Stat 534 Homework 8

Kenny Flagg

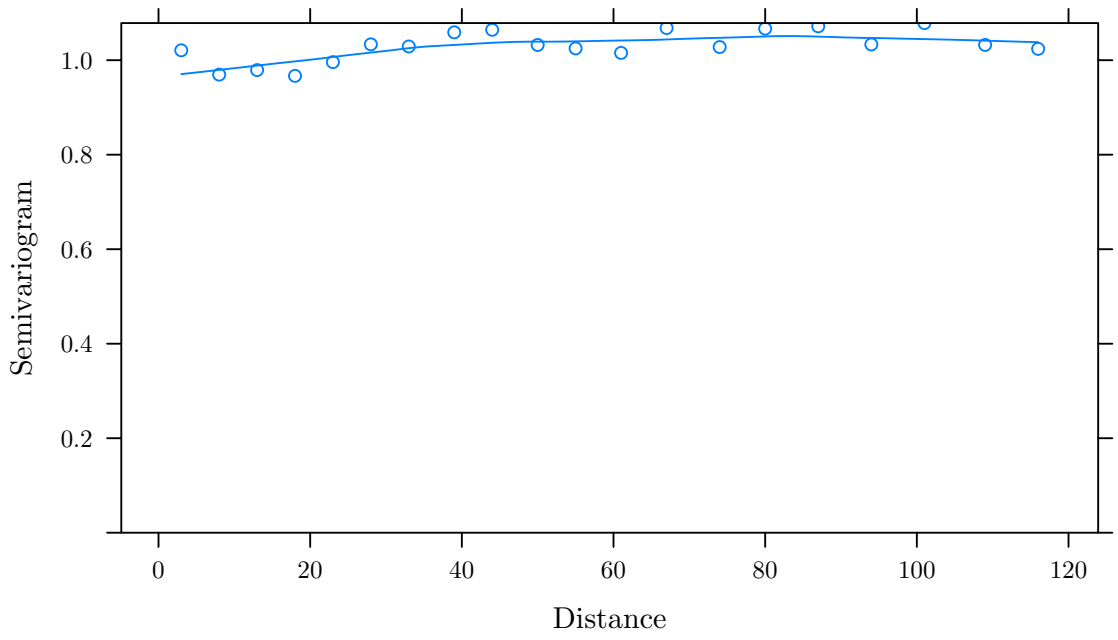
March 31, 2017

1. *The New York Leukemia data are attached along with homework as a comma delimited text file. You are to explore how excluding 3 potential outliers (observations 110, 120, and 121) affects the residual spatial autocorrelation, regression results, and the conclusions obtained from them. Is weighting still necessary? Fit OLS, WLS, and appropriate GLS models following my steps in the example in class.*

   I begin by fitting the OLS model, omitting the outliers, and examining the semivariogram of the residuals. This semivariogram does not look very different from the semivariogram of from the model fit the the entire dataset, so I again use an initial range of 18 and an initial nugget of 0.9. I compare all six models on the next page.

```
library(nlme)
leukemia <- read.csv('leukemia.csv')

leuk_ols <- gls(z ~ pexp + age65 + home, data = leukemia, subset = -c(110, 120, 121))
plot(Variogram(leuk_ols, maxDist = 120))
```



```
# Fit the weighted and spatial models.
leuk_wls <- gls(z ~ pexp + age65 + home, data = leukemia, subset = -c(110, 120, 121),
                weights = varFixed(~1/pop))
leuk_exp <- gls(z ~ pexp + age65 + home, data = leukemia, subset = -c(110, 120, 121),
                correlation = corExp(c(6, 0.9), form = ~x+y, nugget = TRUE))
leuk_sph <- gls(z ~ pexp + age65 + home, data = leukemia, subset = -c(110, 120, 121),
                correlation = corSpher(c(6, 0.9), form = ~x+y, nugget = TRUE))
leuk_exp_w <- gls(z ~ pexp + age65 + home, data = leukemia, subset = -c(110, 120, 121),
                  correlation = corExp(c(6, 0.9), form = ~x+y, nugget = TRUE),
                  weights = varFixed(~1/pop))
leuk_sph_w <- gls(z ~ pexp + age65 + home, data = leukemia, subset = -c(110, 120, 121),
                  correlation = corSpher(c(6, 0.9), form = ~x+y, nugget = TRUE),
                  weights = varFixed(~1/pop))
```

The table below shows the `pexp` coefficient estimates and AICs for all six models. All the slope estimates are similar, but the AICs indicate that the OLS model has the best fit. With the outliers removed, there is no need for weighting.

