1252.84 |
| AICC (smaller is better) | 1253.02 |
| BIC (smaller is better) | 1261.51 |
| CAIC (smaller is better) | 1264.51 |
| HQIC (smaller is better) | 1256.36 |
| Pearson Chi-Square | 947.16 |
| Pearson Chi-Square / DF | 7.29 |

Parameter Estimates

| Effect | Estimate | Standard Error | DF | t Value | Pr > \|t\| |
|---|---|---|---|---|---|
| Intercept | -2.1850 | 0.1712 | 130 | -12.76 | <.0001 |
| mval1000 | -0.6552 | 0.2241 | 130 | -2.92 | 0.0041 |
| pov1000 | 0.6566 | 1.5086 | 130 | 0.44 | 0.6641 |
| Residual | 7.2858 | . | . | . | . |

*and for the model with overdispersion and its standard error we have*

Covariance Parameter Estimates

| Cov Parm | Estimate | Standard Error |
|---|---|---|
| Variance | 7.2858 | 0.9037 |
| SP(EXP) | 0 | . |

Solutions for Fixed Effects

| Effect | Estimate | Standard Error | DF | t Value | Pr > \|t\| |
|---|---|---|---|---|---|
| Intercept | -2.1850 | 0.1712 | 130 | -12.76 | <.0001 |
| mval1000 | -0.6552 | 0.2241 | 130 | -2.92 | 0.0041 |
| pov1000 | 0.6566 | 1.5086 | 130 | 0.44 | 0.6641 |

*The overdispersion parameter is estimated to be* 7.2858 *as in the R output but now we also get a standard error from which we can confidently conclude that it is greater than* 1.

*It is important to understand that this method is appropriate only when the set of covariates is suitable. It also assumes that* $X^2/\sigma^2$ *is approximately chi-squared with* $n - p$ *degrees of freedom which requires "sufficiently" large sample sizes. Actually the models we are looking at here do not fit very well (there are some notable outliers) so the above adjustments are of questionable value. However, these are the only covariates we have.*

*One further note: when we estimated the overdispersion parameter and adjusted the standard errors using the* `quasibinomial` *family argument in R we did not get an AIC value. The reason is that we are no longer working with a valid likelihood but with a quasilikelihood. When we specified*

```
random _residual_
```

*second model the results matched exactly the results from the corresponding R model. But SAS did give us AIC values (see the* `Fit Statistics` *from that output). However, those stastistics were identical to the* `Fit Statistics` *from the first SAS model in which we ignored overdispersion. This is of questionable validity. At the very least we have added another parameter and ought to see the effect of that (but we do not). There are quasi-like AIC adjustments that can be made. I will give you a handout in class that discusses one of those. The R and SAS user who want to use this approach will need to make those modifications on their own. SAS will compute something called a quasi-AIC in GLIMMIX but it is not the same as the QAIC describe in the handout and it is not described or documented in the GLIMMIX documentation. Getting it requires different commands, also. We will not discuss that further here but see me if you want references to some of the literature.*

*As indicated above spatial autocorrelation is one cause of overdispersion. The semivariogram shown above does not suggest strong correlation but we will take a look at how to incorporate a spatial correlation model into a generalized linear model analysis. The analysis we are about to look at is the direct analog to the generalized least squares models we fit earlier to the leukemia data set. I have not been able to find R functions that will do this in a generalized linear models context but there is a function in the* `nlme` *package that will do the nonlinear version of generalized least squares. One can do this in SAS using PROC GLIMMIX. We will look at examples next.*

*For the R application the model is*

$$p(\mathbf{s}_i) = \frac{\exp(\beta_0 + \beta_1 X_1(\mathbf{s}_i) + \beta_2 X_2(\mathbf{s}_i))}{1 + \exp(\beta_0 + \beta_1 X_1(\mathbf{s}_i) + \beta_2 X_2(\mathbf{s}_i))} + \epsilon_i(\mathbf{s}_i)$$

*where* $p(\mathbf{s}_i)$ *is the sample proportion with elevated lead levels in the ith county. The R code to fit this model using the* `gnls` *function.*

```
require(nlme)
resp.i<-lead.dat$elevated72/lead.dat$tested.72 # sample proportions
logit.model<-function(x1,x2,b0,b1,b2){
  a<-exp(b0+b1*x1+b2*x2)
  a/(1+a)}
lead.binom<-gnls(resp.i~logit.model(mval1000,pov1000,b0,b1,b2),data=lead.dat,
   start=c(b0=-2,b1=-.66,b2=.66), corr=corSpher(c(60,.5),
   form=~easting+northing,nugget=T),
   weights=varPower(form=~fitted(.)*(1-fitted(.))/tested72,fixed=.5))
# I used estimates for correlation starting values given in Schabenberger and Gotway
# which come from the semivariogram plot above.
# I used estimates for coefficient starting values from previous fits.
# b0 is the intercept, b1 is for mval1000, and b2 is for pov1000
summary(lead.binom)
Generalized nonlinear least squares fit
  Model: resp.i ~ logit.model(mval1000, pov1000, b0, b1, b2)
  Data: lead.dat
        AIC       BIC   logLik
  -256.4891 -239.1470 134.2446

Correlation Structure: Spherical spatial correlation
 Formula: ~easting + northing
 Parameter estimate(s):
     range     nugget
77.9170365  0.8053385
Variance function:
 Structure: Power of variance covariate
 Formula: ~fitted(.) * (1 - fitted(.))/tested72
```

```
 Parameter estimates:
power
  0.5



Coefficients:
        Value Std.Error    t-value p-value
b0 -2.1044912 0.1957247 -10.752301   0.000
b1 -0.8085939 0.2677174  -3.020326   0.003
b2  1.3500449 1.5718042   0.858914   0.392

Residual standard error: 2.691339
Degrees of freedom: 133 total; 130 residual
```

*The* `varPower` *function assumes a weight function that is equal to*

$$\sigma^2 \left[ \frac{\mu(1-\mu)}{n} \right].$$

*The SAS code is shown below. This code uses PROC GLIMMIX but we are not fitting a mixed model. We are using GLIMMIX to fit a binomial count logistic regression model with fixed effects.*

*We assume* $E\left[p\left(\mathbf{s}_i\right)\right] = \mu\left(\mathbf{s}_i\right)$ *and that*

$$g\left(\mu\left(\mathbf{s}_i\right)\right) = \log\left\{\frac{\mu\left(\mathbf{s}_i\right)}{1-\mu\left(\mathbf{s}_i\right)}\right\} = \beta_0 + \beta_1 X_1\left(\mathbf{s}_i\right) + \beta_2 X_2\left(\mathbf{s}_i\right)$$

*with*

$$Var\left[Z\left(\mathbf{s}_i\right)\right] = \sigma_0^2 \mathbf{V}_{\boldsymbol{\mu}} + \sigma_1^2 \mathbf{V}^{1/2} \mathbf{R} \mathbf{V}^{1/2}.$$

*The spatial correlation structure is incorporated through* $\mathbf{R}$*. Parameters were estimated using pseudoliklihood.*

```
proc glimmix data=VA;
   model elevated72/tested72 = mval1000 pov1000 /
                        ddfm   = residual
                        dist   = binomial
                        solution;
   random _residual_ / subject = intercept
                       type    = sp(sph)(easting northing);
   random _residual_;
   parms (180) (7) (6);
   nloptions tech=nrridg;
run;
```

*We have specified starting values of* 180 *for the range,* 7 *for the sill, and* 6 *for the nugget effect. The* `nloptions` *command tells SAS to use the Newton-Raphson method with ridging (an alternative to garden variety Newton-Raphson that is more stable). The output is shown below.*

Covariance Parameter Estimates

| Cov Parm | Subject | Estimate | Standard Error |
|---|---|---|---|
| SP(SPH) | Intercept | 186.69 | 69.8934 |
| Residual (VC) | | 6.5851 | 0.9469 |
| Residual | | 1.0227 | 0.8304 |

Solutions for Fixed Effects

| Effect | Estimate | Standard Error | DF | t Value | Pr > \|t\| |
|---|---|---|---|---|---|
| Intercept | -2.1194 | 0.2139 | 130 | -9.91 | <.0001 |
| mval1000 | -0.8489 | 0.2858 | 130 | -2.97 | 0.0036 |
| pov1000 | 1.3868 | 1.6844 | 130 | 0.82 | 0.4118 |

*The results from R and SAS are quite similar and both are similar to the earlier results ignoring spatial correlation. The range parameter estimate is different in the SAS and R output but the sill is similar (*6.59 *in SAS and* $2.69^2 = 7.24$ *in R). The biggest change is in the estimated coefficient for poverty but the standard error associated with that estimate is large in all the models. Note also that the standard errors for the estimated regression coefficients are higher in the marginal model with a spatial autocorrelation structure. Which do we use? I do not trust the quasilikelihood adjusted model for reasons specified above. But frankly none of these models fit all that well.*

*Poisson regression: The assumption is made that* $Z(\mathbf{s}_i) \sim Poi(\mu(\mathbf{s}_i))$ *where* $\mu(\mathbf{s}_i) = n(\mathbf{s}_i)\lambda(\mathbf{s}_i)$ *and*

$$\log\{\lambda(\mathbf{s}_i)\} = \beta_0 + \beta_1 X_1(\mathbf{s}_i) + \beta_2 X_2(\mathbf{s}_i)$$

$$g(\mu(\mathbf{s}_i)) = \log\{\mu(\mathbf{s}_i)\} = \log\{n(\mathbf{s}_i)\} + \beta_0 + \beta_1 X_1(\mathbf{s}_i) + \beta_2 X_2(\mathbf{s}_i)$$

$$v(\mu(\mathbf{s}_i)) = \mu(\mathbf{s}_i).$$

*A multiplicative dispersion parameter* $\sigma^2$ *would be incorporated as in the binomial case above. The* $\log\{n(\mathbf{s}_i)\}$ *is called an offset term and is treated as a predictor with known coefficient equal to 1.*

*I fit models in R. I omit the model fit assuming the dispersion parameter is equal to 1. The only difference is in the standard errors which were* 0.06167 *for* $\widehat{\beta}_0$, 0.08140 *for* $\widehat{\beta}_1$, *and* 0.5402 *for* $\widehat{\beta}_2$. *Of course the test results were also different.*

```
> valead.poi<-glm(elevated72 offset(log(tested72))+mval1000+pov1000,
+ data=va2000,family=quasipoisson)
> summary(valead.poi)


Coefficients:
```

| | Estimate | Std. Error | t value | Pr(> \|t\|) |
|---|---|---|---|---|
| (Intercept) | −2.2781 | 0.1606 | −14.185 | $< 2e-16$ |
| mval1000 | −0.6228 | 0.2120 | −2.938 | 0.00391 |
| pov1000 | 0.6293 | 1.4068 | 0.447 | 0.65540 |

```
(Dispersion parameter for quasipoisson family taken to be 6.782545)


Null deviance:  893.54 on 132 degrees of freedom
Residual deviance:  815.87 on 130 degrees of freedom
AIC: NA
```

*The corresponding SAS code and output are shown below.*

```
proc glimmix data=SGdata.VA2000;
   logtested72 = log(tested72);
   model elevated72 = mval1000 pov1000 / solution offset=logtested72 dist=poisson;
   random _residual_;
   run;
```

*This code fits the "Poisson" regression model and estimates the overdispersion parameter. As in the binomial results above the* `Fit Statistics` *are suspect because technically we are looking at a quasilikelihood estimation procedure. The output below is*

Fit Statistics

| | |
|---|---|
| -2 Log Likelihood | 1190.14 |
| AIC (smaller is better) | 1196.14 |
| AICC (smaller is better) | 1196.33 |
| BIC (smaller is better) | 1204.82 |
| CAIC (smaller is better) | 1207.82 |
| HQIC (smaller is better) | 1199.67 |
| Pearson Chi-Square | 881.73 |
| Pearson Chi-Square / DF | 6.78 |

Parameter Estimates

| Effect | Estimate | Standard Error | DF | t Value | Pr > \|t\| |
|---|---|---|---|---|---|
| Intercept | -2.2781 | 0.1606 | 130 | -14.18 | <.0001 |
| mval1000 | -0.6228 | 0.2120 | 130 | -2.94 | 0.0039 |
| pov1000 | 0.6293 | 1.4068 | 130 | 0.45 | 0.6554 |
| Residual | 6.7825 | . | . | . | . |

*I don't show the spatial Poisson SAS code here but for completeness the R nonlinear model results are shown below along with results from both the SAS and R output.*

```
# Spatial Marginal Poisson model
log.model<-function(x1,x2,n.i,b0,b1,b2){
  n.i*exp(b0 + b1*x1 + b2*x2)}
lead.poi<-gnls(elevated72~log.model(mval1000,pov1000,tested72,b0,b1,b2),
data=lead.dat,start=c(b0=-2,b1=-.66,b2=.66),corr=corSpher(c(60,.5),form=~easting+northing,nugget=T),
weights=varPower(form=~fitted(.),fixed=0.5))
# Note that tested72 is in the argument list of log.model. This is an offset term.
# I do not specify it as a parameter and so it is treated as fixed.
# b0 is the intercept, b1 is for mval1000, and b2 is for pov1000
summary(lead.poi)
Generalized nonlinear least squares fit
  Model: elevated72 ~ log.model(mval1000, pov1000, tested72, b0, b1, b2)
  Data: lead.dat
       AIC      BIC    logLik
  798.5805 815.9226 -393.2903
Correlation Structure: Spherical spatial correlation
 Formula: ~easting + northing
 Parameter estimate(s):
     range     nugget
77.7130301  0.8009977
Variance function:
 Structure: Power of variance covariate
 Formula: ~fitted(.)
 Parameter estimates:
power
```

```
  0.5
Coefficients:
        Value Std.Error    t-value p-value
b0 -2.1993349 0.1837770 -11.967414  0.0000
b1 -0.7715992 0.2536714  -3.041727  0.0028
b2  1.2829099 1.4642498   0.876155  0.3826

Residual standard error: 2.595927
Degrees of freedom: 133 total; 130 residual
```

*The SAS results from Schabenberger and Gotway are*

```
Coefficients:
     Value   Std.Error    t-value   p-value
b0 -2.22135  0.2010       -11.01     0.0000
b1 -0.8096   0.2713        -2.98     0.0034
b2  1.3172   1.5736         0.84     0.4041
# The estimated sill is 6.137 and the range estimate is 186.7.
```

*We could fit the Poisson version of the third SAS binomial model with the pure nugget effect spatial correlation model and get an estimate of the standard error for the overdispersion parameter of* 0.8413*. The caveats about the validity of the overdispersion parameter we discussed above are also relevant here.*

*In general R and SAS tell us the same thing. There are differences but given the different fitting methods this is not that surprising. The estimated range in the spatial models is quite different and the estimate of the poverty effect is different (but again the standard errors are large enough there that variability of the magnitude we see here is not surprising).*

*Schabenberger and Gotway discuss this example at some length. They raise the question of whether it is better to model the data as binomial counts or Poisson counts. Overdispersion is to be expected but neither model fits all that well so the quasibinomial adjustments based on a multiplicative dispersion effect are of questionable validity. AIC and QAIC comparisons would not help because the likelihoods are different in the likelihood based models. Schabenberger and Gotway like the Poisson in part because the mean of a binomial is always greater than the variance whereas the mean of a Poisson is equal to the variance so, in general and relatively speaking, overdispersion should not be as great a problem with Poisson data as with binomial data. The data are clearly more appropriately thought of as (quasi)binomial counts but Schabenberger and Gotway justify the Poisson model by appealing to the well known result that binomial probabilities can be approximated by Poisson probabilities if* $n$ *is large and* $p$ *is small. However, county wide* $n$*'s are not all that big and the* $p$*'s are not all that small, at least for all counties. This is one of those times when the approach to take, binomial versus Poisson, is mainly subjective. If something is going on then either one should show it. The take home message at this point is that it appears that lead levels are negatively related to median housing prices implying that as house prices increase exposure to lead decreases (not an earth shattering result). We do have some evidence of spatial correlation and good reason to suspect overdispersion with a not very good way to tackle the latter problem, although it could be argued that attempting to account for it through spatial correlation is better than the quasi-likelihood method. We did fit a spatial binomial model above but there is nothing compelling in the results to suggest one model is better than the other. We will consider another approach based on mixed models but we need to talk about mixed models first.*

*Linear Mixed Models*

- We will first look at mixed models in the standard garden variety linear regression setting and then look at generalized linear mixed models.
- A linear mixed model is expressed as

$$\mathbf{Z}(\mathbf{s}) = \mathbf{X}(\mathbf{s})\boldsymbol{\beta} + \mathbf{U}(\mathbf{s})\boldsymbol{\nu} + \boldsymbol{\epsilon}(\mathbf{s})$$

where $\boldsymbol{\nu}$ is a vector of random effects with $E[\boldsymbol{\nu}] = \mathbf{0}$ and $\text{Var}[\boldsymbol{\nu}] = \mathbf{G}$, and $\boldsymbol{\epsilon}$ has mean $\mathbf{0}$ and $\text{Var}[\boldsymbol{\epsilon}] = \mathbf{R}$. We also assume that $\boldsymbol{\nu}$ and $\boldsymbol{\epsilon}$ are independent of one another. It is not hard to see that

$$\text{Var}[\mathbf{Z}(\mathbf{s})] = \mathbf{V} = \mathbf{U}(\mathbf{s})\mathbf{G}\mathbf{U}(\mathbf{s})^{'} + \mathbf{R}$$

and it can be shown that

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

and

$$\widehat{\boldsymbol{\nu}} = \mathbf{G}\mathbf{U}(\mathbf{s})^{'}\mathbf{V}^{-1}\left(\mathbf{Z}(\mathbf{s}) - \mathbf{X}(\mathbf{s})\widehat{\boldsymbol{\beta}}\right).$$

- *Simple Example: This is not a spatial example but it will help to see examples of the various matrices. You should all have seen randomized complete block designs and discussed analysis of data generated by such designs in the past. These are actually quintessential examples of a mixed model. The data here are from an ergometric study of the effort expended by* 9 *subjects to arise from stools of* 4 *different designs. The treatment is stool type and the blocks are the subjects. There are a total of* 36 *observations. We will assume that stool* `T1` *is a standard type and the other* 3 *types will be compared to it.*

  *Part of the data is shown below.*

```
> ergoStool
Grouped Data: effort ~ Type | Subject
   effort Type Subject
1      12   T1       1
2      15   T2       1
3      12   T3       1
4      10   T4       1
5      10   T1       2
6      14   T2       2
7      13   T3       2
8      12   T4       2
.       .   .        .
.       .   .        .
.       .   .        .
34     13   T2       9
35     10   T3       9
36      8   T4       9
```

  *The model is*

$$\mathbf{Z} = \mathbf{X}\boldsymbol{\beta} + \mathbf{U}\boldsymbol{\nu} + \boldsymbol{\epsilon}.$$

  *We assume that* $\boldsymbol{\nu} \sim N\left(\mathbf{0}, \sigma_{\nu}^2\mathbf{I}\right)$ *and* $\boldsymbol{\epsilon} \sim N\left(\mathbf{0}, \sigma_{\epsilon}^2\mathbf{I}\right)$. *We assume that* `T1` *is the reference type. The resulting* $X$ *matrix is shown on the next page. This matrix was produced by the command*

```
> model.matrix(effort~Type,ergoStool)
```

  *We have a leading column of* 1*'s for the intercept and 3 other columns containing dummy variable coding for types* `T2`, `T3`, `T4`. *The coefficient vector is* $\boldsymbol{\beta}^{'} = [\beta_0 \quad \beta_1 \quad \beta_2 \quad \beta_3]$. *The intercept* $\beta_0$ *will be the mean of the Type 1 stools over the* 9 *subjects and the other coefficients will be the deviation of their means from the Type 1 mean. These are the fixed effects in the model.*

```
   (Intercept) TypeT2 TypeT3 TypeT4
1            1      0      0      0
2            1      1      0      0
```

```
3             1       0       1       0
4             1       0       0       1
5             1       0       0       0
6             1       1       0       0
7             1       0       1       0
8             1       0       0       1
9             1       0       0       0
10            1       1       0       0
11            1       0       1       0
12            1       0       0       1
13            1       0       0       0
14            1       1       0       0
15            1       0       1       0
16            1       0       0       1
17            1       0       0       0
18            1       1       0       0
19            1       0       1       0
20            1       0       0       1
21            1       0       0       0
22            1       1       0       0
23            1       0       1       0
24            1       0       0       1
25            1       0       0       0
26            1       1       0       0
27            1       0       1       0
28            1       0       0       1
29            1       0       0       0
30            1       1       0       0
31            1       0       1       0
32            1       0       0       1
33            1       0       0       0
34            1       1       0       0
35            1       0       1       0
36            1       0       0       1
```

*We have* 9 *different subjects which we can consider to be a random sample from some population of subjects. Thus* `Subject` *is a random effect and the random effect "design" matrix* $\mathbf{U}$ *is shown below.*

```
> U
      [,1] [,2] [,3] [,4] [,5] [,6] [,7] [,8] [,9]
 [1,]    1    0    0    0    0    0    0    0    0
 [2,]    1    0    0    0    0    0    0    0    0
 [3,]    1    0    0    0    0    0    0    0    0
 [4,]    1    0    0    0    0    0    0    0    0
 [5,]    0    1    0    0    0    0    0    0    0
 [6,]    0    1    0    0    0    0    0    0    0
 [7,]    0    1    0    0    0    0    0    0    0
 [8,]    0    1    0    0    0    0    0    0    0
 [9,]    0    0    1    0    0    0    0    0    0
[10,]    0    0    1    0    0    0    0    0    0
[11,]    0    0    1    0    0    0    0    0    0
[12,]    0    0    1    0    0    0    0    0    0
[13,]    0    0    0    1    0    0    0    0    0
[14,]    0    0    0    1    0    0    0    0    0
```

```
[15,]    0    0    0    1    0    0    0    0    0
[16,]    0    0    0    1    0    0    0    0    0
[17,]    0    0    0    0    1    0    0    0    0
[18,]    0    0    0    0    1    0    0    0    0
[19,]    0    0    0    0    1    0    0    0    0
[20,]    0    0    0    0    1    0    0    0    0
[21,]    0    0    0    0    0    1    0    0    0
[22,]    0    0    0    0    0    1    0    0    0
[23,]    0    0    0    0    0    1    0    0    0
[24,]    0    0    0    0    0    1    0    0    0
[25,]    0    0    0    0    0    0    1    0    0
[26,]    0    0    0    0    0    0    1    0    0
[27,]    0    0    0    0    0    0    1    0    0
[28,]    0    0    0    0    0    0    1    0    0
[29,]    0    0    0    0    0    0    0    1    0
[30,]    0    0    0    0    0    0    0    1    0
[31,]    0    0    0    0    0    0    0    1    0
[32,]    0    0    0    0    0    0    0    1    0
[33,]    0    0    0    0    0    0    0    0    1
[34,]    0    0    0    0    0    0    0    0    1
[35,]    0    0    0    0    0    0    0    0    1
[36,]    0    0    0    0    0    0    0    0    1
```

*The random effects vector is* $\boldsymbol{\nu}^{'} = [\nu_1 \;\; \nu_2 \;\; \nu_3 \;\; \cdots \;\; \nu_9]$.

