Stat 534 Homework 9

Kenny Flagg

April 7, 2017

1. *I would like you to try and reproduce some of the results in Table 9.10 on page 395 in Waller and Gotway. They fit several models to the Scottish lipcancer data set. The ones I am interested in are models $S + OD$ and the $MGLM$. I want you to use* `glmmPQL` *to see if you can get close to their fitted results for $S + OD$. I provide you with some code for fitting the nonlinear least squares model to see what you can come up with as an approximation to their $MGLM$. We did something similar with the Virginia lead level data earlier (take a look back at page 50 or thereabouts). I have spent hours every time I teach this course trying to come close to their results. So I have two possible outcomes in mind here: (1) you (or some of you) are able to figure out where I am going wrong which I would love to see, and/or (2) you get a birds-eye view of the issues related to trying to incorporate spatial correlation structures into generalized linear models.*

   The data set they provided on their website is attached. The last 2 columns are the transformed spatial locations they discuss in the text. If you look at the SAS code they used they divided these by 1000. I encourage you to work together on this one. It will go much smoother. I would like for you all to be ready to discuss this on Friday. I have been a bit easy on the exact time homeworks are due, accepting several late over the course of the semester. But I want to be able to talk about this in class on Friday so you need to have them ready by class time.

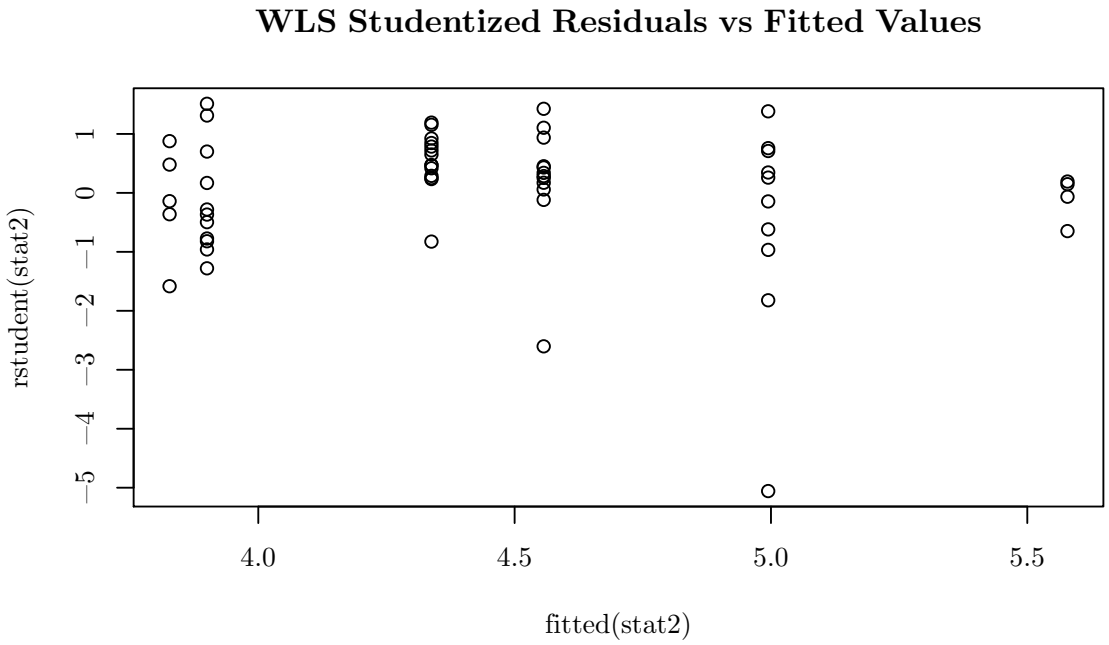
   I first spent some time trying to understand their semivariogram plot. They say they computed it from the studentized residuals of a WLS regression of $\log(\texttt{SMR} + 1)$ against `aff` using the expected counts as weights. I translated their SAS code to get the plot on the next page. Two counties are clear outliers with studentized residuals below -2, and removing them yields a very different semivariogram. Waller and Gotway don't give any indication that they give these observations special handling, so for the purpose of replicating their results I do not omit them.

```
lipcancer <- read.csv('lipcancer.WallerGotway.csv')

# I'm mostly sticking to the same variable names as in their SAS code:
#   http://web1.sph.emory.edu/users/lwaller/book/ch9/scotglms.sas
lipcancer$aff1 <- lipcancer$aff/10
lipcancer$logni <- log(lipcancer$exp)
lipcancer$x <- lipcancer$xcoord/1000
lipcancer$y <- lipcancer$ycoord/1000
lipcancer$logsmr <- log(lipcancer$smr + 1)
lipcancer$expct <- lipcancer$exp
```

```
# Weighted least squares regression.
stat2 <- lm(logsmr ~ aff1, weights = expct, data = lipcancer)
lipcancer$resids <- rstudent(stat2)

plot(rstudent(stat2) ~ fitted(stat2), main = 'WLS Studentized Residuals vs Fitted Values')
```