	$\widehat{\beta}_1$	$SE\left(\widehat{\beta}_1\right)$	AIC	ΔAIC
OLS	0.0750	0.0314	506.741	0.000
WLS	0.0767	0.0274	509.118	2.377
Exponential	0.0864	0.0354	510.020	3.279
Spherical	0.0900	0.0359	509.201	2.460
Exponential, Weighted	0.0934	0.0322	511.072	4.331
Spherical, Weighted	0.0944	0.0320	510.334	3.593

2. *Imagine a lattice process on a 2×3 rectangle. The sites $\mathbf{s}_1$, $\mathbf{s}_2$, and $\mathbf{s}_3$ make up the first row, the remaining sites make up the second row. Assume that the spatial connectivity matrix is given by*

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

For a simultaneous and conditional autoregressive scheme with

$$\text{Var}\left[\mathbf{Z}(\mathbf{s})\right] = \sigma^2 \left(\mathbf{I} - \rho\mathbf{W}\right)^{-1} \left(\mathbf{I} - \rho\mathbf{W}'\right)^{-1}$$

and

$$\text{Var}\left[\mathbf{Z}(\mathbf{s})\right] = \sigma^2 \left(\mathbf{I} - \rho\mathbf{W}\right)^{-1}$$

respectively, and with $\rho = 0.25$ do the following:

(a) *Identify the neighbors of lattice cell $\mathbf{s}_2$.*

The neighbors of $\mathbf{s}_2$, corresponding to the columns of $\mathbf{W}$ that are nonzero in the second row, are $\mathbf{s}_1$, $\mathbf{s}_3$, and $\mathbf{s}_5$.

(b) *Compute the variance-covariance matrices for the SAR and CAR schemes.*

For SAR,

```
W <- matrix(c(0, 1, 0, 1, 0, 0,
              1, 0, 1, 0, 1, 0,
              0, 1, 0, 0, 0, 1,
              1, 0, 0, 0, 1, 0,
              0, 1, 0, 1, 0, 1,
              0, 0, 1, 0, 1, 0),
            nrow = 6)
rho <- 0.25

Sigma_SAR <- solve(diag(6) - rho * W) %*% solve(diag(6) - rho * t(W))
print(Sigma_SAR)
```

```
          [,1]      [,2]      [,3]      [,4]      [,5]      [,6]
[1,] 1.7065624 1.1308206 0.4976735 0.9782448 0.8283631 0.4093559
[2,] 1.1308206 2.2042359 1.1308206 0.8283631 1.3876008 0.8283631
[3,] 0.4976735 1.1308206 1.7065624 0.4093559 0.8283631 0.9782448
[4,] 0.9782448 0.8283631 0.4093559 1.7065624 1.1308206 0.4976735
[5,] 0.8283631 1.3876008 0.8283631 1.1308206 2.2042359 1.1308206
[6,] 0.4093559 0.8283631 0.9782448 0.4976735 1.1308206 1.7065624
```

so

$$\text{Var}\left[\mathbf{Z}(\mathbf{s})\right] = \sigma^2 \begin{bmatrix} 1.707 & 1.131 & 0.498 & 0.978 & 0.828 & 0.409 \\ 1.131 & 2.204 & 1.131 & 0.828 & 1.388 & 0.828 \\ 0.498 & 1.131 & 1.707 & 0.409 & 0.828 & 0.978 \\ 0.978 & 0.828 & 0.409 & 1.707 & 1.131 & 0.498 \\ 0.828 & 1.388 & 0.828 & 1.131 & 2.204 & 1.131 \\ 0.409 & 0.828 & 0.978 & 0.498 & 1.131 & 1.707 \end{bmatrix}.$$

For CAR,

```
Sigma_CAR <- solve(diag(6) - rho * W)
print(Sigma_CAR)
```

```
           [,1]      [,2]       [,3]       [,4]      [,5]       [,6]
[1,] 1.17929607 0.3726708 0.11262940 0.34451346 0.1987578 0.07784679
[2,] 0.37267081 1.2919255 0.37267081 0.19875776 0.4223602 0.19875776
[3,] 0.11262940 0.3726708 1.17929607 0.07784679 0.1987578 0.34451346
[4,] 0.34451346 0.1987578 0.07784679 1.17929607 0.3726708 0.11262940
[5,] 0.19875776 0.4223602 0.19875776 0.37267081 1.2919255 0.37267081
[6,] 0.07784679 0.1987578 0.34451346 0.11262940 0.3726708 1.17929607
```

so

$$\text{Var}\left[\mathbf{Z}(\mathbf{s})\right] = \sigma^2 \begin{bmatrix} 1.1793 & 0.3727 & 0.1126 & 0.3445 & 0.1988 & 0.0778 \\ 0.3727 & 1.2919 & 0.3727 & 0.1988 & 0.4224 & 0.1988 \\ 0.1126 & 0.3727 & 1.1793 & 0.0778 & 0.1988 & 0.3445 \\ 0.3445 & 0.1988 & 0.0778 & 1.1793 & 0.3727 & 0.1126 \\ 0.1988 & 0.4224 & 0.1988 & 0.3727 & 1.2919 & 0.3727 \\ 0.0778 & 0.1988 & 0.3445 & 0.1126 & 0.3727 & 1.1793 \end{bmatrix}.$$

(c) *Determine which of the processes is second-order stationary. Justify your answer.*

Neither process is second-order stationary because the variance is not spatially constant. For both processes, the variance of $Z(\mathbf{s}_2)$ and $Z(\mathbf{s}_5)$ is larger than the variance in other cells.