*The covariance matrices associated with the random effects and the error process are* $\sigma^2_\nu \mathbf{I}$ *and* $\sigma^2_\epsilon \mathbf{I}$ *where* $\mathbf{I}$ *is a* $36 \times 36$ *identity matrix. We have*

$$Var[\mathbf{Z}] = \sigma^2_\nu \mathbf{U}\mathbf{U}^{'} + \sigma^2_\epsilon \mathbf{I}.$$

*This has a block diagonal structure. The piece associated with subjects 1 and 2 are shown below.*

$$\begin{bmatrix}
\sigma^2_\nu + \sigma^2_\epsilon & \sigma^2_\nu & \sigma^2_\nu & \sigma^2_\nu & 0 & 0 & 0 & 0 \\
\sigma^2_\nu & \sigma^2_\nu + \sigma^2_\epsilon & \sigma^2_\nu & \sigma^2_\nu & 0 & 0 & 0 & 0 \\
\sigma^2_\nu & \sigma^2_\nu & \sigma^2_\nu + \sigma^2_\epsilon & \sigma^2_\nu & 0 & 0 & 0 & 0 \\
\sigma^2_\nu & \sigma^2_\nu & \sigma^2_\nu & \sigma^2_\nu + \sigma^2_\epsilon & 0 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 & \sigma^2_\nu + \sigma^2_\epsilon & \sigma^2_\nu & \sigma^2_\nu & \sigma^2_\nu \\
0 & 0 & 0 & 0 & \sigma^2_\nu & \sigma^2_\nu + \sigma\epsilon^2 & \sigma^2_\nu & \sigma^2_\nu \\
0 & 0 & 0 & 0 & \sigma^2_\nu & \sigma^2_\nu & \sigma^2_\nu + \sigma^2_\epsilon & \sigma^2_\nu \\
0 & 0 & 0 & 0 & \sigma^2_\nu & \sigma^2_\nu & \sigma^2_\nu & \sigma^2_\nu + \sigma^2_\epsilon
\end{bmatrix}$$

*We see that subjects* 1 *and* 2 *are independent of one another but that observations within each subject are actually correlated with one another. We can fit this model using the* `lme` *function in R.*

```
> ergo.fit<-lme(effort~Type,data=ergoStool,random=~1|Subject)
> summary(ergo.fit)
Linear mixed-effects model fit by REML
 Data: ergoStool
       AIC      BIC   logLik
  133.1308 141.9252 -60.5654

Random effects:
 Formula: ~1 | Subject
        (Intercept) Residual
StdDev:    1.332465 1.100295
```

```
Fixed effects: effort ~ Type
               Value Std.Error DF   t-value p-value
(Intercept) 8.555556 0.5760123 24 14.853079  0.0000
TypeT2      3.888889 0.5186838 24  7.497610  0.0000
TypeT3      2.222222 0.5186838 24  4.284348  0.0003
TypeT4      0.666667 0.5186838 24  1.285304  0.2110

Standardized Within-Group Residuals:
        Min          Q1         Med          Q3         Max
-1.80200345 -0.64316591  0.05783115  0.70099706  1.63142054

Number of Observations: 36
Number of Groups: 9
```

*The variance components estimates are* $\widehat{\sigma}^2_\nu = (1.332)^2$ *and* $\widehat{\sigma}^2_\epsilon = (1.100)^2$. *The estimate of* $\beta_0$ *is* $\widehat{\beta}_0 = 8.55556$ *which is just the mean of the* 9 *Type* 1 *effort values.*

```
> mean(ergoStool$effort[ergoStool$Type=="T1"])
[1] 8.555556
```

*The estimate of* $\beta_1$ *is* $\widehat{\beta}_1 = 3.88889$ *which is the difference between the mean Type* 1 *values and the Type* 2 *values.*

```
> mean(ergoStool$effort[ergoStool$Type=="T2"])-
+ mean(ergoStool$effort[ergoStool$Type=="T1"])
[1] 3.888889
```

*We can easily look at confidence intervals for all the parameters.*

```
> intervals(ergo.fit)
Approximate 95% confidence intervals

 Fixed effects:
                lower      est.    upper
(Intercept)  7.3667247 8.5555556 9.744386
TypeT2       2.8183781 3.8888889 4.959400
TypeT3       1.1517114 2.2222222 3.292733
TypeT4      -0.4038442 0.6666667 1.737177
attr(,"label")
[1] "Fixed effects:"

 Random Effects:
  Level: Subject
                    lower     est.    upper
sd((Intercept)) 0.7495964 1.332465 2.368559

 Within-group standard error:
    lower      est.     upper
0.8292498 1.1002946 1.4599319
```

*The variance components are generally of most interest rather than the random effects themselves. After all, the random effects would change with new subjects but the variability in those subjects would not. We can look at the random effects if we want though using the* `random.effects` *function. These are not estimates but predictions and the function produces* **B***est* **L***inear* **U***nbiased* **P***redictions or BLUPs.*

```
> random.effects(ergo.fit)
    (Intercept)
8 -1.708716e+00
5 -1.495127e+00
4 -8.543581e-01
9 -2.135895e-01
6  1.008157e-16
3  4.271791e-01
7  4.271791e-01
1  1.708716e+00
2  1.708716e+00
```

*We can also get fitted values and residuals in the normal way. The fitted values are the sums of the relevant coefficient estimates plus the random effect of the associated subject. Thus, each observation of effort is composed of* 4 *pieces:*

1. *the mean effort of the* 9 *Type* 1 *stools plus*
2. *a deviation equal to the mean effort for the particular stool being measured minus the Type* 1 *stool mean effort*
3. *plus a subject random effect*
4. *plus an error term.*

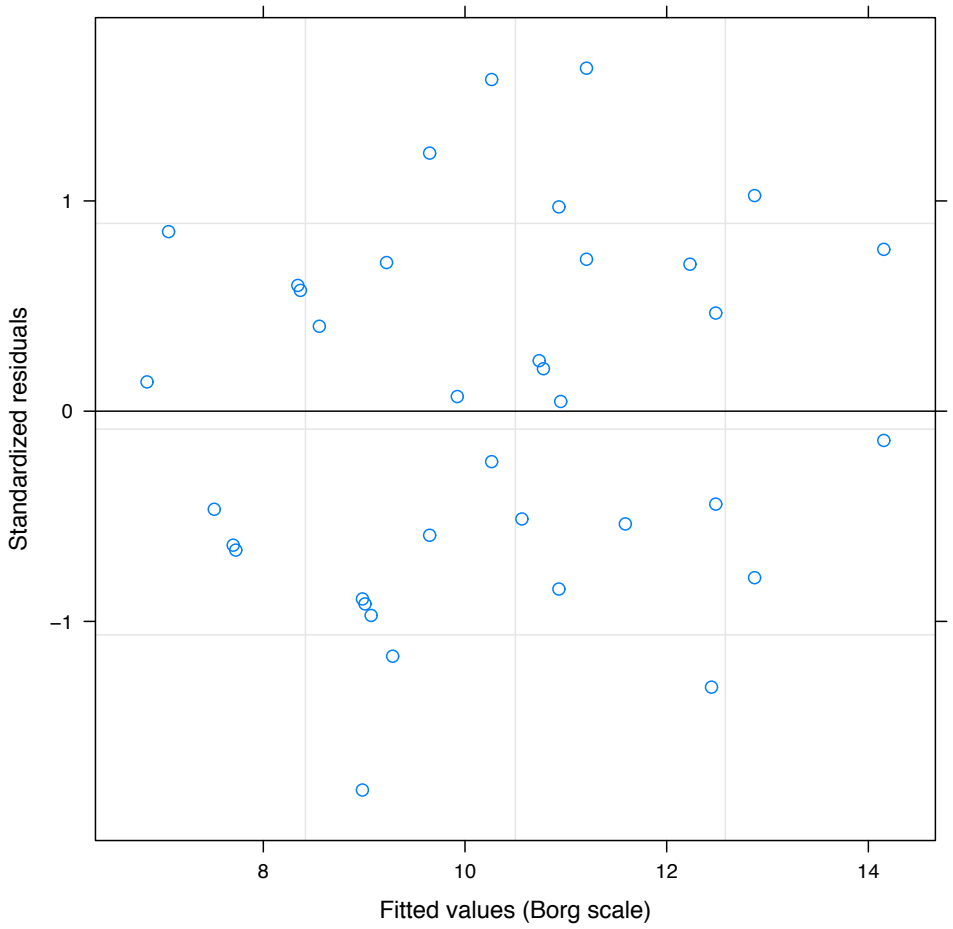
*For example the second observed effort value of* 15 *was made on stool Type* 2 *and Subject* 1. *We see that*

$$8.55556 + 3.88889 + 1.7087 + 0.8468 = 14.15316 + 0.84684 = 15$$

*From R we have*

```
> fitted(ergo.fit)[2]
       1
14.15316
> residuals(ergo.fit)[2]
        1
0.8468393
```

*Diagnostic procedures are similar to those you are already familiar with and are based primarily on residual analysis. For example a plot of the standardized residuals versus the fitted values is shown below and certainly does not indicate any problems.*




- The above example was obviously not a spatial example but I included it because it was a small data set and we could more easily match up some output from fitting a mixed model to the notation. The most basic spatial example would be a randomized block design in an agricultural field trial of some kind. Blocking was in fact envisioned as a rather crude way to control for spatial similarities in such studies. We could look at a simple randomized complete block design but we will look at another spatial example that is more interesting.

- *Example: There are* 16 *transects located in* 3 *soil types:* 8 *transects are in Type* 1, 4 *are in Type* 2, *and* 4 *are in Type* 3. *The area in which the transects are located is contaminated with a toxic compound. At* 11 *locations along each transect the concentration of the compound is measured. The concentration is known to trend from low values at one end of the transect to high values at the other end. The one dimensional coordinates are the same on each transect:* 1, 8, 15, 22, 29, 36, 43, 44, 50, 57, 64. *The model, expressed here in terms of the* $j$*th response on the* $i$*th transect, is*

$$Z_{ij} = \beta_0 + \nu_{0i} + \beta_1 T_{2i} + \beta_2 T_{3i} + (\beta_3 + \nu_{3i})\, s_{ij} + \epsilon_{ij}$$

*where* $T_{2i}$ *is a binary dummy variable taking the value of* 1 *if transect* $i$ *is in Type* 2 *and* 0 *otherwise,* $T_{3i}$ *is a binary dummy variable taking the value of* 1 *if transect* $i$ *is in Type* 3 *and* 0 *otherwise,* $s_{ij}$ *is the spatial location of the* $j$*th response on the* $i$*th transect,* $\nu_{0i}$ *is a random intercept term,* $\nu_{3i}$ *is a random slope term, and* $\epsilon_{ij}$ *is the within group error. We assume that* $\nu$ *and* $\epsilon$ *are normally distributed and independent of one another. The primary question of interest is whether the responses differ among the soil types.*

*I fit a model assuming independent within transect errors.*

```
Conc.lme1<-lme(conc~Type+Space,Conc,random=~Space|Transect)
Linear mixed-effects model fit by REML
  Data: Conc
  Log-restricted-likelihood: -579.6676
  Fixed: conc ~ Type + Space
(Intercept)       Type2       Type3        Space
246.4573397 214.5871794 258.9272322   0.5856833

Random effects:
 Formula: ~Space | Transect
 Structure: General positive-definite, Log-Cholesky parametrization
            StdDev     Corr
(Intercept) 37.3530736 (Intr)
Space        0.3463495 -0.209
Residual     4.4436053

Number of Observations: 176
Number of Groups: 16
```

*We see the estimates of the fixed effects and random effects. The estimate of* $\beta_0$ *is* $\widehat{\beta}_0 = 246.457$. *This is the estimated mean intercept of a population of transects and the standard deviation in the values of the intercepts in the population is estimated to be* 37.353. *Similarly we see that the slope coefficient for* `Space` *is* $\widehat{\beta}_3 = 0.586$. *Again this is an estimated mean slope and the standard deviation in the slopes is estimated to be* 0.346. *We can look at confidence intervals for all components.*

```
> intervals(Conc.lme1)
Approximate 95% confidence intervals

 Fixed effects:
                  lower        est.       upper
(Intercept) 220.6171248 246.4573397 272.2975546
Type2       166.2218633 214.5871794 262.9524955
Type3       210.5619161 258.9272322 307.2925483
Space         0.4113253   0.5856833   0.7600413
attr(,"label")
[1] "Fixed effects:"

 Random Effects:
  Level: Transect
                           lower       est.      upper
sd((Intercept))       24.9982477 37.3530736 55.8139964
sd(Space)               0.2387274  0.3463495  0.5024892
cor((Intercept),Space) -0.7495116 -0.2091962  0.4984027

 Within-group standard error:
   lower     est.    upper
3.958930 4.443605 4.987617
```

*There is considerable evidence for non-zero variance components.*

*The question arises as to whether or not the assumption of independent errors within transects is plausible. A plot of the semivariogram of the standardized residuals (not shown here) casts doubt on that assumption. A model incorporating a spatial error structure was fit.*

```
> Conc.lme2<-update(Conc.lme1,corr=corExp(form=~Space,nugget=T))
```

```
> Conc.lme2
Linear mixed-effects model fit by REML
  Data: Conc
  Log-restricted-likelihood: -568.0859
  Fixed: conc ~ Type + Space
(Intercept)       Type2       Type3       Space
245.0052289 217.8119916 261.3569753   0.5984438

Random effects:
 Formula: ~Space | Transect
 Structure: General positive-definite, Log-Cholesky parametrization
            StdDev     Corr
(Intercept) 37.6035642 (Intr)
Space        0.3462589 -0.242
Residual     5.4422329

Correlation Structure: Exponential spatial correlation
 Formula: ~Space | Transect
 Parameter estimate(s):
     range     nugget
13.7962694  0.1850731
Number of Observations: 176
Number of Groups: 16
> intervals(Conc.lme2)
Approximate 95% confidence intervals

 Fixed effects:
                  lower        est.       upper
(Intercept) 218.9932189 245.0052289 271.0172390
Type2       169.3870595 217.8119916 266.2369237
Type3       212.9320432 261.3569753 309.7819074
Space         0.4185985   0.5984438   0.7782892
attr(,"label")
[1] "Fixed effects:"

 Random Effects:
  Level: Transect
                           lower       est.      upper
sd((Intercept))       24.8917948 37.6035642 56.8069939
sd(Space)              0.2325370  0.3462589  0.5155964
cor((Intercept),Space)-0.7753818 -0.2418633  0.4931242

 Correlation structure:
            lower       est.       upper
range  1.13383576 13.7962694 167.8700345
nugget 0.06035852  0.1850731   0.4453446
attr(,"label")
[1] "Correlation structure:"

 Within-group standard error:
    lower      est.     upper
 2.898408  5.442233 10.218677
```

*Did we do better? We did according to AIC comparisons.*

```
> AIC(Conc.lme1)
[1] 1175.335
> AIC(conc.lme2)
[1] 1156.172
```

- Mixed effects models can be difficult to fit and the results can be very bad but not in any immediately obvious way. Here is an example that involves fitting a mixed model to the leukemia data set. This is a simple model that ignores spatial correlation entirely. I fit a random intercept model in R.

```
> leuk.lme<-lme(z~exp+age65+home,random=~1|key,data=leuk.dat)
> leuk.lme
Linear mixed-effects model fit by REML
  Data: leuk.dat
  Log-restricted-likelihood: -283.9471
  Fixed: z ~ exp + age65 + home
(Intercept)         exp       age65        home
-0.51727634  0.04883627  3.95088956 -0.56004134

Random effects:
 Formula: ~1 | key
        (Intercept)  Residual
StdDev:   0.6153029 0.2307386

Number of Observations: 281
Number of Groups: 281
```

There are no obvious problems apparent. We do note one strange thing. We haev 281 groups and the same number of observations, i.e. our group size is 1. That should be one tip-off that something is a bit off. We will look at the intervals.

```
> intervals(leuk.lme)
Approximate 95% confidence intervals

 Fixed effects:
                  lower        est.      upper
(Intercept) -0.82940652 -0.51727634 -0.2051462
exp         -0.02018850  0.04883627  0.1178610
age65        2.75892674  3.95088956  5.1428524
home        -0.89530376 -0.56004134 -0.2247789
attr(,"label")
[1] "Fixed effects:"

 Random Effects:
  Level: key
                     lower      est.    upper
sd((Intercept)) 0.02877582 0.6153029 13.15680

 Within-group standard error:
       lower         est.        upper
8.084425e-11 2.307386e-01 6.585538e+08
```

The interval for the within group error runs from 0 to about 658 million. Something is definitely not right. Below is the SAS code and output for the same model.

```
proc glimmix noprofile;
class indx;
model z=home age exp/solution dist=n ddfm=none;
random intercept/sub=indx;
nloptions tech=nrridg;
run;
```

Fit Statistics

| | |
|---|---|
| -2 Res Log Likelihood | 567.89 |
| AIC (smaller is better) | 571.89 |
| AICC (smaller is better) | 571.94 |
| BIC (smaller is better) | 579.17 |
| CAIC (smaller is better) | 581.17 |
| HQIC (smaller is better) | 574.81 |
| Generalized Chi-Square | 105.94 |
| Gener. Chi-Square / DF | 0.38 |

Covariance Parameter Estimates

| Cov Parm | Subject | Estimate | Standard Error |
|---|---|---|---|
| Intercept | indx | 0.04939 | 0.03669 |
| Residual | | 0.3825 | . |

Solutions for Fixed Effects

| Effect | Estimate | Standard Error | DF | t Value | Pr > \|t\| |
|---|---|---|---|---|---|
| Intercept | -0.5173 | 0.1586 | Infty | -3.26 | 0.0011 |
| home | -0.5600 | 0.1703 | Infty | -3.29 | 0.0010 |
| age | 3.9509 | 0.6055 | Infty | 6.53 | <.0001 |
| exp | 0.04884 | 0.03506 | Infty | 1.39 | 0.1637 |

The results pretty much match those we saw in R with one subtle exception. The output does not indicate any problems. We are told for example in the output that all convergence criteria were satisfied. The SAS log file also does not indicate any problem. We do see one red flag. There is no standard error given for the residual term in the covariance parameter summary information. The table below shows one other strange result when the SAS and R output random effects results are compared.

| Variance Component | SAS | R |
|---|---|---|
| Intercept | 0.0494 | 0.376 |
| Residual | 0.3825 | 0.0532 |

SAS and R have switched the within and between group error terms. The problem is that the variance

components are not *identifiable*. We are trying to do the impossible, fit 281 separate regressions, one for each census tract, when there is only one response in each tract. It is difficult to fit a regression line through a single point. It seems rather odd that the programs would even run. But, aside from some rather subtle red flags, it is not immediately apparent that there is a problem in either program.