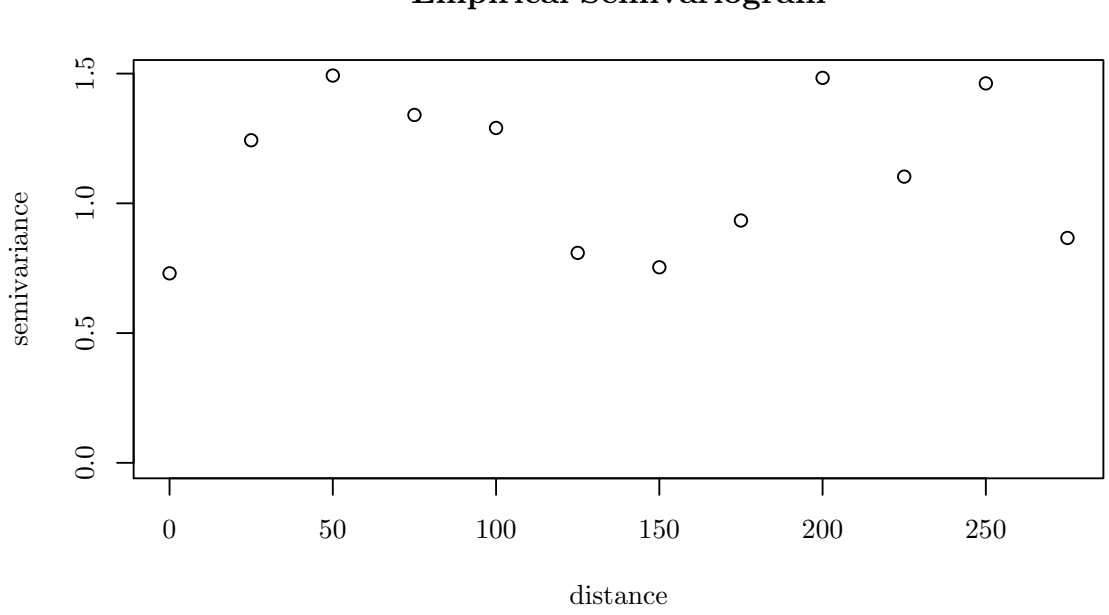
Note that the smallest distance in the dataset is 7.1 km, but their semivariogram has three points at distances smaller than that (the caption claims that the distances are indeed in km). It would seem like either they plotted the semivariogram of a different dataset or they rescaled the distances somehow. The largest distance is 644.4 km, so I took a wild guess that they estimated the semivariogram up to about half the largest possible distance, and used units of 10 km for some reason. I eyeballed the distances on their plot and set manual bins, resulting in the plot on the next page. The vertical scales are slightly different, but the shape is very similar. At the top of page 395 they say there is negative dependence at 72 km to 105 km, but I don't know where they see that.

```
# Trying to recreate their empirical semivariogram.
resid_geodata <- as.geodata(lipcancer, coords.col = c('x', 'y'), data.col = 'resids')

# uvec specifies the bin centers.
resid_vario <- variog(resid_geodata, uvec = seq(0, 275, 25))

variog: computing omnidirectional variogram

plot(resid_vario, main = 'Empirical Semivariogram')
```



Next, I checked that I understood what I was doing by fitting their non-spatial models. The R Poisson regression results below match their SAS results.

```
# Poisson regression without adjustment for overdispersion.
PR <- glm(observed ~ aff1, family = poisson, offset = logni, data = lipcancer)
summary(PR) # Agrees!


Call:
glm(formula = observed ~ aff1, family = poisson, data = lipcancer,
    offset = logni)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-4.7632  -1.2156   0.0967   1.3362   4.7130

Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept) -0.54227    0.06952   -7.80 6.21e-15
aff1         0.73732    0.05956   12.38  < 2e-16

(Dispersion parameter for poisson family taken to be 1)

    Null deviance: 380.73  on 55  degrees of freedom
Residual deviance: 238.62  on 54  degrees of freedom
AIC: 450.6

Number of Fisher Scoring iterations: 5
```

Their Poisson regression with overdispersion results are matched by a quasiPoisson regression in R, as seen below.

```
# Poisson regression with adjustment for overdispersion.
PR_OD <- glm(observed ~ aff1, family = quasipoisson, offset = logni, data = lipcancer)
summary(PR_OD) # Agrees!


Call:
glm(formula = observed ~ aff1, family = quasipoisson, data = lipcancer,
    offset = logni)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-4.7632  -1.2156   0.0967   1.3362   4.7130

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  -0.5423     0.1542  -3.517 0.000893
aff1          0.7373     0.1321   5.583 7.89e-07

(Dispersion parameter for quasipoisson family taken to be 4.917964)

    Null deviance: 380.73  on 55  degrees of freedom
Residual deviance: 238.62  on 54  degrees of freedom
AIC: NA

Number of Fisher Scoring iterations: 5
```