(d) *Describe the correlation patterns that result. Are observations equicorrelated that are the same distance apart? Do correlations decrease with increasing lag distance?*

The SAR correlation matrix is

```
D_SAR_invsqrt <- diag(1 / sqrt(diag(Sigma_SAR)))
D_SAR_invsqrt %*% Sigma_SAR %*% D_SAR_invsqrt
```

```
          [,1]      [,2]      [,3]      [,4]      [,5]      [,6]
[1,] 1.0000000 0.5830470 0.2916234 0.5732253 0.4271010 0.2398717
[2,] 0.5830470 1.0000000 0.5830470 0.4271010 0.6295155 0.4271010
[3,] 0.2916234 0.5830470 1.0000000 0.2398717 0.4271010 0.5732253
[4,] 0.5732253 0.4271010 0.2398717 1.0000000 0.5830470 0.2916234
[5,] 0.4271010 0.6295155 0.4271010 0.5830470 1.0000000 0.5830470
[6,] 0.2398717 0.4271010 0.5732253 0.2916234 0.5830470 1.0000000
```

and the CAR correlation matrix is

```
D_CAR_invsqrt <- diag(1 / sqrt(diag(Sigma_CAR)))
D_CAR_invsqrt %*% Sigma_CAR %*% D_CAR_invsqrt
```

```
           [,1]      [,2]       [,3]       [,4]      [,5]       [,6]
[1,] 1.00000000 0.3019223 0.09550562 0.29213483 0.1610252 0.06601124
[2,] 0.30192232 1.0000000 0.30192232 0.16102524 0.3269231 0.16102524
[3,] 0.09550562 0.3019223 1.00000000 0.06601124 0.1610252 0.29213483
[4,] 0.29213483 0.1610252 0.06601124 1.00000000 0.3019223 0.09550562
[5,] 0.16102524 0.3269231 0.16102524 0.30192232 1.0000000 0.30192232
[6,] 0.06601124 0.1610252 0.29213483 0.09550562 0.3019223 1.00000000
```

so we see that the correlations generally decrease as the lag increases, but not all pairs at a given lag are equally correlated. Neighbors within a row are more highly correlated than neighbors across rows, but pairs of cells in the same row with lag 2 are less correlated than pairs in different rows with lag 2.

Assuming square grid cells, the lag distance is equivalent to the Manhattan distance. The correlation structure appears to reflect the Euclidean distance because diagonally adjacent cells (lag 2; cells $\mathbf{s}_1$ and $\mathbf{s}_5$, for example) are more correlated than pairs at lag 2 but in the same row (e.g., cells $\mathbf{s}_1$ and $\mathbf{s}_3$), and the latter have a larger Euclidean distance.

3. *We have 5 binomial count responses at 5 locations, i.e. the number of successes out of $n(\mathbf{s}_i)$ trials. Assume a single covariate X_1 and that there is no overdispersion. Find the diagonal elements of $\mathbf{V}_\mu^{1/2}$ being sure to express these in terms of β_0 and β_1.*

The model is

$$\begin{aligned} Z(\mathbf{s}_i) &\sim \text{Binomial}(n(\mathbf{s}_i), \mu(\mathbf{s}_i)); \\ \log\left(\frac{\mu(\mathbf{s}_i)}{1-\mu(\mathbf{s}_i)}\right) &= \beta_0 + \beta_1 x_1(\mathbf{s}_i) \end{aligned}$$

for $i = 1, \ldots, 5$. Then, assuming no overdispersion, the ith diagonal element of $\mathbf{V}_\mu^{1/2}$ is

$$\begin{aligned} \sqrt{v(\mu(\mathbf{s}_i))} &= \sqrt{n(\mathbf{s}_i)\mu(\mathbf{s}_i)(1-\mu(\mathbf{s}_i))} \\ &= \sqrt{n(\mathbf{s}_i)\left(\frac{\exp\left[\beta_0 + \beta_1 x_1(\mathbf{s}_i)\right]}{1+\exp\left[\beta_0 + \beta_1 x_1(\mathbf{s}_i)\right]}\right)\left(\frac{1}{1+\exp\left[\beta_0 + \beta_1 x_1(\mathbf{s}_i)\right]}\right)} \\ &= \frac{\sqrt{n(\mathbf{s}_i)\exp\left[\beta_0 + \beta_1 x_1(\mathbf{s}_i)\right]}}{1+\exp\left[\beta_0 + \beta_1 x_1(\mathbf{s}_i)\right]}. \end{aligned}$$