Bivand, Pebesma, and Gomez-Rubio make this error in their text "Applied Spatial Data Analysis with R". They discuss mixed models and give an example with the leukemia data set. They fit a random intercept model and also incorporate a within group spatial error component; a gaussian spatial correlation function with no nugget. They give the following results

```
Linear mixed-effects model fit by REML
 Data: leuk.dat
       AIC      BIC    logLik
  581.8942 607.2623 -283.9471

Random effects:
 Formula: ~1 | key
        (Intercept)   Residual
StdDev:   0.6547277 0.05629857

Correlation Structure: Gaussian spatial correlation
 Formula: ~x + y | key
 Parameter estimate(s):
     range
0.03396634
Fixed effects: z ~ exp + age65 + home
                Value Std.Error  DF   t-value p-value
(Intercept) -0.517276 0.1585572 277 -3.262396  0.0012
exp          0.048836 0.0350635 277  1.392795  0.1648
age65        3.950890 0.6054983 277  6.525022  0.0000
home        -0.560041 0.1703080 277 -3.288403  0.0011
 Correlation:
      (Intr) exp    age65
exp   -0.411
age65 -0.587 -0.075
home  -0.741  0.082  0.147

Number of Observations: 281
Number of Groups: 281
```

and discuss them as if there is no problem. Here are the intervals.

```
Approximate 95% confidence intervals

 Fixed effects:
                   lower        est.      upper
(Intercept) -0.82940652 -0.51727635 -0.2051462
exp         -0.02018850  0.04883627  0.1178610
age65        2.75892674  3.95088956  5.1428524
home        -0.89530376 -0.56004134 -0.2247789
attr(,"label")
[1] "Fixed effects:"
```

```
Random Effects:
  Level: key
                    lower      est.    upper
sd((Intercept)) 0.343092 0.6547277 1.249427

 Correlation structure:
             lower       est.        upper
range 8.947363e-39 0.03396634 1.289444e+35
attr(,"label")
[1] "Correlation structure:"

 Within-group standard error:
       lower         est.        upper
1.267645e-39 5.629857e-02 2.500328e+36
```

Note the intervals on the range parameter and on the within-group standard error.

*Generalized Linear Mixed Models*

- We present a general discussion first. We suppose that we have a vector $\mathbf{Z}$ of responses, a vector $\boldsymbol{\beta}$ of fixed effects, and a vector $\boldsymbol{\nu}$ of random effects. The assumption is made that the mean function is such that

$$g\left\{E\left[\mathbf{Z}|\boldsymbol{\nu}\right]\right\} = \mathbf{X}\boldsymbol{\beta} + \mathbf{U}\boldsymbol{\nu}$$

where $\mathbf{X}$ is a full rank design matrix for the fixed effects, $\mathbf{U}$ is a design matrix for the random effects, and $g$ is an appropriate link function. Unconditionally, the random effects are assumed to be normally distributed with mean $\mathbf{0}$ and variance-covariance matrix $\boldsymbol{\Sigma}_{\boldsymbol{\nu}}$. This type of model is sometimes referred to as a conditional model in the generalized linear mixed model literature. Conditional on the random effects we have

$$\text{Var}\left[\mathbf{Z}|\boldsymbol{\nu}\right] = \boldsymbol{\Sigma}\left(\boldsymbol{\theta}\right) = \sigma^2\mathbf{V}_{\boldsymbol{\mu}}^{1/2}\mathbf{R}\left(\boldsymbol{\theta}\right)\mathbf{V}_{\boldsymbol{\mu}}^{1/2}$$

although often $\mathbf{R}$ is taken to be $\mathbf{I}$.

- Waller and Gotway discuss generalized linear mixed models on pages 383 to 399. They use notation to emphasize the spatial setting and they consider only random intercept models of the form:

$$g\left\{E\left[\mathbf{Z}\left(\mathbf{s}\right)|\mathbf{S}\left(\mathbf{s}\right)\right]\right\} = \mathbf{X}\left(\mathbf{s}\right)\boldsymbol{\beta} + \mathbf{S}\left(\mathbf{s}\right)$$

where $\mathbf{S}$ is normally distributed with an appropriately specified covariance matrix. $\mathbf{S}$ is a latent smooth Gaussian random field and the responses are noisy measurements of that underlying process.

- The distribution of each response is thus conditioned on the underlying latent (unobserved) spatial process $\mathbf{S}\left(\mathbf{s}\right)$ which determines the spatial correlation structure. The latent process is typically assumed to be Gaussian with mean 0 and covariance function $\sigma_S^2\rho\left(\mathbf{s}_i - \mathbf{s}_j;\boldsymbol{\theta}\right)$. This process defines the "spatial similarity between observations." The assumption of normality for $S$ is separate from the assumption of the distribution of the random error terms associated with each observation in a *GLM* as quantified in $\boldsymbol{\Sigma}\left(\boldsymbol{\theta}\right)$.

  We consider the conditional model

$$E\left[Z\left(\mathbf{s}\right)|S\left(\mathbf{s}\right)\right] = \mu\left(\mathbf{s}\right)$$

with the link function now not only relating the mean to a linear combination of predictors but also to the underlying process $S$,

$$\begin{aligned} g\left(\mu\left(\mathbf{s}\right)\right) &= \beta_0 + \beta_1 X_1 + \cdots + \beta_{p-1}X_{p-1} + S\left(\mathbf{s}\right) \\ &= \left\{\beta_0 + S\left(\mathbf{s}\right)\right\} + \beta_1 X_1 + \cdots + \beta_{p-1}X_{p-1} \end{aligned}$$

Note that $S$

1. adds spatially structured noise to the mean and
2. is a random addition to the intercept term.

This is an example of a Generalized Linear Mixed Model ($GLMM$) with fixed effects $\mathbf{x}^{'}\boldsymbol{\beta}$ and random intercept adjustment $S$.

We assume that the observations are independent of one another given $S$ so that

$$\mathrm{Var}\left[Z\left(\mathbf{s}\right)|S\left(\mathbf{s}\right)\right]=\sigma^{2}v\left(\mu\left(\ \mathbf{s}\right)\right).$$

The marginal (fixed effect) model and conditional model are different both structurally and in interpretation. For example, letting $m\left(\mathbf{s}\right)=\exp\left[\mathbf{x}\left(\mathbf{s}\right)^{'}\boldsymbol{\beta}\right]$ with a log link function it can be shown that, for the mixed model

$$\begin{aligned}
E\left[Z\left(\mathbf{s}\right)\right] &= m\left(\mathbf{s}\right)\exp\left(\sigma_{S}^{2}/2\right)\\
\mathrm{Var}\left[Z\left(\mathbf{s}\right)\right] &= m\left(\mathbf{s}\right)\left[\sigma^{2}\exp\left(\sigma_{S}^{2}/2\right)+m\left(\mathbf{s}\right)\exp\left(\sigma_{S}^{2}\right)\left(\exp\left(\sigma_{S}^{2}\right)-1\right)\right]\\
\mathrm{Cov}\left[Z\left(\mathbf{s}_{i}\right),Z\left(\mathbf{s}_{j}\right)\right] &= m\left(\mathbf{s}_{i}\right)m\left(\mathbf{s}_{j}\right)\exp\left(\sigma_{S}^{2}\right)\left\{\exp\left[\sigma_{S}^{2}\rho_{S}\left(\mathbf{s}_{i}-\mathbf{s}_{j}:\boldsymbol{\theta}\right)\right]-1\right\}
\end{aligned}$$

which contrasts with the results from the fixed effects model:

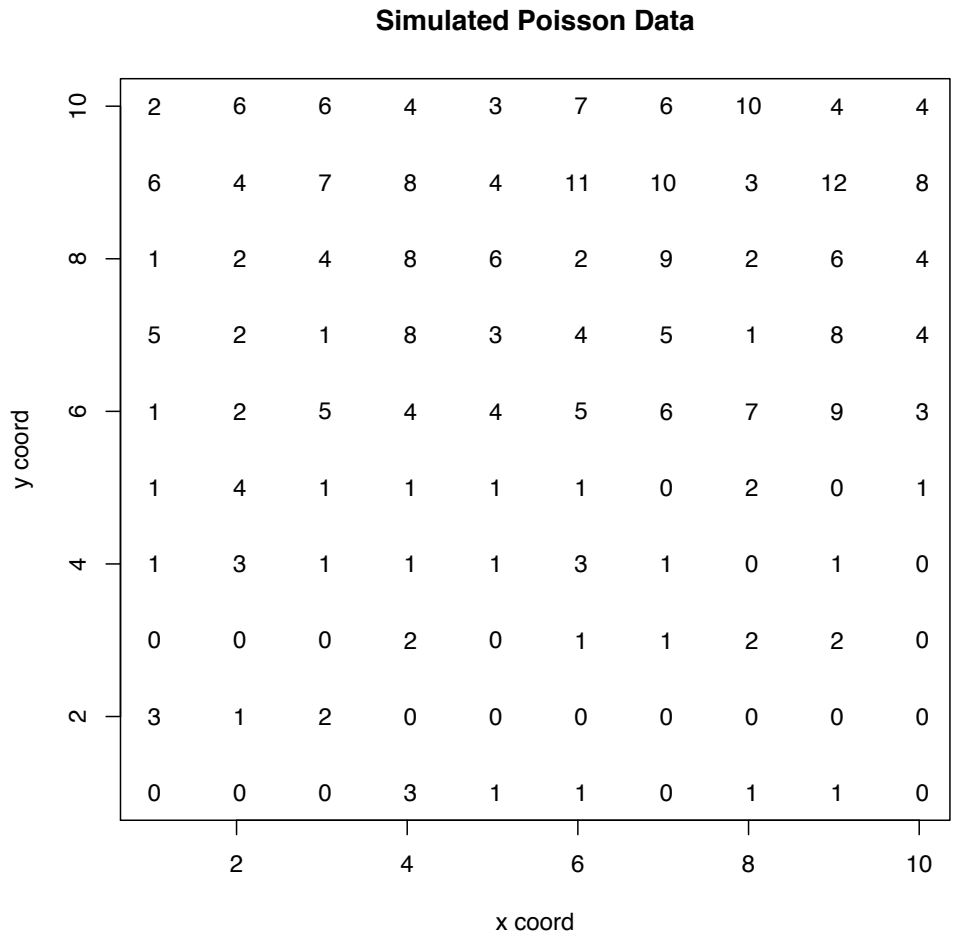
$$\begin{aligned}
E\left[Z\left(\mathbf{s}\right)\right] &= m\left(\mathbf{s}\right)\\
\mathrm{Var}\left[Z\left(\mathbf{s}\right)\right] &= \sigma^{2}m\left(\mathbf{s}\right)\\
\mathrm{Cov}\left[Z\left(\mathbf{s}_{i}\right),\left(Z\left(\mathbf{s}_{j}\right)\right)\right] &= \sigma^{2}\sqrt{m\left(\mathbf{s}_{i}\right)m\left(\mathbf{s}_{j}\right)}\rho\left(\mathbf{s}_{i}-\mathbf{s}_{j};\boldsymbol{\theta}\right)
\end{aligned}$$

- The model under consideration may seem strange. Here is a simple example of one using some functions available in `geoR`. This example is given in some documentation for the R package `geoRglm` by Christensen and Ribeiro. First we simulate "data" from a Gaussian random field.

```
sim.g<-grf(grid=expand.grid(x=seq(1,10,l=10),y=seq(1,10,l=10)),cov.pars=c(0.1,0.2))
```

We want to simulate a Poisson model with a log link.

```
> sim <- list(coords = sim.g$coords, units.m = c(rep(1,
+ 50), rep(5, 50)))
> attr(sim, "class") <- "geodata"
> sim$data <- rpois(100, lambda = sim$units.m * exp(sim.g$data))
> plot(sim$coords[, 1], sim$coords[, 2], type = "n")
> text(sim$coords[, 1], sim$coords[, 2], format(sim$data))
```




Letting $\mathbf{Z}(\mathbf{s})$ be the response we have

$$\mathbf{Z}(\mathbf{s}) | \mathbf{S}(\mathbf{s}) \sim \text{Poisson}(\lambda(\mathbf{s}))$$

$$\log(\lambda(\mathbf{s})) = \beta_0 + \beta_1 X(\mathbf{s}) + S(\mathbf{s})$$

$$\mathbf{S}(\mathbf{s}) \sim N\left(\mathbf{0}, \sigma_s^2 \mathbf{R}_S\right)$$

where $\sigma_S^2 = 0.1$ is the sill and $\mathbf{R}_S$ is an exponential correlation matrix with a range of 0.2. The predictor $X$ is a $100 \times 1$ vector of 50 0's and 50 log(5)'s (note $\beta_0 = 0$ in this example).

- *Estimation*: As we indicated above in our discussion of fixed effects GLMs maximum likelihood estimation requires specification of a joint distribution, which is impossible when observations are spatially correlated with one another. Even in the conditional setting where it is theoretically possible (conditional on $S$) maximum likelihood estimation presents formidable problems (see page 359 in Schabenberger and Gotway). Likelihood *like* approaches are generally used. The psuedo-likelihood method discussed earlier is repeated below with notational changes to incorporate the random effects. In what follows we will let $\mu_i = \mu(\mathbf{s}_i)$ and $Z_i = Z(\mathbf{s}_i)$.
  - *Pseudolikelihood Estimation*: The term pseudolikelihood is used in a confusingly large number of ways in statistics. In the context in which we are working it is introduced as a method of producing estimates of parameters in a conditional $GLMM$ although it is also possible to get psedolikelihood estimators of parameters in a marginal model. Pseudolikelihood differs from quasi-likelihood in that true joint likelihoods are being used to estimate parameters at each step in the iterative process.

A first order Taylor series expansion of $g$ about $\mu$ is used to create pseudo-data:

$$Z_i^{(p)} = g\left(\widehat{\mu}_i\right) + g^{'}\left(\widehat{\mu}_i\right)\left(Z_i - \widehat{\mu}_i\right)$$

where $\widehat{\mu}$ is a current estimate of $\mu$. For the conditional model, given some not unreasonable assumptions

$$\begin{aligned} E\left[\mathbf{Z}^{(p)}|\boldsymbol{\beta},\mathbf{S}\right] &= \mathbf{X}\boldsymbol{\beta}+\mathbf{S} \\ \text{Var}\left(\mathbf{Z}^{(p)}|\boldsymbol{\beta},\mathbf{S}\right) &= \boldsymbol{\Sigma}_{\widehat{\boldsymbol{\mu}}} \end{aligned}$$

where

$$\boldsymbol{\Sigma}_{\widehat{\boldsymbol{\mu}}} = \sigma^2 \boldsymbol{\Delta}_{\widehat{\boldsymbol{\mu}}} \mathbf{V}_{\widehat{\boldsymbol{\mu}}}^{1/2} \mathbf{R} \mathbf{V}_{\widehat{\boldsymbol{\mu}}}^{1/2}$$

with $\boldsymbol{\Delta}_{\widehat{\boldsymbol{\mu}}}$ a diagonal matrix with diagonal elements $\delta_{ii} = \partial g\left(\mu_i\right)/\partial\mu_i$ evaluated at $\widehat{\boldsymbol{\mu}}$. The marginal (unconditional) mean and variance of the pseudo-data are

$$\begin{aligned} E\left[\mathbf{Z}^{(p)}\right] &= \mathbf{X}\boldsymbol{\beta} \\ \text{Var}\left(\mathbf{Z}^{(p)}\right) &= \boldsymbol{\Sigma}_S + \boldsymbol{\Sigma}_{\widehat{\boldsymbol{\mu}}} \end{aligned}$$

with the $(i,j)$th element of $\boldsymbol{\Sigma}_S$ being $\sigma_S^2 \rho_S\left(\mathbf{s}_i - \mathbf{s}_j\right)$. In other words, we have an ordinary linear regression model with spatially correlated errors which can be fit using generalized least squares. Define $\boldsymbol{\Sigma}_{\mathbf{Z}^{(p)}} = \boldsymbol{\Sigma}_S + \boldsymbol{\Sigma}_{\widehat{\boldsymbol{\mu}}}$ and proceed as follows:

1. Compute an initial estimate of $\widehat{\boldsymbol{\mu}}$ using a nonspatial $GLM$.
2. Compute the pseudodata.
3. Use maximum likelihood (or $REML$) with the pseudodata to get estimates of the spatial correlation parameters and $\sigma_S^2$.
4. Use these estimates to get generalized least squares estimates of $\boldsymbol{\beta}$ and $\sigma^2$ and, with a conditional model, to predict $\mathbf{S}$:

$$\begin{aligned} \widehat{\boldsymbol{\beta}} &= \left(\mathbf{X}\boldsymbol{\Sigma}_{\mathbf{Z}^{(p)}}^{-1}\mathbf{X}\right)^{-1}\mathbf{X}^{'}\boldsymbol{\Sigma}_{\mathbf{Z}^{(p)}}^{-1}\mathbf{Z}^{(p)} \\ \widehat{\mathbf{S}} &= \boldsymbol{\Sigma}_S\boldsymbol{\Sigma}_{\mathbf{Z}^{(p)}}^{-1}\left(\mathbf{Z}^{(p)} - \mathbf{X}\widehat{\boldsymbol{\beta}}\right) \\ \widehat{\sigma}^2 &= \frac{1}{n}\left(\mathbf{Z}^{(p)} - \mathbf{X}\boldsymbol{\beta}\right)^{'}\boldsymbol{\Sigma}_{\mathbf{Z}^{(p)}}^{-1}\left(\mathbf{Z}^{(p)} - \mathbf{X}\boldsymbol{\beta}\right) \end{aligned}$$

5. Use the inverse link function to update $\widehat{\boldsymbol{\mu}}$:

$$\widehat{\boldsymbol{\mu}} = g^{-1}\left(\mathbf{X}\widehat{\boldsymbol{\beta}} + \widehat{\mathbf{S}}\right).$$

6. Repeat until convergence.

Letting $\widehat{\boldsymbol{\Sigma}}_{\mathbf{Z}^{(p)}} = \boldsymbol{\Sigma}_S + \boldsymbol{\Sigma}_{\widehat{\boldsymbol{\mu}}}$ the standard errors for the fixed effects can be approximated as

$$\widehat{\text{Var}}\left(\widehat{\boldsymbol{\beta}}\right) = \widehat{\sigma}^2\left(\mathbf{X}^{'}\widehat{\boldsymbol{\Sigma}}_{\mathbf{Z}^{(p)}}^{-1}\mathbf{X}\right)^{-1}$$

where $\widehat{\boldsymbol{\Sigma}}_{\mathbf{Z}^{(p)}}^{-1}$ is defined to be $\boldsymbol{\Sigma}_{\mathbf{Z}^{(p)}}^{-1}$ using the final converged solution of the spatial correlation parameters $\widehat{\boldsymbol{\theta}}$.

- We now have 2 choices for modeling spatial autocorrelation: we can do it through $\mathbf{R}$ in a marginal or fixed effects model (with $\mathbf{S} = \mathbf{0}$) or $\boldsymbol{\Sigma}_S$ in the conditional mixed effects model (with $\mathbf{R} = \mathbf{I}$). Actually we could theoretically try to use both but we will keep it simple.

- *Example: This example involves a famous data set - the Scottish lip cancer data. This data set is discussed in Waller and Gotway on pages 393 to 399 and we will go over it in more detail below (or I will assign it to you in some form). The data set consists of both observed and expected numbers of lip cancer cases in* 56 *counties in Scotland between* 1975 *and* 1980*. The simple model below is actually a case study discussed in the SAS documentation of PROC GLIMMIX. Let $Z_i$ be the number of incident cases observed in county $i$. Let $E_i$ be the expected number of cases in county $i$. The standardized morbidity ratio (SMR) for county $i$ is $r_i = Z_i/E_i$. Associated with each county is an unobserved spatially varying relative risk $\gamma_i$. These are assumed to be random effects (the S terms in the discussion above).*

  *Conditional on $\gamma_i$, the observed counts are independent Poisson random variables with mean $\mu_i = E_i\gamma_i$. An elementary mixed model for $Z_i$ specifies only a random intercept for each county, in addition to a fixed intercept. We incorporate a covariate that measures the percentage of employees in agriculture, fishing, and forestry. This variable is a surrogate for sun exposure. The expanded conditional model for the observed counts is*

$$\mu_i = \exp\left(\log(E_i) + \beta_0 + \beta_1 x_i/10 + \gamma_i\right), \quad i = 1, \cdots, 56$$

  *where $\beta_0$ and $\beta_1$ are fixed effects, and the $\gamma_i$ are county random effects. We assume that the random effects are normally distributed with mean $\mathbf{0}$ and covariance matrix $\mathbf{\Sigma}_\gamma(\boldsymbol{\theta})$. For this example, we simply assume that $\mathbf{\Sigma}_\gamma = \sigma^2_\gamma \mathbf{I}$. The term $\log(E_i)$ is referred to in the literature as an offset term and is assumed to be a covariate with know coefficient of* 1*. This is an example of a rate model.*

  *The SAS code is:*

```
proc glimmix data=lipcancer;
class county;
x = employment / 10;
logn = log(expCount);
model observed = x / dist=poisson offset=logn
solution ddfm=none;
random county;
run;
```

  *The output is*

Fit Statistics

| | |
|---|---|
| -2 Res Log Pseudo-Likelihood | 127.51 |
| Generalized Chi-Square | 55.99 |
| Gener. Chi-Square / DF | 1.04 |

Covariance Parameter Estimates

| Cov Parm | Estimate | Standard Error |
|---|---|---|
| county | 0.3567 | 0.09869 |

Solutions for Fixed Effects

| Effect | Estimate | Standard Error | DF | t Value | Pr > \|t\| |
|---|---|---|---|---|---|
| Intercept | -0.4406 | 0.1572 | Infty | -2.80 | 0.0051 |
| x | 0.6799 | 0.1409 | Infty | 4.82 | <.0001 |

*Note the* `Fit Statistics`. *The parameters were estimated using the psuedo-likelihood method and the usual model comparison techniques are not applicable. There is not even an ad hoc psuedo-AIC statistic we can use. The estimate of* $\sigma^2_\gamma$ *is* $0.3567$ *with a standard error of around* $0.1$ *providing fairly strong evidence (if all the numerous assumptions are reasonable) of county to county variability in risk of lip cancer. The estimates of the fixed effects suggests that the risk of lip cancer increases with an increase in the proportion of people working in outdoor jobs (proportion of people exposed to sun).*

- *Example: Blood Lead Levels in Children*

  *We will look at 3 models:*

  1. *Random Effects Model: The model is a conditional model with* $Z(\mathbf{s}_i)|S(\mathbf{s}_i) \sim Poi(n(\mathbf{s}_i)\lambda(\mathbf{s}_i))$ *where*

  $$\log\{\lambda(\mathbf{s}_i)\} = \beta_0 + \beta_1 X_1(\mathbf{s}_i) + \beta_2 X_2(\mathbf{s}_i) + S(\mathbf{s}_i).$$

  *The conditional variance is*

  $$Var[\mathbf{Z}(\mathbf{s})|\mathbf{S}(\mathbf{s})] = \sigma^2 \mathbf{V}_{\boldsymbol{\mu}}$$

  *and the latent spatial process* $S$ *is distributed as*

  $$\mathbf{S}(\mathbf{s}) \sim N\left(\mathbf{0}, \sigma^2_S \mathbf{I}\right).$$

  *The assumption is that there is no spatial autocorrelation but that each county has its own random intercept* $\beta_0 + S(\mathbf{s})$. *This is analogous to the lip cancer model we discussed above. The results are shown below.*

| `Effect` | `Estimate` | `Std. Error` | `t-value` | `p-value` |
|---|---|---|---|---|
| `Intercept` | $-2.6112$ | $0.1790$ | $-14.59$ | $< 0.0001$ |
| `Median Value` | $-0.2672$ | $0.2239$ | $-1.19$ | $0.2349$ |
| `Poverty` | $-0.9688$ | $2.900$ | $-0.33$ | $0.7389$ |
| $\widehat{\sigma}^2_S = 0.5047$ | $0.1478$ | | | |
| $\widehat{\sigma}^2 = 1.1052$ | $0.3858$ | | | |

  *The magnitude of the estimated coefficients is comparable to that seen in the fixed effects model although the partial z-tests for the 2 predictors are not significant. The estimate of the intercept* $\widehat{\beta}_0 = -2.6112$ *is an estimate of the mean intercept of all 133 counties and* $\widehat{\sigma}^2_S = 0.5047$ *is an estimate of the variability around that mean. The standard error of* $0.148$ *suggests that we have significant county to county variation in blood lead levels. Of course the fixed effects model suggested this also but the county to county variation was explained in part by the median value of homes. The estimate of the dispersion parameter* $\sigma^2$ *is much smaller in the mixed model but keep in mind that this is a conditional dispersion parameter whereas in the fixed effects model it is a marginal dispersion parameter.*

  2. *Conditional Spatial GLMM Model: The model assumes that conditional on a latent spatial process* $S$

  $$Z(\mathbf{s}_i)|S(\mathbf{s}_i) \sim Poi(n(\mathbf{s}_i)\lambda(\mathbf{s}_i))$$

  *with*

  $$\log\{\lambda(\mathbf{s}_i)\} = \beta_0 + \beta_1 X_1(\mathbf{s}_i) + \beta_2 X_2(\mathbf{s}_i) + S(\mathbf{s}_i).$$

  *The conditional variance is*

  $$Var(Z(\mathbf{s}_i)|S(\mathbf{s}_i)) = \sigma^2 \mathbf{V}_{\boldsymbol{\mu}}$$

  *and* $\mathbf{S}(\mathbf{s}_i) \sim G\left(\mathbf{0}, \sigma^2_S \mathbf{R}_S(\boldsymbol{\theta})\right)$. *A spherical correlation model was used for* $\mathbf{R}_S$. *The results are*

| `Effect` | `Estimate` | `Std. Error` | `t-value` | `p-value` |
|---|---|---|---|---|
| `Intercept` | $-2.4246$ | $0.3057$ | $-7.93$ | $< 0.0001$ |
| `Median Value` | $-0.7835$ | $0.3619$ | $-2.17$ | $0.0318$ |
| `Poverty` | $3.5216$ | $2.5933$ | $1.36$ | $0.1768$ |
| $\widehat{\sigma}^2_S$ | $0.8806$ | $0.1949$ | | |
| $\widehat{\sigma}^2$ | $0.7010$ | $0.1938$ | | |
| `range par` | $78.81$ | $6.51$ | | |

*This model assigns spatial structure to the random effects, i.e. it assumes a spatially correlated latent process. The residual process is assumed to be uncorrelated with overdispersion parameter* $\sigma^2$*. The mean intercept is estimated to be* $\widehat{\beta}_0 = -2.425$ *with normally distributed errors having variance estimated to be* $\widehat{\sigma}_S^2 = 0.881$*, larger than in the mixed effects model with no spatial correlation. We see much stronger evidence for an effect due to median housing value whereas poverty is still seemingly not important.*

3. *Marginal Spatial GLMM: The model is* $Z(\mathbf{s}_i) \sim Poi(n(\mathbf{s}_i)\lambda_i(\mathbf{s}_i))$ *where*

$$\log\{\lambda(\mathbf{s}_i)\} = \beta_0 + \beta_1 X_1(\mathbf{s}_i) + \beta_2 X_2(\mathbf{s}_i)$$

*and*

$$v(\mu(\mathbf{s}_i)) = \mu(\mathbf{s}_i).$$

*The covariance matrix of* $\mathbf{Z}$ *is*

$$Var(\mathbf{Z}(\mathbf{s})) = \sigma_0^2 \mathbf{V}_{\boldsymbol{\mu}} + \sigma_1^2 \mathbf{V}_{\boldsymbol{\mu}}^{1/2} \mathbf{R}(\boldsymbol{\theta}) \mathbf{V}_{\boldsymbol{\mu}}^{1/2}.$$

*The results are*

| Effect | Estimate | Std. Error | t-value | p-value |
|---|---|---|---|---|
| Intercept | $-2.2135$ | 0.2010 | $-11.01$ | $< 0.0001$ |
| Median Value | $-0.8096$ | 0.2713 | $-2.98$ | 0.0034 |
| Poverty | 1.3172 | 1.5736 | 0.84 | 0.4041 |
| $\widehat{\sigma}_0^2$ | 6.1374 | 0.8846 | | |
| $\widehat{\sigma}_1^2$ | 0.9633 | 0.7815 | | |
| range par | 186.65 | 70.04 | | |

*These results are similar to the marginal Poisson model we fit earlier with no spatial autocorrelation. Note that the overdispersion parameter in the spatial GLMM is* $\widehat{\sigma}_0^2 + \widehat{\sigma}_1^2 = 7.1007$ *as compared to the estimate of* 6.7825 *above. Note the lack of evidence for a partial sill greater than 0 (i.e. the lack of evidence for anything other than a pure nugget effect model) although the evidence for the range parameter being greater than 0 is strong enough that many would claim it to be "significant". Frankly, though it is not clear at all that more complex models incorporating spatial correlation are necessary.*

- *Summary and Discussion of Results*: We have looked at the results from 4 models. The main results for the Poisson models are summarized in the table below.

| Model | $\widehat{\beta}_0$ | $\widehat{\beta}_1$ | $\widehat{\beta}_2$ | $\widehat{\sigma}_S^2$ | $\widehat{\sigma}^2$ |
|---|---|---|---|---|---|
| GLM | $-2.28 \pm 0.06$ | $-0.62 \pm 0.08$ | $0.63 \pm 0.54$ | | |
| GLM-D | $-2.28 \pm 0.16$ | $-0.62 \pm 0.21$ | $0.63 \pm 1.41$ | | 6.78 |
| GLMM-RE | $-2.61 \pm 0.18$ | $-0.27 \pm 0.22$ | $-0.97 \pm 2.90$ | 0.51 | 1.11 |
| GLMM-C | $-2.43 \pm 0.31$ | $-0.78 \pm 0.36$ | $3.52 \pm 2.59$ | 0.88 | 0.79 |
| GLM-M | $-2.21 \pm 0.20$ | $-0.81 \pm 0.27$ | $1.32 \pm 1.57$ | | 7.10 |

A summary table can be bit misleading because the parameter estimates have different interpretations depending on the model.

1. The coefficient estimates are identical in the nonspatial fixed effects models but the standard errors are much higher when overdispersion is accounted for.
2. The estimates for $\beta_1$ and $\beta_2$ in GLMM-RE are quite different from the other models. This is the only model that does not exhibit a significant effect for median value of a home. The sign on the poverty index coefficient is in the wrong direction but the estimate has a very high standard error associated with it. The variance component $\sigma_S^2$ appears to be significant (standard error of 0.15). Which of models GLM-D and GLMM-RE is best? They may in fact be equivalent in their predictive ability, i.e. they may both be *valid* models (we do not explore that here). GLM-D explains the variability primarily through the mean structure and GLMM-RE explains it primarily through the random effects structure.

3. The coefficient estimates associated with the spatial models are closer to those of GLM-D. GLMM-RE is essentially nested in GLMM-C with different assumptions made about the spatial structure (or lack of it) in the latent process $S(\mathbf{s})$. We do not have any easy way to compare the models (i.e. AIC or Likelihood Ratio Test). We could base a comparison on predictive ability (using cross-validation for example). Note that the standard errors in GLMM-C are larger than in GLMM-RE.
4. GLM-D is nested in GLM-M. Frankly, these 2 models appear to be quite close to one another although the predictions do differ a bit at the county level.

These results may leave you scratching your head. Waller and Gotway have this to say about the matter (page 396) and these comments are relevant for all the spatial regression models we have discussed:

*All regression models partition the variation in the data into variation that can be attributed to systematic changes in fixed covariates and the remaining random variation. When we are modeling spatial variation with spatially varying covariates, this partitioning is not unique and it can be difficult to decide what part of the spatial variation belongs to the covariates and what part should be treated as residual autocorrelation. The variation that one attributes to the covariates may be attributed to random variation in another model, and both models may be valid! This is even more difficult with the inclusion of spatially structured random effects, since we now have to partition our variation into three parts instead of two. If the spatial variation in the random effects is related to that in the covariates, some of the covariate effect will be assigned to the random effects. Thus, we have to take great care in interpreting the results from spatial regression models by understanding how the components of our models can affect our interpretation. If the signal in the data is strong, all models should give similar conclusions, even if the particular values of parameter estimates and standard errors differ.*

- *Marginal (fixed effects) or Conditional (mixed) Models?*: The 2 types of models are different. Some things to consider:
  1. Marginal models are sometimes referred to as *population averaged* models because they describe the relationship between mean response and selected covariates. Conditional models are *subject-specific*.
  2. The existence of a latent spatially structured process may or may not make sense in a particular application in which case a marginal model will be more appropriate.
  3. Is covariate interpretation really of interest? If the goal is a spatially smoothed map of local counts or rates adjusted for known confounding variables then a conditional $GLMM$ may be more appropriate than a marginal model which will tend to oversmooth. It can be shown that the fitted means from a conditional model are equivalent to smoothed empirical Bayes estimates of mean responses.

- Autoregressive Models: The above spatial models all took a geostatistical approach to modeling spatial correlation. The lead data though are clearly regional, i.e. this is lattice data and geostatistically motivated correlation functions may not be the most appropriate measures of spatial nearness. We discussed briefly the use of SAR and CAR models for spatial linear regression models. Might not that be a more appropriate approach here? We can specify autoregressive models for the latent spatial process $\mathbf{S}(\mathbf{s})$. We simply choose $\mathbf{\Sigma}_S = \sigma^2 (\mathbf{I} - \rho\mathbf{W})^{-1}$ for a CAR model and $\mathbf{\Sigma}_S = \sigma^2 (\mathbf{I} - \rho\mathbf{W})^{-1} \left(\mathbf{I} - \rho\mathbf{W}'\right)^{-1}$ for a SAR model. We are not modeling the *data* as a CAR process (there are sizable theoretical problems with attempting to do so). A marginal version of such a model would have $\mathbf{R}$ equal to the SAR or CAR covariance models above. Schabenberger and Gotway carry out a marginal CAR analysis of the Virginia lead data using a binary connectivity matrix constructed assuming the "Queen's" definition of contiguity. The model is

$$E[Z(\mathbf{s})] = \mu(\mathbf{s}) = n(\mathbf{s})\lambda(\mathbf{s})$$
$$\log[\lambda(\mathbf{s})] = \beta_0 + \beta_1 X_1(\mathbf{s}) + \beta_2 X_2(\mathbf{s})$$
$$\text{Var}[\mathbf{Z}(\mathbf{s})] = \sigma_0^2 \mathbf{V}_{\boldsymbol{\mu}} + \sigma_1^2 \mathbf{V}_{\boldsymbol{\mu}}^{1/2} (\mathbf{I} - \rho\mathbf{W})^{-1} \mathbf{V}_{\boldsymbol{\mu}}^{1/2}$$

  where $\mathbf{W}$ is the spatial contiguity matrix. They provide SAS code but it is complicated. They fit a nonspatial Poisson model to generate starting values for fitting the CAR model and then use those in a SAS program they wrote using PROC IML. The results are shown below.

```
--------------- Final Estimates --------------

      Estimate   StdErr   t-value    p-value

    Intercept -2.14722 0.179643    -11.95   <0.00001
     mval1000  -0.75326 0.247112     -3.05    0.0014
     pov1000   0.585616 1.441671      0.40    0.6574
     rho       0.099238        .
     s2_1      6.249709        .
     s2_0             0        .

----------------------------------------------
```

They do not discuss the fitting results or why the standard errors are missing for the variance components and $\widehat{\rho}$. The basic results are similar to those seen earlier. As Schabenberger and Gotway and Waller and Gotway have both argued it is more important to try to account for spatial correlation in some defensible way than to ignore it.

- We have seen a number of ways to model this data and come to no conclusion about which is "best". Certainly one comparison to make with data such as these is to look at maps of predicted values. For the marginal (fixed effects) Poisson models these are easy to get as the fitted values are

$$\widehat{\boldsymbol{\mu}} = \exp\left(\mathbf{X}(\mathbf{s})\widehat{\boldsymbol{\beta}}\right)$$

and for the conditional (mixed effects) models we have

$$\widehat{\boldsymbol{\mu}} = \exp\left(\mathbf{X}(\mathbf{s})\widehat{\boldsymbol{\beta}} + \widehat{\mathbf{S}}(\mathbf{s})\right).$$

I will have a handout in class showing maps from some of the models.

- First we will look at a short "primer" of Bayesian statistics focusing mostly on the Binomial and Poisson distributions. We will also look at methods for sampling from posterior distributions. The material below may raise more questions than it answers for you. Just as in the mixed models presentation above we will be leaving out a lot of detail and if you want to explore that detail more then you need to take a course in Bayesian statistics such as STAT 532 which will be taught next fall. Also, the notation below will not match that in your text.

- Bayesian Statistics: One well known text states: "The essential characteristic of Bayesian methods is their explicit use of probability for quantifying uncertainty in inferences based on statistical data analysis." This may seem a strange statement. After all, don't all statistical methods use probability explicitly to quantify uncertainty? Yes and No.

  *Example: Let $y_i, i = 1, \cdots, n$ be independent Bernoulli random variables with $\theta = P(y_i = 1), i = 1, \cdots, n$. The "frequentist" approach to the estimation of $\theta$ is to assume that it is an unknown but fixed parameter. We know that $y = \sum y_i \sim Bin(n, \theta)$. The best estimator of $\theta$ (in some sense) is*

$$\widehat{\theta} = \frac{y}{n}.$$

  *If $n$ is large enough then, by the Central Limit Theorem,*

$$\widehat{\theta} \dot{\sim} N\left(\theta, \frac{\theta(1-\theta)}{n}\right).$$

  *An approximate large sample 95% confidence interval for $\theta$ is*

$$\widehat{\theta} \pm 1.96\sqrt{\frac{\widehat{\theta}(1-\widehat{\theta})}{n}}.$$

  *Over the long run this interval will capture the true value of $\theta$ approximately 95% of the time and it will miss approximately 5% of the time. This property does not depend on $\theta$. However, in a single trial, i.e. after collecting data and computing the bounds for the interval we can no longer make probability statements. We say instead that we are approximately 95% confident that the computed interval contains $\theta$ because the process that produced the interval works approximately 95% of the time. The only place we have explicitly used probability is in specification of the probability model for $y$. However, following data collection, we do not explicitly use probability to quantify our uncertainty about $\theta$ relying instead on* confidence.

  Bayesian statistics differs from the frequentist approach by setting up, in the words of the text, "a *full probability model* for all observable and unobservable quantities in a problem." The observable quantities in the above example are the data $(y_i, i = 1, \cdots, n)$ and the unobservable quantity is $\theta$. We subsequently use the observed data to obtain a conditional probability distribution for $\theta$. All information about the unknown $\theta$ is contained in this conditional distribution. A computed Bayesian interval estimate for $\theta$, a computation based on the post-data conditional distribution of $\theta$, would have a probabilistic interpretation.

- Uncertainty about a parameter can be quantified through the use of a probability distribution. The parameter could be "fixed but unknown," as the frequentist says, but even if the parameter is a constant, our knowledge (or lack of knowledge) can be expressed using a probability distribution.

  Let $f(\cdot|\cdot)$ denote a conditional probability density and $f(\cdot)$ denote a marginal density.

  There are 4 probability distributions to think about.

  1. The prior distribution for $\theta$, $f(\theta)$. Strong prior beliefs make for tight (high precision, low variance) prior distributions. Lack of knowledge would make us choose non-informative (unprecise, high variance) priors. Ideally, we would like final results to be robust to the choice of different priors.
  2. The data have a distribution which depends on the parameters, $f(y|\theta)$. Viewed as a function of the parameters for fixed data, this is called the likelihood.

3. The marginal distribution for the data,

$$f(y) = \begin{cases} \sum_\theta f(\theta) f(y|\theta) & \theta \text{ discrete} \\ \int f(\theta) f(y|\theta) d\theta & \theta \text{ continuous} \end{cases}$$

We are summing (or integrating) $\theta$ out of the joint distribution $f(y, \theta)$. .

4. The posterior distribution of $\theta$ which is, by Bayes Rule

$$f(\theta|y) = \frac{f(y|\theta) f(\theta)}{f(y)}.$$

Recognizing that $f(y)$ is a constant we will sometimes write the unnormalized posterior distribution,

$$f(\theta|y) \propto f(y|\theta) f(\theta).$$

The posterior distribution $f(\theta|y)$ is the ultimate goal of a Bayesian analysis and it contains all the information about $\theta$.

- *Example:*

*In 1786, Laplace wanted to estimate the proportion of births of male children in Paris. He had the data:*

| | |
|---|---|
| *Males:* | *251527* |
| *Females:* | *241945* |
| *Total* | *493472* |

*We will follow his assumption that the number of male births, given $\theta$ = proportion male, is $Bin(493472, \theta)$. The likelihood is:*

$$p(y = 251527|\theta) = \binom{493472}{251527} \theta^{251527} (1-\theta)^{241945}.$$

*In general, we might assume a Beta prior distribution, Beta($\alpha$, $\beta$) where $\alpha = \beta$. Equality of the prior parameters ensures that we are not biased toward lower or higher proportions of male births.*

*Posterior:*

$$f(\theta|y) = \frac{f(y|\theta) f(\theta)}{\int_0^1 f(y|\theta) f(\theta) d\theta} \quad (1)$$

$$= \frac{\binom{n}{y} \theta^y (1-\theta)^{n-y} \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)} \theta^{\alpha-1} (1-\theta)^{\beta-1}}{constant} \quad (2)$$

*The kernel of the density is:*

$$\theta^{y+\alpha-1} (1-\theta)^{n-y+\beta-1}$$

*We recognize this as a Beta distribution with parameters: $y + \alpha$ and $n - y + \beta$. The posterior mean is $(y + \alpha)/(n + \alpha + \beta)$ and the posterior variance is*

$$\frac{(y+\alpha)(n-y+\beta)}{(n+\alpha+\beta)^2 (n+\alpha+\beta+1)}.$$

*Note that the prior and posterior are both Beta distributions. This is a special case - we say that the Beta family is a conjugate family of prior distributions for the binomial parameter. For large n, the values chosen for prior parameters $\alpha$ and $\beta$ make little difference in the posterior. This is the ideal result in that for large n (lots of data) the posterior is robust to the choice of prior parameters.*

*A reasonable and noninformative prior is a Uniform(0,1) distribution which is also a Beta(1,1). If we place a Uniform (0,1) prior on $\theta$ then the posterior distribution of $\theta$ is Beta(251528,241946). Are there more female births than male births? Using R we find*

$$P(\theta \leq 1/2|y) \approx 0$$

*We are actually performing a Bayesian test of the hypothesis that $\theta \leq 1/2$ and we can say that the probability the hypothesis is true is approximately 0, which is how everyone (not just STAT 216 students) want to interpret a p-value.*

*We would compute an approximate frequentist 95%* **confidence** *interval for $\theta$ to be $(0.5083, 0.5111)$ and a 95% Bayesian* **credible** *interval for $\theta$ is $(0.5083, 0.5111)$. They are identical, but the interpretation is very different. For the confidence interval we say that we are approximately 95% confident that $\theta$ lies in the interval because we know that around 95% of all intervals computed using this procedure will contain $\theta$. But $\theta$ either lies in the interval or it does not. For the credible interval the interpretation is that the probability $\theta$ lies in the interval is 95%.*

*Note: I found the Bayesian credible interval by finding the 2.5th and 97.5th of the posterior distribution.*

```
> qbeta(.975,251528,241946)
[1] 0.5111035
> qbeta(.025,251528,241946)
[1] 0.5083139
```

- Poisson model: Recall that if $y \sim Poisson(\theta)$ then

$$f(y \mid \theta) = \frac{\theta^y \exp(-\theta)}{y!}, \quad y = 1, 2, \cdots$$

If we have $n$ iid observations $y = (y_1, y_2, \cdots, y_n)$ then

$$f(\ y \mid \theta) = \prod_{i=1}^{n} \frac{\theta^{y_i} \exp(-\theta)}{y_i!} \propto \theta^{t(y)} \exp(-n\theta)$$

where $t(y) = \sum_{i=1}^{n} y_i$.

- It can be shown that the Gamma family of distributions is conjugate for the Poisson parameter and the posterior distribution is then also Gamma. Denoting the prior $\theta \sim Gamma(\alpha, \beta)$. Then

$$\theta \mid y \sim Gamma(\alpha + n\overline{y}, \beta + n)$$

- Rate/exposure models: We assume that

$$y_i \sim Poisson\,(x_i\theta)$$

where $\theta$ is a rate parameter and $x_i$ the exposure of the $i$th unit. The likelihood is

$$p\,(y \mid \theta) \propto \prod_{i=1}^{n} (x_i\theta)^{y_i} \exp\,(-x_i\theta) \propto \theta^{t(y)} \exp\left(-\theta \sum_{i=1}^{n} x_i\right)$$

where $t(y) = \sum y_i$. Again, the gamma distribution is conjugate and with prior $\theta \sim Gamma(\alpha, \beta)$ the posterior is

$$\theta \mid y \sim Gamma\left(\alpha + \sum_{i=1}^{n} y_i, \beta + \sum_{i=1}^{n} x_i\right).$$

- *Example: This is the example discussed on pages 53-55 in the Bayesian text referred to above. We will go over it in some detail.*

*An analysis of causes of death finds that 3 people in a city of $200,000$ people died of asthma. The crude death rate is 1.5 per $100,000$ people. We assume that the number of deaths follows a Poisson distribution with parameter $\theta$, where $\theta$ is the death rate per $100,000$. The exposure for the city is 2 because there are $200,000$ people in the city. We have observed $y = 3$ so the likelihood is*

$$p(y = 3 \mid \theta) = \frac{(2\theta)^3 \exp\,(-2\theta)}{6}.$$

*Quite a lot is known about asthma mortality rates. In particular it is known that typical rates in Western countries are around* $0.6$ *deaths per* $100,000$ *people. Assuming that this is a typical city, then a prior that gives a mean of* $0.6$ *would be reasonable. Further, it is known that rates higher than* $1.5$ *per* $100,000$ *are quite rare so little of the mass of the distribution should be in the upper tail of the prior distribution. The authors chose a* $Gamma(\alpha = 3.0, \beta = 5.0)$ *distribution with a mean of* $0.6$*, mode of* $0.4$ *and a variance of* $0.12$*. Only* $2.5\%$ *of the probability is above* $1.46$*.*

*There are still details missing but the idea is hopefully clear. The epidemiologists should have a pretty good idea of the distribution of asthma death rates in large cities comparable to the city under consideration. The death rate for that city could then be considered a random draw from the distribution of death rates. Knowledge of summary characteristics of the distribution such as mean, median, mode, quantiles, etc. allow selection of the particular pair of hyperparameters* $\alpha$ *and* $\beta$ *that come close to matching this distribution.*

*The posterior is then*

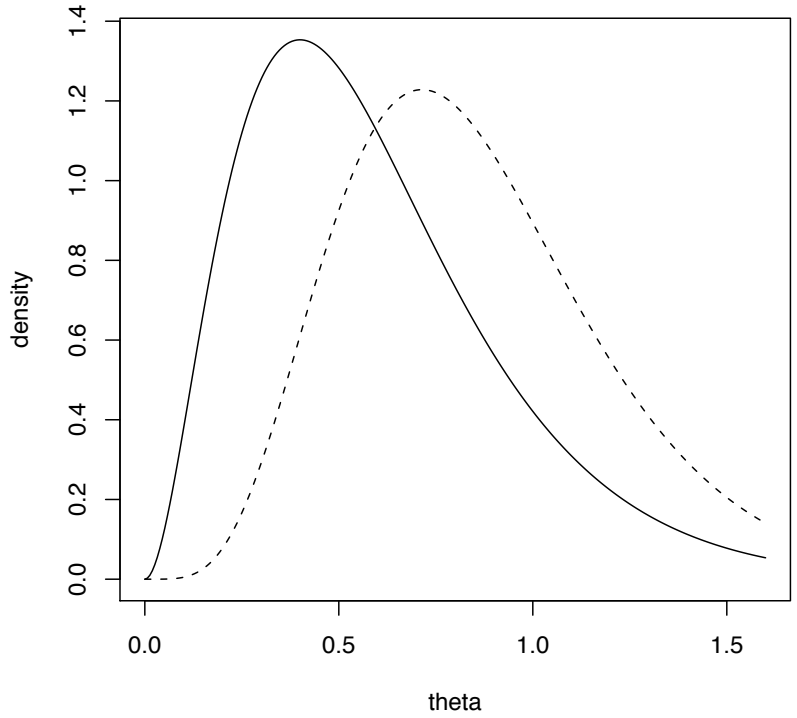
$$\begin{aligned} p(\theta \mid y = 3) &\propto \theta^{(\alpha-1)} \exp(-\beta\theta) \cdot \theta^3 \exp(-2\theta) \\ &\propto \theta^{(3-1+3)} \exp(-(5+2)\theta) \\ &\propto \theta^{(6-1)} \exp(-7\theta) \end{aligned}$$

*i.e. the posterior is*

$$\theta \mid y = 3 \sim Gamma(6, 7).$$

*The mean of the posterior is* $6/7 = 0.857$ *and variance* $6/49 = 0.122$*. A sketch of the prior distribution (solid line) and the posterior distribution (dashed line) is shown below.*

```
> theta=seq(0,1.6,l=250)
> plot(theta,dgamma(x,3,5),type="l",xlab="theta",ylab="density")
> lines(theta,dgamma(x,6,7),lty=2)
```



*Note that the informative prior in this example tends to dominate the data shrinking the observed value of* 1.5 *towards the prior mean of* 0.6*. The data are not entirely overwhelmed, however. The variance is virtually unchanged, and whereas only* 2.5% *of the probability in the prior lies above* 1.46 *the amount of probability above that value in the posterior is about* 0.06*.*

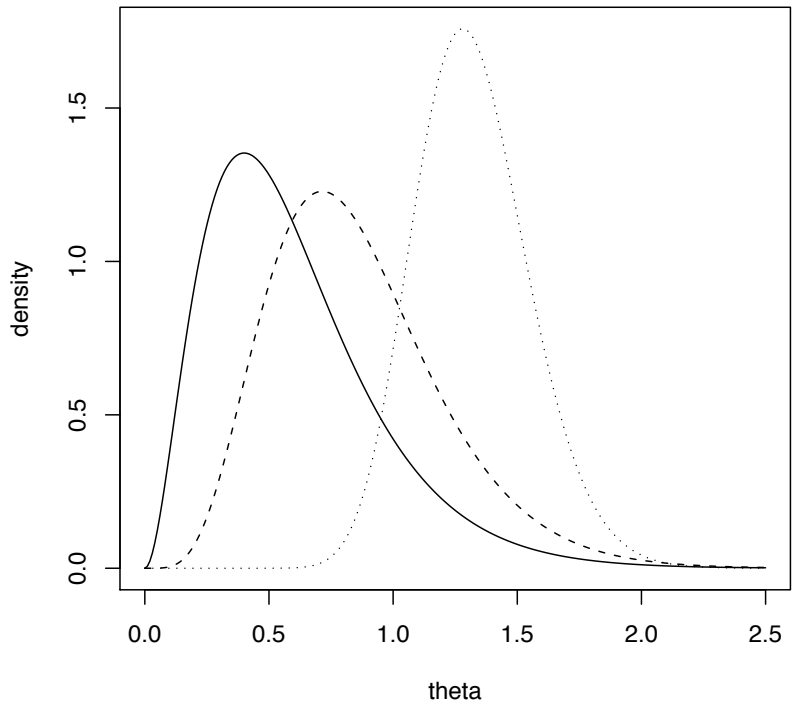
```
> 1-pgamma(1.456899,6,7)
[1] 0.05994606
```

*Suppose we have* 10 *years of data,* $y_1, \cdots, y_{10} \sim Poisson(2\theta)$*. We observe* $y = \sum y_i = 30$ *deaths. The posterior is*

$$\theta \mid y \sim Gamma\left(\alpha + \sum_{i=1}^{n} y_i, \beta + \sum_{i=1}^{n} x_i\right) = Gamma(3 + 30, 5 + 20) = Gamma(33, 25).$$

*The mean is* $33/25 = 1.32$ *and the variance is* $33/625 = 0.053$*.*

*I have sketched the prior (solid line), the posterior from one year's data (dashed line), and the posterior from 10 years data (dotted line) below.*




*Note the impact of the additional data; the data now dominate the prior.*

- *A Spatial Example: We will look at a simple example based on the lip cancer data set. As before we assume that* $Z_i$ *is the number of incident cases of lip cancer in county* $i$ *and* $E_i$ *is the expected number of cases in county* $i$*. The Standardized Morbidity Ratio (SMR) for the* $i$*th county is* $r_i = Z_i/E_i$*. There are spatially varying relative risks* $\gamma_i, i = 1, \cdots, 56$ *and conditional on these we have*

$$Z_i|\gamma_i \sim Poi\left(E_i\gamma_i\right).$$

*The MLE for $\gamma_i$ is $r_i$. In a Bayesian context we need a prior distribution on $\gamma_i$ and we will choose a Gamma conjugate prior, i.e.*

$$\gamma_i \sim Gamma(\alpha, \beta).$$

*Note the hierarchical structure.*

*Based on the results above the posterior distribution of $\gamma_i|z_i \sim Gamma\,(\alpha + z_i, \beta + E_i)$. Note that there is a posterior distribution for each of the 56 counties. The mean of the posterior distribution is one common way to summarize it and we have*

$$E\left[\gamma_i|z_i\right] = \frac{\alpha + z_i}{\beta + E_i}.$$

*It is more instructive to reexpress this as*

$$E\left[\gamma_i|z_i\right] = w_i SMR_i + (1 - w_i)\,\mu$$

*where $w_i = E_i/\left[E_i + (\mu/\sigma^2)\right]$, $\mu = \alpha/\beta$, and $\sigma^2 = \alpha/\beta^2$. Note that $0 < w_i < 1$ and that the posterior mean is a weighted average of the observed SMR values (the data) and the prior mean $\mu$. If $w_i$ is large then most of the weight goes to the observed SMR (the data dominate and/or the prior is uninformative) and if $w_i$ is small most of the weight goes to the prior mean (the data are sparse and/or the prior is informative). We will see small $w_i$ if either $E_i$ or $\sigma^2$ are small and $w_i$ will be large if either $E_i$ or $\sigma^2$ is large. Generally $\sigma^2$ (the prior variance) will be fixed so the $w_i$ values will depend on the county wide $E_i$ values in which case the prior will dominate when data are sparse and the data will dominate otherwise.*

*Carlin and Louis, in their Bayesian data analysis text, give an example of a Gamma prior that has been used for this type of problem. The Gamma distribution is parametrized so that a $Gamma(\alpha, \beta)$ distribution has mean $\mu = \alpha/\beta$ and variance $\sigma^2 = \alpha/\beta^2$. They choose $\mu = 1$ (on average there is no spatial variation in relative risk) and a variance of $\sigma^2 = (0.5)^2$ which they argue is large given the scale. This results in values of $\alpha = \beta = 4$. Thus the posterior mean is*

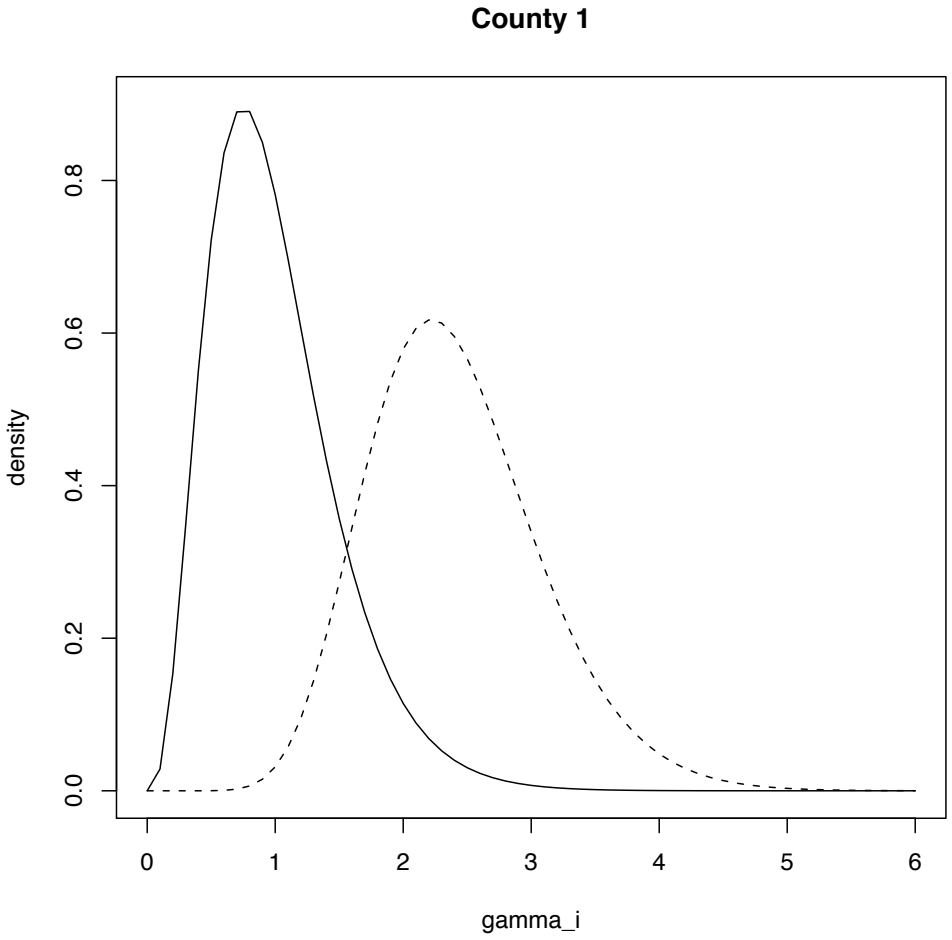
$$E\left[\gamma_i|z_i\right] = \frac{4 + z_i}{4 + E_i}.$$

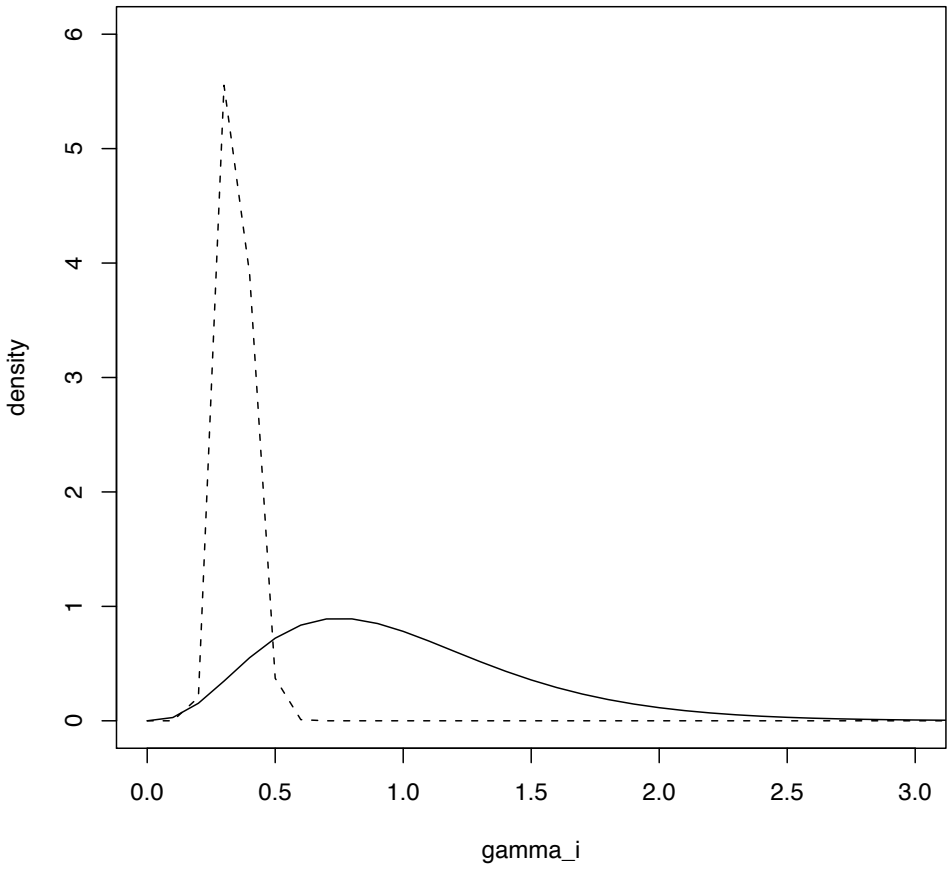
*The lip cancer data set has the observed SMR values and the observed and expected counts. It is thus easy to generate the 56 posterior means and compare them to the observed data. There are a number of ways to do such comparisons. We can look at the results for individual counties. For example, county 1 has an observed $SMR_1 = 6.43$ and $E_1 = 1.4$ which is fairly small. The observed count is $Z_1 = 9$. The mean of the posterior distribution is*

$$E\left[\gamma_1|Z_1 = 9\right] = \frac{4 + 9}{4 + 1.4} = 2.41.$$

*The impact of the small $E_1 = 1.4$ value and the prior is to shrink the raw estimate toward the prior mean of 1. By contrast, we have $E_{49} = 88.7$, $Z_{49} = 28$, $SMR_{49} = 0.32$, and the mean of the posterior is 0.35. The data dominate in this county.*

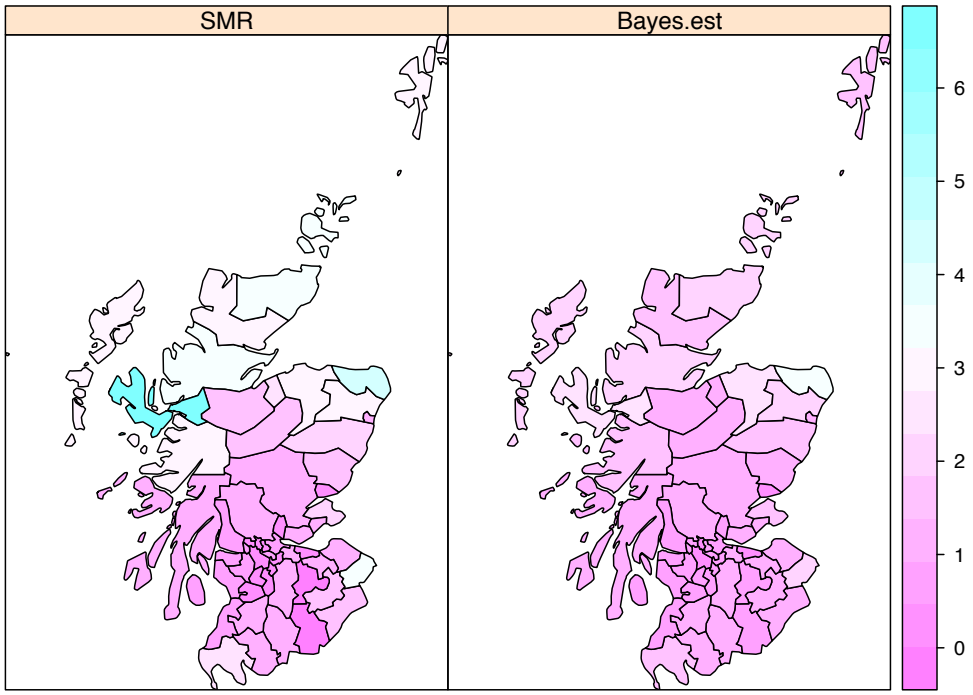
*We have a posterior distribution for each county. We can look at the county level posteriors and provide answers to questions that may be of interest. We can look at histograms of the prior and posterior distributions.*


County 1
density
0.0
0.2
0.4
0.6
0.8
0
1
2
3
4
5
6
gamma_i

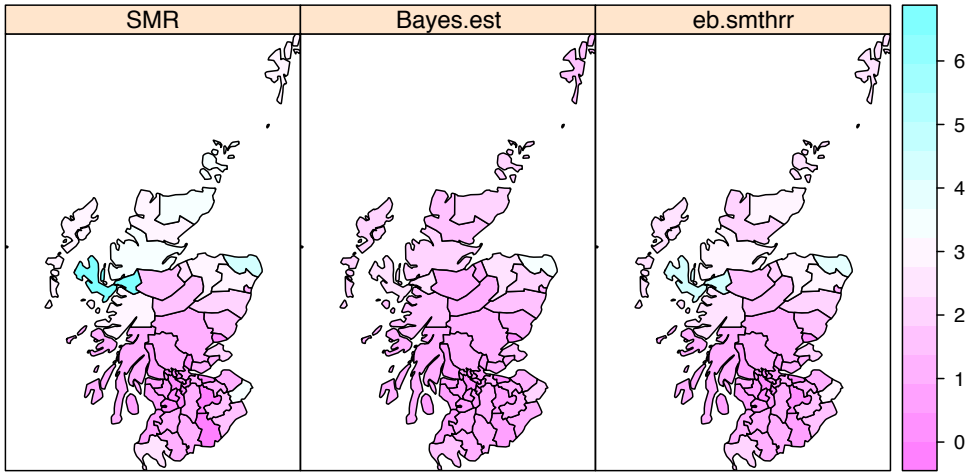



*The solid line is the prior and the dashed line is the posterior. Recall that the observed SMR in County 1 was* 6.43*. You can see that the prior dominates with the mean of the posterior of* 2.41 *being shrunk towards* 1*. The mean of the posterior for County* 49 *was* 0.35 *showing that the prior was dominated by the data. We can estimate* $P(\gamma_1 > 1|Z) = 0.996$ *whereas the probability for County* 49 *is* $P(\gamma_1 > 1|Z) \approx 0$*. Credible intervals (*95*%) are* $(1.28, 3.88)$ *and* $(0.24, 0.47)$ *for counties* 1 *and* 49*, respectively.*

*A map comparing the raw* $SMR$ *and smoothed Bayes estimates of the relative risk is shown below. You can see that the Bayes estimates are smoother than the raw counts. Notice for example, that the range of SMR values ranges from close to* 0 *to over* 6 *but the Bayes estimates are all in the range from* 0 *to around* 3*.*




*Bayesian methods are commonly used to produce smoothed disease maps in spatial epidemiology. Often so-called empirical Bayes methods are used for this purpose (see Chapter 4 in Waller and Gotway). In an empirical Bayes approach the data would be used to determine values for the parameters in the prior distribution as opposed to the pure Bayesian approach we took here. The most intuitive way to proceed would be to set the sample mean SMR value equal to $\alpha/\beta$ and the sample variance equal to $\alpha/\beta^2$ and solve for $\alpha$ and $\beta$. The resulting "method of moments" estimators would yield values of $\alpha = 1.354$ and $\beta = 0.889$. These would be the values we use for the prior distribution. R has a function* `empbaysmooth` *in the* `DCluster` *library that will do these calculations for us. It uses the crude method of moments estimators above as starting points in an iterative method to generate empirical Bayes estimates for the prior parameters that have better properties. The results are shown below.*

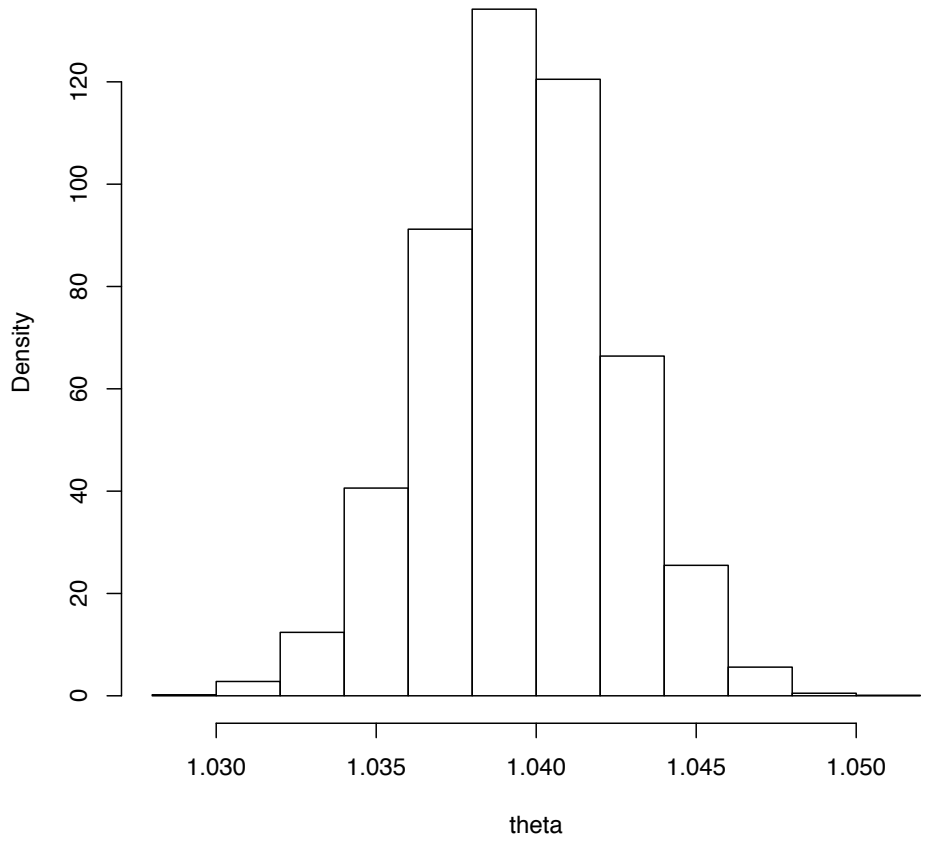



*You see less smoothing with the empirical Bayes approach which makes sense because we are letting the data play a major role in determining the prior distribution.*

- As mentioned earlier the posterior is the goal of all Bayesian analyses. It makes inference on parameters easy but it also makes inference on transformation of parameters easy. Consider the Laplace example. The posterior for $\theta$ was $Beta(251528, 241946)$. What if we were interested in the odds $\theta/(1-\theta)$? This is a change of variable problem in 421 or 501. Given $\theta \sim Beta(251528, 241946)$ find the distribution of $\theta/(1-\theta)$. The result would not be a recognizable distribution. You could find the mean and variance but it would involve some messy integration. It is easy in the Bayesian setting to work with transformations like this. I drew 5000 $\theta$'s from the posterior distribuion:

```
> theta<-rbeta(5000,251528,241946)
```

A histogram of the corresponding 5000 odds is shown below. This is a good approximation to the true posterior distribution of the odds.




We can easily get good approximations to the mean, variance, median, quantiles, etc.

```
> mean(theta/(1-theta))
[1] 1.039523
> var(theta/(1-theta))
[1] 8.51715e-06
> median(theta/(1-theta))
[1] 1.039522
> quantile((theta/(1-theta)),p=c(0.025,0.975))
    2.5%    97.5%
1.033694 1.045209
```

We see for example, that an approximate 95% credible interval for the odds is $(1.034, 1.045)$.

- Thus, one reason for the popularity of conjugate families is that it makes posterior inference much simpler. But things become more complex if we move away from conjugate priors. Further, conjugate priors are not suitable for multiparameter problems. We need methods for sampling from complex posterior distributions and that is the next topic.

*Markov Chain Monte Carlo*

- Imagine a person taking a random walk over the values $1, 2, 3, 4, 5, 6$. If he is at one of the endpoints he stays there with probability $1/2$ or moves to the adjacent location with probability $1/2$. If he is at one of the interior

points he remains where he is with probability 1/2, moves left with probability 1/4 or right with probability 1/4. This is a simple example of a discrete Markov chain. The process has 6 states and the chain describes the probabilities of movement among the states. The essence of a Markov chain is that where one moves next depends only on the current location or state and not on any previous locations.

More formally, we consider a family of random variables $\{x^n\}$ with $n \in \{0, 1, 2, \cdots, \}$. Such a family is called a discrete *time* stochastic process. The range of possible values of $x$ denotes the state space of the process and $x^n$ denotes the state at *time* $n$. The state space can be discrete (finite or infinite) or continuous. By way of a brief introduction we will focus on finite valued state space. If

$$Pr\left(x^{n+1} = j | x^n = i, x^{n-1} = i^{n-1}, \cdots, x^0 = i^0\right) = Pr\left(x^{n+1} = j | x^n = i\right)$$

for all time steps $n$ and all states $i^0, i^1, \cdots, i^{n-1}, i, j$, then the stochastic process is a Markov chain. The probability that $x^{n+1} = j$ given $x^n = i$ is the one-step transition probability $p_{ij}$. We assume for convenience that these do not depend on $n$. These probabilities are typically summarized in a transition probability matrix $P$.

In the simple random walk example we have

$$P = \begin{bmatrix} 0.50 & 0.50 & 0 & 0 & 0 & 0 \\ 0.25 & 0.50 & 0.25 & 0 & 0 & 0 \\ 0 & 0.25 & 0.50 & 0.25 & 0 & 0 \\ 0 & 0 & 0.25 & 0.50 & 0.25 & 0 \\ 0 & 0 & 0 & 0.25 & 0.50 & 0.25 \\ 0 & 0 & 0 & 0 & 0.50 & 0.50 \end{bmatrix}$$

The first row contains the probabilities of moving from state 1 to any of the states 1 through 6 in one step. The second row contains the probabilities of moving from state 2 to any of the states in a single step, and so on. Note that the rows sum to 1, i.e. each row must be a valid probability distribution. It is easy to find $n$ step transition probabilities by calculating the matrix

$$P^n = \underbrace{P \times P \times \cdots \times P}_{n \text{ times}}$$

Let $u^n = \left(p_1^n, \cdots, p_N^n\right)^T$ be a probability vector whose elements denote the probability of being in an indicated state at time $n$. Then $P^T u^n = u^{n+1}$.

Markov chains can be characterized by their properties. The most important property for us is regularity. A finite discrete state Markov chain is said to be regular if there exists an integer $k$ such that all the elements of $P^k$ are positive. Associated with all regular chains is a limiting distribution $\pi = \left(\pi_0, \pi_1, \cdots, \pi_N\right)^T$ where $\pi_j$ is the approximate long run probability of the chain being in state $j$:

$$\lim_{n \to \infty} Pr\left(x^n = j | x^0 = i\right) = \pi_j > 0.$$

Note that this is independent of the starting state. The limiting distribution is such that

$$P^T \pi = \pi.$$

The limiting distribution is stationary in the sense that if the starting states are determined with probability distribution $\pi$ then the probability of being in state $j$ at any future time is $\pi_j$.

The limiting distribution for the random walk example is

$$\pi = \left(0.1, 0.2, 0.2, 0.2, 0.2, 0.1\right)^T.$$

- The above can be generalized to Markov chains with continuous valued state spaces. The goal of MCMC methods is to set up a chain whose stationary distribution is equal to the posterior density from which we wish to sample. We start sampling at an initial point and at some future time step we should be sampling from the stationary distribution.

Specifically, we will be drawing values of a parameter vector $\theta$ from the Markov chain. These are drawn sequentially in such a way that the distribution of a particular sample depends only on the previous value. A sequence $\theta^t, t = 1, 2, \cdots$ initiates from a specified starting value $\theta^0$ and subsequent draws are made from the transition distribution $T_t\left(\theta^t \mid \theta^{t-1}\right)$. These transition probability distributions are constructed in such a way that the chain converges to its stationary distribution, $f(\theta \mid y)$. The key is to sample long enough so that the distribution from which draws are being made is approximately equal to the posterior density of interest. The devil is in the details.

- Gibbs Sampler: We have a $d$-dimensional parameter $\theta = (\theta_1, \theta_2, \cdots, \theta_d)^T$. At each iteration $\theta_j^t$ is sampled, conditional on all the other components:

$$f\left(\theta_j |\ \theta_1^t, \theta_2^t, \cdots, \theta_{j-1}^t, \theta_{j+1}^{t-1}, \cdots, \theta_d^{t-1}, y\right).$$

Each $\theta_j$ is updated conditional on the latest values of the other components which will the value at the $t$th iteration for those components updated ahead of $\theta_j$ and the value at the $(t-1)$st iteration for components yet to be updated. Gibbs sampling is convenient to use when it is possible to draw directly from conditional distributions. Here is a simple example from the text.

*Bivariate normal distribution: We suppose that we have a posterior distribution*

$$\left.\begin{pmatrix} \theta_1 \\ \theta_2 \end{pmatrix}\right| y \sim N\left(\begin{pmatrix} y_1 \\ y_2 \end{pmatrix}, \begin{pmatrix} 1 & 0.8 \\ 0.8 & 1 \end{pmatrix}\right)$$

*distribution. From the properties of the multivariate normal distribution we know*

$$\begin{aligned} \theta_1|\theta_2, y &\sim N\left(y_1 + 0.8\left(\theta_2 - y_2\right), 0.36\right) \\ \theta_2|\theta_1, y &\sim N\left(y_2 + 0.8\left(\theta_1 - y_1\right), 0.36\right) \end{aligned}$$

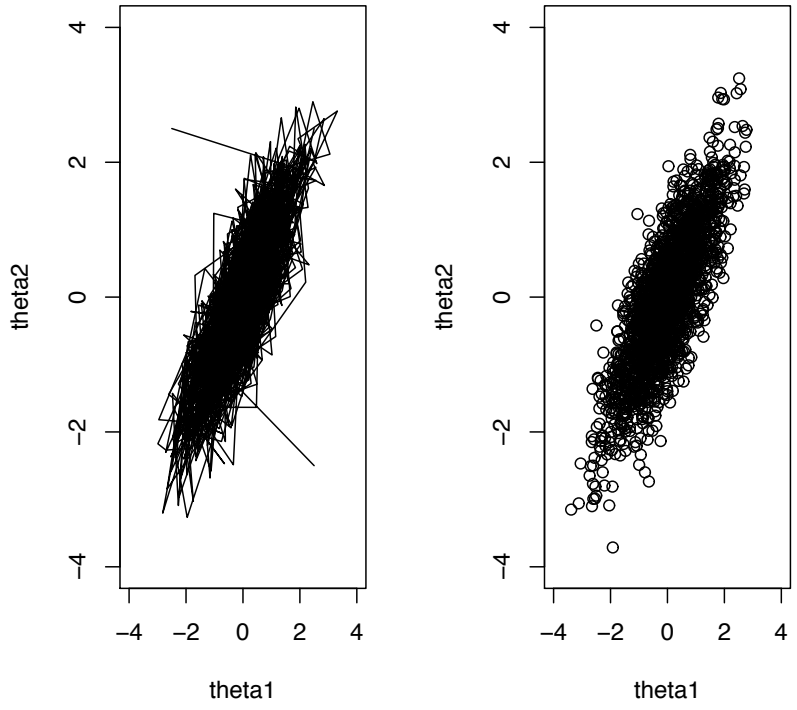
*We have one observation $y_1 = y_2 = 0$. We proceed iteratively by drawing first from $\theta_1|\theta_2, y$ followed by $\theta_2|\theta_1, y$. One of the major problems in MCMC simulations is to confirm convergence of the chain. One suggested method is to run several independent sequences using different starting points. If, after a suitable 'burn-in' period, all the chains are producing simulated values in the same area of the parameter space then this is taken as one sign of convergence. For this example, we will evaluate 4 sequences started from $(\pm 2.5, \pm 2.5)$.*

```
> gibbs.fun=function(th.1,th.2,th.10,th.20,n){
+ th.1[1]=th.10
+ th.2[1]=th.20
+ for(i in 2:n){
+ th.1[i]=rnorm(1,0.8*(th.2[i-1]),sqrt(.36))
+ th.2[i]=rnorm(1,0.8*th.1[i],sqrt(.36))}
+ return(list(th.1=th.1,th.2=th.2))}
```

*The vectors* `th.1` *and* `th.2` *contain* 1000 *simulated values of $\theta_1$ and $\theta_2$ each and* `th.10` *and* `th.20` *are the starting values. We know that early simulated values of $\theta_1$ and $\theta_2$ will not be drawn from the posterior distribution but at some point the chain should converge. The first panel in the figure below shows the results after* 500 *iterations for all 4 sequences. They have reached apparent convergence. The second panel shows the* 2000 *pairs of simulated $(\theta_1, \theta_2)$ values. These are to be considered samples from the posterior distribution of interest. The R commands below generated the simulated values and the plots:*

```
> par(mfrow=c(2,1))
> th.11=gibbs.fun(th.1,th.2,2.5,2.5,1000)
> th.12=gibbs.fun(th.1,th.2,2.5,-2.5,1000)
> th.22=gibbs.fun(th.1,th.2,-2.5,-2.5,1000)
> th.21=gibbs.fun(th.1,th.2,-2.5,2.5,1000)
> plot(c(-4,4),c(-4,4),type='n',xlab="theta1",ylab="theta2")
> lines(th.21$th.1[1:500],th.21$th.2[1:500])
> lines(th.22$th.1[1:500],th.22$th.2[1:500])
```

```
> lines(th.12$th.1[1:500],th.12$th.2[1:500])
> lines(th.11$th.1[1:500],th.11$th.2[1:500])
> plot(c(-4,4),c(-4,4),type='n',xlab="theta1",ylab="theta2")
> points(th.11$th.1[501:1000],th.11$th.2[501:1000])
> points(th.21$th.1[501:1000],th.21$th.2[501:1000])
> points(th.22$th.1[501:1000],th.22$th.2[501:1000])
> points(th.12$th.1[501:1000],th.12$th.2[501:1000])
```




*I computed the means, variances, and the correlation of the* 2000 *values left over after discarding the burn-in values. The means are* $-0.036, -0.049$*, the variances are* $0.976, 0.982$ *and the correlation is* $0.798$ *which can be compared to the true means of* $0$*, variances of* $1$ *and correlation of* $0.8$*. We really did not run this chain all that long and running it longer with perhaps a longer burn-in period should yield even better results.*

- Metropolis-Hastings Algorithms: These are methods of sampling from a posterior $f(\theta \mid y)$ using a proposal density (referred to as a jumping density). The choice of the type of jumping density is what distinguishes among the methods. In general, we proceed as follows:

  1. Draw a starting value $\theta^0$ from a starting distribution or more simply just specify a value. (In practice several values should be specified and independent sequences are generated. Evaluating the behavior of multiple sequences starting from different locations helps in assessing the convergence properties of the chain.)
  2. Simulate a candidate value $\theta^*$ from the jumping distribution $J_t\left(\theta^* \mid \theta^{t-1}\right)$.
  3. Compute the ratio

$$r = \frac{f\left(\theta^* \mid y\right) J_t\left(\theta^{t-1} \mid \theta^*\right)}{f\left(\theta^{t-1} \mid y\right) J_t\left(\theta^* \mid \theta^{t-1}\right)}.$$

4. Determine whether or not to accept the proposed value $\theta^*$ according to the rule:

$$\text{Set } \theta^t = \begin{cases} \theta^* & \text{with probability } \min(r, 1) \\ \theta^{t-1} & \text{otherwise} \end{cases}$$

   If the acceptance probability is $r < 1$ then this step requires drawing a $Unif(0, 1)$ random variable and comparing it to $r$.

Under specified and not unreasonable regularity conditions involving the properties of the Markov chain and $J_t$, the sequence $\theta^t, t = 1, 2, 3, \cdots$ will converge to a sequence of random variables distributed according to the posterior density.

Different choices of $J_t$ lead to different algorithms. If $J_t$ is chosen to be symmetric, i.e. $J_t\left(\theta^* \mid \theta^{t-1}\right) = J_t\left(\theta^{t-1} \mid \theta^*\right)$ then

$$r = \frac{f\left(\theta^* \mid y\right)}{f\left(\theta^{t-1} \mid y\right)}.$$

This is sometimes referred to in the literature as a Metropolis-Hastings algorithm with a random walk chain. Some texts reserve the term Metropolis-Hastings for nonsymmetric $J_t$'s. Even here there are choices to be made. For example, choosing $J_t\left(\theta^* \mid \theta^{t-1}\right) = J_t\left(\theta^*\right)$ results in a Metropolis-Hastings algorithm with an independence chain. One choice of a jumping distribution actually leads to Gibbs sampling. The Gibbs sampler and the simpler Metropolis algorithm can serve as building blocks for more complex chains.

The following are properties of a good jumping distribution:

1. it is easy to sample from
2. $r$ is easy to compute
3. each jump goes a reasonable distance (but not too far) in the parameter space
4. jumps are not rejected too frequently

Proof of convergence of the Markov chains to a stationary density follows immediately from the fact that, under relatively mild regularity conditions, the Markov chains involved have the properties required for convergence from the theory of stochastic processes. Proof that the stationary distribution is in fact the posterior of interest is a bit more involved.

- MCMC output analysis: The theoretical results provide no guidance on how to decide if a given chain has converged. Inference can also be made more difficult by properties of the chain that we have not yet discussed.
  - Burn-in period: The sequence of $\theta^t$ values converges to the posterior density but the rate at which it converges varies. Certainly the initial simulated values have not been drawn from the target distribution. How many of these initial values should be discarded? A conservative approach is to discard the first half of the sequence. Even then trace plots of simulated values against iteration number should be examined to provide more confidence that the chain has in fact converged.
  - Dependencies in the chain: We are looking at a Markov chain and pairs $\left(\theta^{t-1}, \theta^t\right)$ are dependent. If the dependence is strong then the chain may converge slowly and in fact may appear to have converged before all of the parameter space has been adequately 'explored'. The degree of autocorrelation can be assessed by plotting the autocorrelation function. A common recommendation is that chains be thinned by accepting values in a systematic manner, i.e. accept only every $k$th value. If convergence has occurred the chain can still be used for inference as even in the presence of autocorrelation.
  - It is always a good idea, as we have already mentioned above, to simulate independent sequences of the chain with different starting points. Choosing starting values that are known to not be close to the truth is actually a good idea (within reason). Such choices may result in chains that converge more slowly but if they all converge to the same place then one has more confidence in the final results. We also have less chance that areas of the parameter space will be missed.

– Convergence can be monitored by visual inspection of trace plots but quantitative methods are also available. One such method is based on comparisons of within and between sequence variability. That is, we generate $m$ independent sequences of the chain, each of length $n$. Let $\psi_{ij}$ be the $i$th iterate from the $j$th sequence. Then the between sequence variance is

$$B = \frac{n}{m-1}\sum_{j=1}^{m}\left(\overline{\psi}_{.j} - \overline{\psi}_{..}\right)^2$$

where

$$\overline{\psi}_{.j} = \frac{1}{n}\sum_{i=1}^{n}\psi_{ij}, \quad \overline{\psi}_{..} = \frac{1}{m}\sum_{j=1}^{m}\overline{\psi}_{.j}.$$

The within sequence variance is

$$W = \frac{1}{m}\sum_{j=1}^{m}s_j^2$$

where

$$s_j^2 = \frac{1}{n-1}\sum_{i=1}^{n}\left(\psi_{ij} - \overline{\psi}_{.j}\right)^2.$$

The marginal posterior variance of $\psi$ is

$$\widehat{var}^{+}\left(\psi \mid y\right) = \frac{n-1}{n}\,W + \frac{1}{n}\,B.$$

Convergence is monitored by computing

$$\widehat{R}(n) = \sqrt{\frac{\widehat{var}^{+}\left(\psi \mid y\right)}{W}}.$$

$W$ underestimates the marginal posterior variance for finite $n$ (because the $m$ sequences have not had time to fully explore the posterior density) and $\widehat{var}^{+}$ overestimates the marginal posterior variance if we have chosen widely dispersed starting values. Both estimates converge to the marginal variance as $n \to \infty$ and so $\widehat{R}(n) > 1$ and

$$\lim_{n\to\infty}\widehat{R}(n) \to 1.$$

$\widehat{R}$ is referred to as a measure of scale reduction. It can be computed at various values of $n$ and if the potential reduction is high then further simulations are indicated whereas if $\widehat{R}(n) \approx 1$ further simulations are (probably) not needed. In the multiple parameter settings of most practical problems these assessments need to be carried out for all parameters and simulation runs are continued until *all* parameters show convergence.

– Simulation accuracy: The dependencies among the iterates means that the effective sample size is actually less than the observed sample size. The effective sample size can be computed as

$$n_{eff} = mn\left[\frac{\widehat{var}^{+}\left(\psi \mid y\right)}{B}\right].$$

We could also evaluate simulation accuracy using the autocorrelation structure of the sequence.

- *Here is a brief example. The target is a bivariate normal distribution with mean vector* $(0,0)$ *and* $2\times 2$ *variance-covariance matrix* $I$*, i.e. we are looking at 2 independent normally distributed random variables. The jumping distribution from which candidates* $\theta^*$ *will be drawn is also normal (and hence symmetric). We can generate the MCMC results easily here given the simplicity of the target.*

  1. *Specify a starting value - the value of* $\theta$ *at time* $t = 0$*. Ideally we would choose several starting values but we will only look at a single chain.*

2. *For* $t = 1, 2, \cdots, n$ *draw a candidate value* $\theta^*$ *from a bivariate normal distribution* $N\left(\theta^{t-1}, (0.2)^2 I\right)$. *Choosing a variance of* $0.2$ *means that the chain will not move very fast.*

3. *Compute the density ratio*

$$r = \frac{N\left(\theta^* | 0, I\right)}{N\left(\theta^{t-1} | 0, I\right)}.$$

*The ratio does not involve the jumping distribution because it is symmetric and cancels out of the numerator and denominator.*

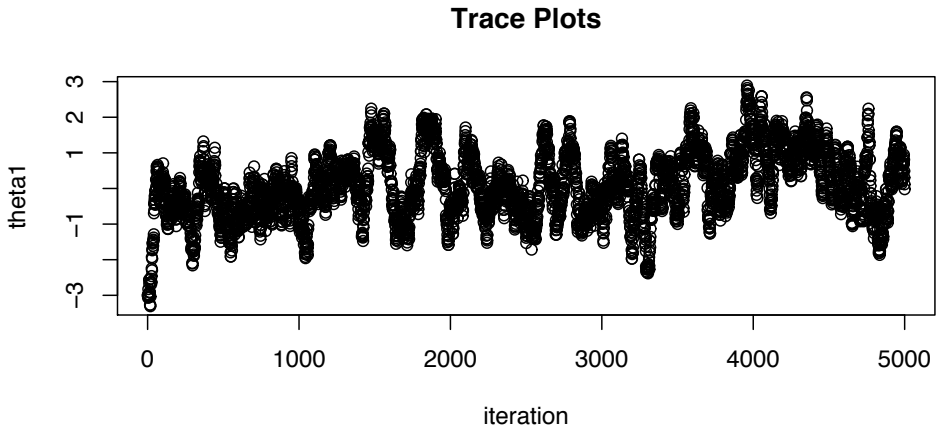
4. *Generate a* $Unif(0, 1)$ *random variable* $U$ *and set*

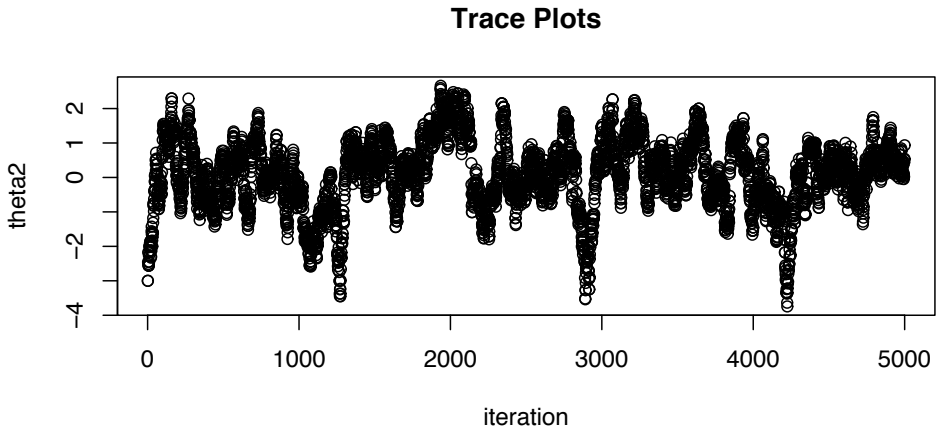
$$\theta^t = \begin{cases} \theta^* & U \leq \min(r, 1) \\ \theta^{t-1} & otherwise \end{cases}$$

5. *Iterate the process until the desired number of iterations is achieved.*

```
theta<-matrix(0,nrow=5000,ncol=2)
# specify a starting value
theta[1,]<-c(-3,-3)
#
for(i in 2:5000){
    theta.star<-c(rnorm(1,theta[(i-1),1],.2),rnorm(1,theta[(i-1),2],.2))
    r<-(dnorm(theta.star[1])*dnorm(theta.star[2]))/
                        (dnorm(theta[(i-1),1])*dnorm(theta[(i-1),2]))
    U<-runif(1)
    theta[i,1]<-ifelse(U<=min(r[i],1),theta.star[1],theta[(i-1),1])
    theta[i,2]<-ifelse(U<=min(r[i],1),theta.star[2],theta[(i-1),2])}
```

*Trace plots are shown below.*




*Convergence was slow and may not have been achieved even after* 5000 *iterations. I threw away the first* 3000 *iterations. The* 2.5*th and* 97.5*th percentiles and the correlation matrix are shown below. These do not look all that good.*

```
> quantile(theta[-(1:3000),2],p=c(0.025,0.975))
     2.5%     97.5%
-1.742841  1.676432
> quantile(theta[-(1:3000),1],p=c(0.025,0.975))
     2.5%     97.5%
-1.548241  2.026814
>  cor(theta[-(1:3000),])
           [,1]       [,2]
[1,]  1.0000000 -0.1196260
[2,] -0.1196260  1.0000000
```

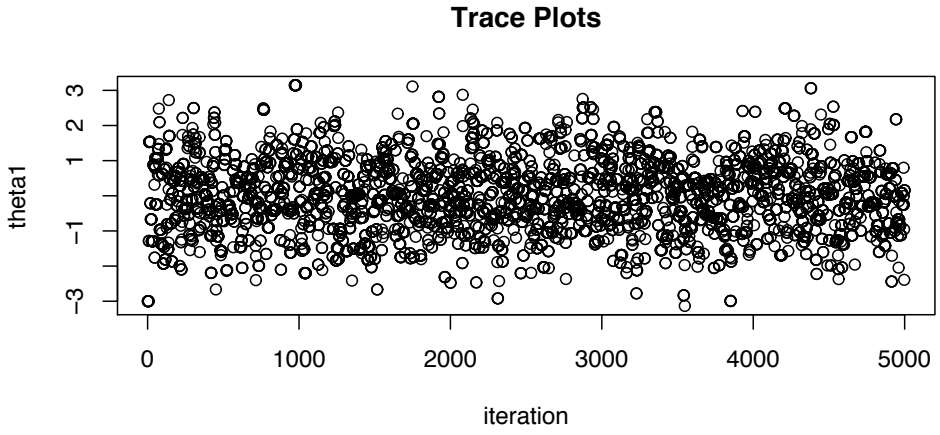
*I modified the jumping distribution by increasing the variance to* 4*.*

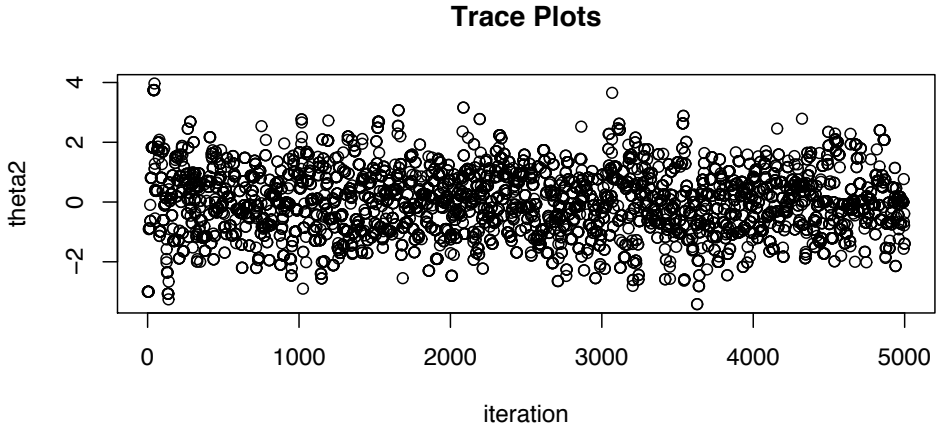
```
# specify a starting value
theta[1,]<-c(-3,-3)
#
for(i in 2:5000){
```

```
    theta.star<-c(rnorm(1,theta[(i-1),1],2),rnorm(1,theta[(i-1),2],2))
    r[i]<-(dnorm(theta.star[1])*dnorm(theta.star[2]))/
                    (dnorm(theta[(i-1),1])*dnorm(theta[(i-1),2]))
    U<-runif(1)
    theta[i,1]<-ifelse(U<=min(r[i],1),theta.star[1],theta[(i-1),1])
    theta[i,2]<-ifelse(U<=min(r[i],1),theta.star[2],theta[(i-1),2])}
```

*Trace plots look better. The resulting credible intervals and the correlation matrix for the latter* 2000 *iterations are also better.*

```
>   cor(theta[-(1:2999),])
          [,1]      [,2]
[1,] 1.0000000 0.0354492
[2,] 0.0354492 1.0000000
>  quantile(theta[-(1:2999),1],p=c(0.025,0.975))
     2.5%     97.5%
-1.874000  1.898624
>  quantile(theta[-(1:2999),2],p=c(0.025,0.975))
     2.5%     97.5%
-2.000128  2.005781
```




*There is a lot more to diagnostics than this but we will not go into those here.*

- *Lip cancer example: We looked at a Poisson-Gamma example above. We will now look at* 2 *models with both incorporating the covariate and one incorporating spatial coorelation. Incorporating correlation necessitates a different type of model - a Poisson-Lognormal model. Waller and Gotway discuss this model starting on page 412.*

  *We assume*

$$Z_i|\gamma_i \sim Poi\left(E_i\gamma_i\right)$$

  *with*

$$\log\left(\gamma_i\right) = \beta_0 + \beta_1 X_1 + \psi_i.$$

  *The model is thus*

$$Z_i|\beta_0, \beta_1, \psi_i \sim Poi\left(E_i \exp\left\{\beta_0 + \beta_1 X_1 + \psi_i\right\}\right).$$

  *As Bayesians we need to specify prior distributions for* $\boldsymbol{\beta} = \left[\beta_0, \beta_1\right]'$ *and* $\boldsymbol{\psi} = \left[\psi_1, \psi_2, \cdots, \psi_{56}\right]'$*. The joint posterior is*

$$f\left(\boldsymbol{\psi}, \boldsymbol{\theta}_\psi, \boldsymbol{\beta}|\mathbf{Z}\right) = f\left(\mathbf{Z}|\boldsymbol{\beta}, \boldsymbol{\psi}\right) f\left(\boldsymbol{\psi}|\boldsymbol{\theta}_\psi\right) f\left(\boldsymbol{\beta}\right) f\left(\boldsymbol{\theta}_\psi\right)$$

  *where* $\boldsymbol{\theta}_\psi$ *is a vector of parameters that define the spatial correlation structure for the "random effects". This is equation 9.70 on page 411 in Waller and Gotway. The model has a hierarchical structure, i.e. we assume that the prior distribution of* $\boldsymbol{\psi}$ *is dependent on* $\boldsymbol{\theta}_\psi$ *and assign a (hyper)prior distribution to* $\boldsymbol{\theta}_\psi$ *itself. We also assume a prior for* $\boldsymbol{\beta}$ *and that* $\boldsymbol{\theta}_\psi$ *and* $\boldsymbol{\beta}$ *are independent of one another. In all the models below we will assume that* $\beta_0$ *and* $\beta_1$ *are have independent normal distributions with mean* 0 *and variance* 20000*. These are noninformative priors.*

  *Model 1: We assume that* $\psi_i \sim N\left(0, \nu_\psi\right)$*. This induces a nonspatial overdispersion among the observed counts, the* $Z_i$*'s. We need to specify a hyperprior for* $\nu_\psi$*. Waller and Gotway illustated the Bayesian method on the leukemia data and chose a so-called inverse gamma distribution*

$$\tau_\psi = \frac{1}{\nu_\psi} \sim Gamma\left(0.5, 0.0005\right).$$

  *As they note on page 424, this prior were chosen for illustrative purposes and no claim is made as to their optimality. I am going to choose a prior used by Carlin and Louis in their discussion of Bayesain modeling of the lip cancer data:*

$$\tau_\psi \sim Gamma\left(0.001, 0.001\right).$$

  *Both of these are uninformative in that they have large variances.*

  *We will use MCMC methods to estimate the joint posterior distribution for the* 59 *parameters in this model* $\psi_1, \cdots, \psi_{56}, \beta_0, \beta_1, \tau_\psi$*. Bugs and the use of Bugs in R is described in Appendix C in Gelman et al. The basic idea is to specify a full Bayesian model including all prior information. We will need to provide this model to Bugs, along with data, the parameters whose posterior distributions we want, and some initial values at which to start a chain. It is possible to run multiple chains with a single execution of an R function. The appendix in Gelman et al. will be handed out to you in class.*

  *First we need Bugs code for the hierarchical model. The code is shown below. The file is saveds as a* `.bug` *file in the R working directory.*

```
model
{
   for (i in 1 : regions) {
      O[i] ~ dpois(mu[i])
      log(mu[i]) <- log(E[i]) + beta0 + beta1*aff[i]/10 + psi[i]
      psi[i] ~ dnorm(0.0,nu.psi)
      SMRhat[i] <- 100 * mu[i] / E[i]
      SMRraw[i] <- 100* O[i] / E[i]
   }
   beta0 ~ dnorm(0.0, 1.0E-5)  # vague prior on grand intercept
   beta1 ~ dnorm(0.0, 1.0E-5)  # vague prior on covariate effect
```

```
   nu.psi ~ dgamma(1.0E-3,1.0E-3)  #
}
```

*This looks a lot like R code. The* `dnorm` *command specifies the mean and precision of the prior distribution. The precision is the inverse of the variance. All priors in Bugs must be proper distributions. The priors specified here are proper but are "diffuse" or "vague".*

*We need to tell Bugs how to distinguish between data and parameters, and we need to tell it how to choose initial or starting values.*

```
> # set up the data in R
> O<-lipcancer2.dat2$Observed
> E<-lipcancer2.dat2$Expected
> aff<-lipcancer2.dat$PCAFF # this is will be transformed to X1
> regions<-56
> # set up commands to identify and pass  data, parameters, and
> # initial values to Bugs
> lipcancer.data<-list("O","E","regions","aff")
> lipcancer.parameters<-c("beta0","beta1","psi","'tau.psi"')
> lipcancer.inits<-function(){
+ list(beta0=rnorm(1),beta1=rnorm(1),theta=rnorm(regions),psi=rnorm(regions),
+ tau.psi=rnorm(1,1,.2)}
> # load the required libraries
> library(arm)
> library(BRugs)
Welcome to BRugs running on OpenBUGS version 3.0.3
> lipcancer.model1<-bugs(lipcancer.data,lipcancer.inits,lipcancer.parameters,
+ "lipcancer.model1.bug",n.chains=5,n.iter=10000)
```

*When you execute the* `bugs` *functions the R command window is frozen and Bugs window opens. In the first attempt or so it pays to focus on the log window. It will tell you how things are going and it will provide debugging information when problems crop up. Initially it will also help to specify only a few iterations and chains along with the argument* `debug=T` *which keeps the log window open after Bugs is done. Once you have the function working you can delete the* `debug` *argument.*

*It did not take long for the above model to fit. There is a lot of ouput one could look at. It is always worth looking at basic summary information.*

```
> lipcancer.model1
Inference for Bugs model at "lipcancer.model1.bug", fit using WinBUGS,
 5 chains, each with 10000 iterations (first 5000 discarded), n.thin = 25
 n.sims = 1000 iterations saved
          mean   sd  2.5%   25%   50%   75% 97.5% Rhat n.eff
beta0     -0.5  0.2  -0.9  -0.6  -0.5  -0.4  -0.2    1   650
beta1      0.7  0.2   0.4   0.6   0.7   0.8   1.0    1   900
psi[1]     0.9  0.4   0.1   0.6   0.9   1.1   1.6    1   800
psi[2]     0.8  0.2   0.4   0.7   0.8   1.0   1.2    1  1000
psi[3]     0.8  0.3   0.1   0.6   0.9   1.1   1.4    1   680
psi[4]     0.1  0.4  -0.7  -0.2   0.1   0.3   0.7    1  1000
psi[5]     0.9  0.3   0.3   0.7   0.8   1.0   1.4    1   970
psi[6]     0.0  0.4  -0.7  -0.2   0.0   0.3   0.7    1  1000
psi[7]     0.9  0.2   0.4   0.7   0.9   1.0   1.3    1  1000
psi[8]     0.7  0.4   0.0   0.5   0.7   1.0   1.5    1  1000
psi[9]     0.7  0.4  -0.1   0.4   0.7   1.0   1.4    1  1000
psi[10]    0.4  0.3  -0.1   0.3   0.4   0.6   0.9    1  1000
```

```
psi[11]    0.9  0.3   0.3   0.7   0.9   1.1   1.4    1   900
psi[12]    0.2  0.4  -0.6   0.0   0.2   0.5   1.0    1   600
psi[13]    0.3  0.5  -0.7   0.0   0.3   0.7   1.2    1  1000
psi[14]   -0.2  0.4  -1.0  -0.5  -0.2   0.0   0.4    1  1000
psi[15]    0.7  0.3   0.2   0.5   0.7   0.8   1.1    1  1000
psi[16]    0.0  0.3  -0.6  -0.2   0.0   0.2   0.6    1  1000
psi[17]    0.1  0.5  -0.8  -0.2   0.2   0.5   1.1    1   520
psi[18]    0.3  0.3  -0.4   0.1   0.4   0.6   1.0    1   780
psi[19]    0.4  0.3  -0.3   0.2   0.4   0.6   0.9    1  1000
psi[20]    0.2  0.3  -0.5   0.0   0.2   0.4   0.8    1   260
psi[21]    0.3  0.3  -0.2   0.2   0.3   0.5   0.8    1   810
psi[22]   -0.3  0.2  -0.7  -0.4  -0.3  -0.1   0.1    1   620
psi[23]    0.0  0.3  -0.6  -0.2   0.0   0.2   0.5    1  1000
psi[24]    0.2  0.3  -0.5   0.0   0.2   0.4   0.8    1  1000
psi[25]    0.5  0.3   0.0   0.4   0.5   0.7   1.0    1   750
psi[26]    0.5  0.3  -0.1   0.3   0.5   0.7   1.1    1  1000
psi[27]    0.1  0.3  -0.5  -0.1   0.1   0.3   0.7    1  1000
psi[28]    0.1  0.3  -0.6  -0.1   0.1   0.3   0.6    1  1000
psi[29]   -0.1  0.3  -0.6  -0.3  -0.1   0.1   0.4    1  1000
psi[30]   -0.1  0.3  -0.7  -0.3  -0.1   0.1   0.4    1  1000
psi[31]    0.0  0.4  -0.8  -0.3   0.0   0.3   0.7    1   610
psi[32]   -0.7  0.4  -1.6  -1.0  -0.7  -0.4   0.0    1   520
psi[33]   -0.2  0.3  -0.9  -0.4  -0.2   0.0   0.4    1   780
psi[34]   -0.1  0.3  -0.7  -0.3  -0.1   0.2   0.5    1  1000
psi[35]   -0.1  0.3  -0.6  -0.3  -0.1   0.1   0.4    1  1000
psi[36]    0.3  0.3  -0.4   0.1   0.3   0.5   0.9    1  1000
psi[37]   -0.3  0.3  -0.8  -0.5  -0.3  -0.1   0.2    1   560
psi[38]    0.2  0.3  -0.5   0.0   0.2   0.4   0.8    1  1000
psi[39]   -0.6  0.3  -1.3  -0.8  -0.6  -0.4   0.0    1  1000
psi[40]    0.1  0.4  -0.8  -0.2   0.1   0.3   0.8    1  1000
psi[41]   -0.2  0.3  -0.8  -0.4  -0.2   0.0   0.4    1   840
psi[42]   -1.0  0.3  -1.6  -1.2  -1.0  -0.8  -0.5    1   430
psi[43]   -0.7  0.4  -1.6  -1.0  -0.7  -0.4   0.0    1   790
psi[44]   -0.3  0.3  -1.0  -0.5  -0.3  -0.1   0.3    1  1000
psi[45]   -0.5  0.3  -1.0  -0.7  -0.5  -0.3   0.0    1   890
psi[46]   -0.6  0.4  -1.4  -0.8  -0.6  -0.4   0.1    1  1000
psi[47]   -0.3  0.4  -1.2  -0.6  -0.3   0.0   0.5    1  1000
psi[48]   -0.4  0.4  -1.2  -0.7  -0.4  -0.1   0.3    1   870
psi[49]   -0.6  0.2  -1.1  -0.8  -0.6  -0.4  -0.2    1  1000
psi[50]   -0.6  0.3  -1.2  -0.8  -0.5  -0.3   0.1    1  1000
psi[51]   -0.3  0.5  -1.4  -0.6  -0.3   0.0   0.7    1  1000
psi[52]   -0.3  0.5  -1.5  -0.6  -0.3   0.0   0.6    1  1000
psi[53]   -0.5  0.4  -1.5  -0.8  -0.5  -0.2   0.3    1   430
psi[54]   -0.6  0.4  -1.5  -0.9  -0.6  -0.3   0.2    1   970
psi[55]   -1.1  0.5  -2.1  -1.4  -1.1  -0.8  -0.2    1   950
psi[56]   -0.5  0.5  -1.6  -0.9  -0.5  -0.2   0.3    1   830
tau.psi    2.8  0.8   1.5   2.2   2.7   3.2   4.7    1   370
deviance 270.2 11.2 250.3 262.5 269.6 276.8 294.5    1   400

For each parameter, n.eff is a crude measure of effective sample size,
and Rhat is the potential scale reduction factor (at convergence, Rhat=1).

DIC info (using the rule, pD = Dbar-Dhat)
pD = 39.7 and DIC = 309.9
```

`DIC is an estimate of expected predictive error (lower deviance is better)`

*We only used* 1000 *of the* 50000 *simulated values. The first half of each of the 5 chains was discarded and then the default thinning (*`n.thin=25`*) indicates that only every* 25*th value was used. A little care is needed here because R is (for some strange reason) rounding everything to one decimal place. What you see is a summary of the joint posterior distribution. Each line of the output shows the mean, standard deviation, and* 5 *quantiles. You can easily read off* 95% *credible intervals. The* `Rhat` *values are all rounded to* 1 *and the effective sample sizes are all reasonably high suggesting that the chain converged. We see a lot of county to county variability. We can see the unrounded output by entering the command* `lipcancer.model1$summary`*. The* `Rhat` *values are all very close to* 1 *so not too much rounding was done. Summary information for the regression coefficients is shown below.*

| *Coefficient* | *Posterior Mean* | *Posterior SD* | 95% *Credible interval* | $\widehat{R}$ | `n.eff` |
|---|---|---|---|---|---|
| $\beta_0$ | $-0.511$ | 0.168 | $(-0.861, -0.192)$ | 1.003 | 650 |
| $\beta_1$ | 0.695 | 0.152 | $(0.388, 0.999)$ | 1.002 | 900 |

*This is the Bayesian analog to the simple mixed model we fit to this data set earlier (see pages 69 and 70). The estimated coefficients and standard errors were* $\widehat{\beta}_0 = -0.441$ $(0.157)$ *and* $\widehat{\beta}_1 = 0.680$ $(0.141)$.

*There is obviously a lot of county to county variation as evidenced by the variation in the* $\psi_i$ *means. You can think of these as being "random" additions to the grand (fixed effect) intercept.*

*The results below show the expected counts, the raw SMR values and the modeled SMR values.*

```
psi[1]   1.4 6.4285714 4.3679669
psi[2]   8.7 4.4827586 4.1875230
psi[3]   3.0 3.6666667 2.8039895
psi[4]   2.5 3.6000000 3.3702744
psi[5]   4.3 3.4883721 2.8229045
psi[6]   2.4 3.3333333 3.2085387
psi[7]   8.1 3.2098765 2.8612011
psi[8]   2.3 3.0434783 2.0277131
psi[9]   2.0 3.0000000 1.9612862
psi[10]  6.6 3.0303030 2.7877123
psi[11]  4.4 2.9545455 2.3265476
psi[12]  1.8 2.7777778 2.2834983
psi[13]  1.1 2.7272727 1.6921620
psi[14]  3.3 2.4242424 2.4896444
psi[15]  7.8 2.1794872 1.8843974
psi[16]  4.6 1.9565217 1.8238506
psi[17]  1.1 1.8181818 1.3862154
psi[18]  4.2 1.6666667 1.3842353
psi[19]  5.5 1.6363636 1.3996644
psi[20]  4.4 1.5909091 1.4299443
psi[21] 10.5 1.5238095 1.3788745
psi[22] 22.7 1.3656388 1.3790154
psi[23]  8.8 1.2500000 1.1936484
psi[24]  5.6 1.2500000 1.1437502
psi[25] 15.5 1.2258065 1.1111726
psi[26] 12.5 1.2000000 1.0579732
psi[27]  6.0 1.1666667 1.0684057
psi[28]  9.0 1.1111111 1.0315289
psi[29] 14.4 1.1111111 1.0973107
psi[30] 10.2 1.0784314 1.0779852
psi[31]  4.8 1.0416667 0.9749365
```

```
psi[32]  2.9 1.0344828 1.5522988
psi[33]  7.0 1.0000000 1.0155038
psi[34]  8.5 0.9411765 0.9204860
psi[35] 12.3 0.8943089 0.9025271
psi[36] 10.1 0.8910891 0.7844556
psi[37] 12.7 0.8661417 0.8843448
psi[38]  9.4 0.8510638 0.7633810
psi[39]  7.2 0.8333333 1.0093637
psi[40]  5.3 0.7547170 0.6476567
psi[41] 18.8 0.5319149 0.5411346
psi[42] 15.8 0.5063291 0.6622312
psi[43]  4.3 0.4651163 0.8804469
psi[44] 14.6 0.4109589 0.4450161
psi[45] 50.7 0.3747535 0.3928164
psi[46]  8.2 0.3658537 0.5305133
psi[47]  5.6 0.3571429 0.4607572
psi[48]  9.3 0.3225806 0.4206741
psi[49] 88.7 0.3156708 0.3284214
psi[50] 19.6 0.3061224 0.3691449
psi[51]  3.4 0.2941176 0.4721751
psi[52]  3.6 0.2777778 0.4339157
psi[53]  5.7 0.1754386 0.3839701
psi[54]  7.0 0.1428571 0.3420427
psi[55]  4.2 0.0000000 0.6199531
psi[56]  1.8 0.0000000 0.6964520
```

*You can see the smoothing. Note that the higher the expected counts the less the smoothing.*

*Model 2: Model 1 was nonspatial. We will fit a CAR model to the data.*

*We assume that*

$$\psi_i|\psi_{j\neq i} \sim N\left(\frac{\sum_{j\neq i} c_{ij}\psi_j}{\sum_{j\neq i} c_{ij},}, \frac{1}{\nu_{car}\sum_{j\neq i} cij}\right).$$

*The prior mean of each $\psi_i$ is a weighted average of the $\psi$'s of the surrounding neighbors. The hyperparameter $\nu_{car}$ is related to the conditional variance (see Waller and Gotway, page 415). We need to specify a hyperprior for $\nu_{car}$. I am going to choose a prior used by Carlin and Louis in their discussion of Bayesian modeling of the lip cancer data:*

$$\frac{1}{\nu_{car}} \sim Gamma\,(0.1, 0.1)\,.$$

*This is relatively uninformative. The $c_{ij}$'s are the weights assigned to neighbors $j$ of region $i$.*

```
model
{
   for (i in 1 : regions) {
      O[i] ~ dpois(mu[i])
      log(mu[i]) <- log(E[i]) + beta0 + beta1*aff[i]/10 + psi[i]
      SMRhat[i] <- 100 * mu[i] / E[i]
      SMRraw[i] <- 100* O[i] / E[i]
   }
   psi[1:regions] ~ car.normal(adj[], wts[], num[], tau.car)
   beta0 ~ dnorm(0.0, 1.0E-5)  # vague prior on grand intercept
   beta1 ~ dnorm(0.0, 1.0E-5)  # vague prior on covariate effect
   tau.car ~ dgamma(1.0E-1,1.0E-1)  #   (1999, Bayesian Statistics 6)
}
```

*WinBugs has a* `car.normal` *distribution. The arguments are* `adj` *which contains the observed counts in identified neighbors,* `wts` *which contains the weights* $c_{ij}$*, and* `num` *contains the number of neighbors associated with each region. For example, the first region has* 4 *neighbors with observed counts of* 5, 9, 11, 19 *and weights of* 1 *each. All of the other* 51 *regions are given a weight of* 0*.*

```
> wts[1:4]
[1] 1 1 1 1
> adj[1:4]
[1]  5  9 11 19
> num[1]
[1] 4
```

*All the data are available in a text file on a web site maintained by Carlin. The R code to fit the model is given below.*

```
> lipcancer.data<-c("O","E","adj","wts","num","aff")
> lipcancer.parameters<-c("beta0","beta1","psi","tau.car")
> lipcancer.inits
function(){
list(beta0=rnorm(1),beta1=rnorm(1),psi=rnorm(regions),tau.car=rnorm(1,1,.1))}
> lipcancer.model2<-bugs(lipcancer.data,lipcancer.inits,lipcancer.parameters,
+ "lipcancer.model2.bug",n.chains=5,n.iter=10000)
```

*The output is given below.*

```
> (lipcancer.model2)
Inference for Bugs model at "lipcancer.model2.bug", fit using WinBUGS,
 5 chains, each with 10000 iterations (first 5000 discarded), n.thin = 25
 n.sims = 1000 iterations saved
          mean  sd  2.5%   25%   50%   75% 97.5% Rhat n.eff
beta0     -0.2 0.1  -0.4  -0.3  -0.2  -0.1   0.0    1  1000
beta1      0.3 0.1   0.1   0.3   0.4   0.4   0.6    1   720
psi[1]     1.1 0.3   0.6   0.9   1.1   1.3   1.7    1   270
psi[2]     1.1 0.2   0.8   1.0   1.1   1.2   1.5    1  1000
psi[3]     1.0 0.3   0.5   0.8   1.0   1.2   1.5    1  1000
psi[4]     0.4 0.3  -0.3   0.2   0.4   0.6   0.9    1   890
psi[5]     1.0 0.2   0.6   0.9   1.0   1.1   1.4    1  1000
psi[6]     0.7 0.3   0.0   0.4   0.7   0.9   1.3    1  1000
psi[7]     0.9 0.2   0.6   0.8   0.9   1.0   1.3    1  1000
psi[8]     0.9 0.4   0.2   0.7   1.0   1.2   1.6    1   350
psi[9]     0.7 0.3   0.2   0.5   0.7   0.9   1.2    1  1000
psi[10]    0.7 0.2   0.3   0.6   0.7   0.8   1.1    1   510
psi[11]    1.0 0.2   0.5   0.8   1.0   1.1   1.4    1   690
psi[12]    0.8 0.3   0.2   0.6   0.8   1.0   1.4    1  1000
psi[13]    0.8 0.3   0.1   0.6   0.8   1.0   1.4    1   860
psi[14]    0.0 0.3  -0.5  -0.2   0.0   0.2   0.6    1   520
psi[15]    0.5 0.2   0.1   0.3   0.5   0.7   0.9    1  1000
psi[16]    0.4 0.2  -0.1   0.2   0.4   0.5   0.8    1  1000
psi[17]    0.5 0.3   0.0   0.3   0.6   0.7   1.1    1  1000
psi[18]    0.1 0.3  -0.4  -0.1   0.1   0.2   0.6    1   210
psi[19]    0.6 0.2   0.2   0.5   0.6   0.8   1.0    1  1000
psi[20]    0.1 0.3  -0.5   0.0   0.2   0.3   0.7    1  1000
psi[21]    0.2 0.2  -0.2   0.1   0.3   0.4   0.7    1  1000
psi[22]    0.0 0.2  -0.4  -0.1   0.0   0.1   0.4    1  1000
```

```
psi[23]    0.0 0.2  -0.5  -0.2   0.0   0.1   0.4    1   780
psi[24]   -0.3 0.2  -0.7  -0.4  -0.3  -0.1   0.2    1   440
psi[25]    0.3 0.2  -0.1   0.2   0.3   0.5   0.7    1  1000
psi[26]    0.1 0.2  -0.3   0.0   0.1   0.3   0.6    1  1000
psi[27]   -0.2 0.3  -0.7  -0.4  -0.2   0.0   0.4    1  1000
psi[28]    0.0 0.3  -0.6  -0.2   0.0   0.1   0.4    1  1000
psi[29]    0.0 0.2  -0.3  -0.1   0.0   0.2   0.4    1  1000
psi[30]   -0.3 0.2  -0.8  -0.5  -0.3  -0.2   0.1    1  1000
psi[31]   -0.3 0.2  -0.7  -0.4  -0.3  -0.1   0.2    1   850
psi[32]   -0.3 0.3  -1.0  -0.5  -0.3  -0.1   0.3    1  1000
psi[33]   -0.2 0.3  -0.8  -0.4  -0.2  -0.1   0.3    1  1000
psi[34]   -0.3 0.2  -0.7  -0.5  -0.3  -0.2   0.1    1  1000
psi[35]   -0.2 0.2  -0.7  -0.4  -0.2  -0.1   0.2    1  1000
psi[36]   -0.1 0.3  -0.7  -0.3  -0.1   0.1   0.4    1   420
psi[37]   -0.3 0.2  -0.8  -0.5  -0.3  -0.1   0.1    1  1000
psi[38]   -0.4 0.3  -1.0  -0.6  -0.4  -0.2   0.1    1   590
psi[39]   -0.4 0.2  -0.9  -0.5  -0.4  -0.2   0.1    1  1000
psi[40]   -0.4 0.3  -1.0  -0.6  -0.4  -0.3   0.1    1  1000
psi[41]   -0.5 0.2  -1.0  -0.6  -0.5  -0.3  -0.1    1  1000
psi[42]   -0.7 0.2  -1.1  -0.8  -0.7  -0.5  -0.3    1  1000
psi[43]   -0.5 0.3  -1.1  -0.7  -0.5  -0.3   0.1    1   670
psi[44]   -0.6 0.3  -1.1  -0.8  -0.6  -0.4  -0.1    1   720
psi[45]   -0.7 0.2  -1.1  -0.8  -0.7  -0.5  -0.3    1  1000
psi[46]   -0.6 0.3  -1.1  -0.8  -0.6  -0.4  -0.1    1  1000
psi[47]   -0.6 0.3  -1.2  -0.9  -0.6  -0.4  -0.1    1   660
psi[48]   -0.7 0.3  -1.4  -0.9  -0.7  -0.5  -0.1    1  1000
psi[49]   -0.8 0.2  -1.2  -0.9  -0.8  -0.7  -0.5    1  1000
psi[50]   -0.5 0.3  -1.0  -0.7  -0.5  -0.3   0.0    1  1000
psi[51]   -0.6 0.4  -1.3  -0.9  -0.6  -0.4   0.0    1   740
psi[52]   -0.7 0.4  -1.4  -0.9  -0.7  -0.4   0.1    1  1000
psi[53]   -0.8 0.3  -1.6  -1.0  -0.8  -0.6  -0.2    1  1000
psi[54]   -0.8 0.3  -1.5  -1.0  -0.8  -0.6  -0.2    1  1000
psi[55]   -0.5 0.3  -1.2  -0.7  -0.5  -0.3   0.0    1  1000
psi[56]   -0.5 0.3  -1.0  -0.6  -0.4  -0.3   0.1    1   840
tau.car    2.0 0.7   1.0   1.5   1.9   2.4   3.5    1  1000
deviance 268.6 9.8 250.8 261.3 268.1 275.3 287.6    1  1000

For each parameter, n.eff is a crude measure of effective sample size,
and Rhat is the potential scale reduction factor (at convergence, Rhat=1).

DIC info (using the rule, pD = Dbar-Dhat)
pD = 28.8 and DIC = 297.4
DIC is an estimate of expected predictive error (lower deviance is better).
```

*The chain appears to have converged.*

*The expected counts, raw SMR values and modeled SMR values are*

```
> a<-exp(log(E)+lipcancer.model2$summary[3:58,1]-0.206+0.349*aff/10)
> cbind(E,O/E,a/E)
psi[1]   1.4 6.4285714 4.4223499
psi[2]   8.7 4.4827586 4.2876586
psi[3]   3.0 3.6666667 3.1877037
psi[4]   2.5 3.6000000 2.7098985
psi[5]   4.3 3.4883721 3.1096784
```

```
psi[6]   2.4 3.3333333 3.6207809
psi[7]   8.1 3.2098765 2.8786179
psi[8]   2.3 3.0434783 2.6561631
psi[9]   2.0 3.0000000 2.0973119
psi[10]  6.6 3.0303030 2.8483319
psi[11]  4.4 2.9545455 2.7374672
psi[12]  1.8 2.7777778 3.2333890
psi[13]  1.1 2.7272727 2.5079631
psi[14]  3.3 2.4242424 1.9116828
psi[15]  7.8 2.1794872 1.7182924
psi[16]  4.6 1.9565217 2.0457056
psi[17]  1.1 1.8181818 1.9910818
psi[18]  4.2 1.6666667 1.1107450
psi[19]  5.5 1.6363636 1.9437470
psi[20]  4.4 1.5909091 1.3254833
psi[21] 10.5 1.5238095 1.3295246
psi[22] 22.7 1.3656388 1.4308434
psi[23]  8.8 1.2500000 1.1217586
psi[24]  5.6 1.2500000 0.7834026
psi[25] 15.5 1.2258065 1.1510373
psi[26] 12.5 1.2000000 0.9668397
psi[27]  6.0 1.1666667 0.8878764
psi[28]  9.0 1.1111111 1.0032277
psi[29] 14.4 1.1111111 1.2114191
psi[30] 10.2 1.0784314 0.8349504
psi[31]  4.8 1.0416667 0.7996204
psi[32]  2.9 1.0344828 1.3602045
psi[33]  7.0 1.0000000 0.9090818
psi[34]  8.5 0.9411765 0.7382260
psi[35] 12.3 0.8943089 0.8163616
psi[36] 10.1 0.8910891 0.7244157
psi[37] 12.7 0.8661417 0.8479132
psi[38]  9.4 0.8510638 0.5515079
psi[39]  7.2 0.8333333 0.9577039
psi[40]  5.3 0.7547170 0.5225849
psi[41] 18.8 0.5319149 0.5108063
psi[42] 15.8 0.5063291 0.7296206
psi[43]  4.3 0.4651163 0.8906313
psi[44] 14.6 0.4109589 0.4396217
psi[45] 50.7 0.3747535 0.4271306
psi[46]  8.2 0.3658537 0.5606171
psi[47]  5.6 0.3571429 0.4419185
psi[48]  9.3 0.3225806 0.4036008
psi[49] 88.7 0.3156708 0.3524438
psi[50] 19.6 0.3061224 0.5125485
psi[51]  3.4 0.2941176 0.4435849
psi[52]  3.6 0.2777778 0.4141181
psi[53]  5.7 0.1754386 0.3686002
psi[54]  7.0 0.1428571 0.3849613
psi[55]  4.2 0.0000000 0.8288068
psi[56]  1.8 0.0000000 0.7303320
```

*Once again we see the smoothing effect. But the effect is different. Neighbors now have more of an impact than before. In effect Model 1 could be viewed as modeling extra Poisson variability in a region wide manner (as*

*discussed by Banerjee, Carlin, and Gelfand) and Model 2 is modeling overdispersion more locally (a clustering approach).*

*It is possible to incorporate both types of "random effects" in one model in a Bayesian approach. This would not be possible in a frequentist approach because the region wide random effect and the local random effect would not be identifiable. However, we can circumvent this problem in a Bayesian approach by assigning different priors to the* 2 *random effects. We will not look at that here although Waller and Gotway mention it and show results of fitting such a model to the leukemia data set.*

*We have left out a lot and there are many questions that are raised by the above analysis. Which model is "better"? Where did those priors and hyperpriors come from and how can they be justified? How robust are the results to selection of priors and hyperpriors? Did the chains actually converge to the correct posterior distribution? How would different definitions of neighbors and different neighbor weighting schemes affect the results? What about a geostatistical approach? We will go no further with Bayesian models. Good texts to consult are "Bayesian Methods for Data Analysis" (3rd edition) by Carlin and Louis, "Bayesian Data Analysis" by Gelman, Carlin, Stern, and Rubin, and "Hierarchical Modeling and Analysis for Spatial Data" by Banerjee, Carlin, and Gelfand.*