The nonspatial model with uncorrelated random effects and "adjusted for overdispersion" is run-of-the-mill GLMM, because `glmmPQL` estimates a within-group residual variance parameter (even though that isn't strictly a "Poisson" model because of the overdispersion parameter). The coefficients and standard errors match those in the book, and squaring the random effect and residual standard deviations gives values that are off from the variances in the book by only about 0.02.

```
# Poisson GLMM adjusted for overdispersion.
GLMMI <- glmmPQL(observed ~ aff1 + offset(logni), random = ~1|county,
                 family = poisson, data = lipcancer)
summary(GLMMI) # Coefs/SEs match, variances are close.

Linear mixed-effects model fit by maximum likelihood
 Data: lipcancer
  AIC BIC logLik
   NA  NA     NA

Random effects:
 Formula: ~1 | county
        (Intercept) Residual
StdDev:   0.5202324 1.216389

Variance function:
 Structure: fixed weights
```

```
 Formula: ~invwt
Fixed effects: observed ~ aff1 + offset(logni)
                 Value Std.Error DF   t-value p-value
(Intercept) -0.4260939 0.1552171 54 -2.745148  0.0082
aff1         0.6801942 0.1382437 54  4.920255  0.0000
 Correlation:
     (Intr)
aff1 -0.796

Standardized Within-Group Residuals:
         Min           Q1          Med           Q3          Max
-1.636260822 -0.504104920 -0.005553504  0.383545967  1.310047597

Number of Observations: 56
Number of Groups: 56
```

Waller and Gotway appear to construct non-overdispersed spatial models in SAS by fixing one of the covariance parameters, but I don't understand their `parms` statement. The effect of their `parms` statement might be to fix the nugget (within-county variance) at 1, but that is a wild guess. If that guess is correct, and if SAS uses the nugget as a variance multiplier (analogous to the quasiPoisson dispersion), then this should have the correct effect: conditional on the random intercept, the observed count has a Poisson distribution without overdispersion.

I would construct these models by placing the correlation structure on the random effects, instead of placing it on the "residuals", and estimating the nugget. Unfortunately, I don't know of any R functions that allow that. So I'm skipping the non-overdispersion spatial mixed models.

The spatial GLMM with overdispersion is confusing. The nugget should not be necessary because `glmmPQL` estimates a residual variance to account for overdispersion. I initially tried to include both parameters anyway; I played around with the `lmeControl` options and got a variety of cryptic error messages, and never got the optimizer to happily converge.

When leaving out the nugget and using `optim` to do the optimization, `glmmPQL` finds estimates in only 7 reweighting iterations. The BFGS and L-BFGS algorithms both lead to the same estimates, but they are very different from the SAS results.

With the `nlminb` optimizer, `glmmPQL` gets stuck in a feedback loop where the working response changes a bit every iteration, so the initial coefficient values change, then the final coefficient estimates for the iteration are different from the previous iteration, and so the working responses change again. One time I ran `glmmPQL` for 10,000 iterations and it showed no sign of converging. It's strange that it converges so easily using `optim`, so I am suspicious about the `optim` results and curious if SAS has convergence issues too.

```
# Conditional spatial GLMM adjusted for overdispersion, spherical covariance.
#corr_S_OD <- corSpher(value = c(263, 0.38), form = ~x+y, nugget = TRUE)
corr_S_OD <- corSpher(value = 263, form = ~x+y)
corr_S_OD <- Initialize(corr_S_OD, lipcancer)
S_OD <- glmmPQL(observed ~ aff1 + offset(logni), random = ~1|county,
                correlation = corr_S_OD, family = poisson,
                control = lmeControl(opt = 'optim', 'BFGS'),
                data = lipcancer)
```

```
summary(S_OD) # Way off...

Linear mixed-effects model fit by maximum likelihood
 Data: lipcancer
  AIC BIC logLik
   NA  NA     NA

Random effects:
 Formula: ~1 | county
           (Intercept) Residual
StdDev: 0.000001332513 2.687085

Correlation Structure: Spherical spatial correlation
 Formula: ~x + y | county
 Parameter estimate(s):
   range
91.75018
Variance function:
 Structure: fixed weights
 Formula: ~invwt
Fixed effects: observed ~ aff1 + offset(logni)
                 Value Std.Error DF   t-value p-value
(Intercept) -0.5524042 0.1957897 54 -2.821417  0.0067
aff1         0.5002428 0.1283416 54  3.897746  0.0003
 Correlation:
     (Intr)
aff1 -0.498

Standardized Within-Group Residuals:
       Min         Q1        Med         Q3        Max
-1.2006085 -0.2876362  0.2953889  0.7357674  3.1015412

Number of Observations: 56
Number of Groups: 56
```

Finally, I think the marginal model with overdispersion is conceptually the easiest to understand of their spatial models. It just allows nearby rates to be similar to each other, without going to the trouble of predicting a latent mean for each county. It's a relief than `gnls` gives the same estimates as SAS.

```
# R code for MGLM.
log.model <- function(x1, expct, b0, b1){exp(b0 + log(expct) + b1 * x1)}

corr_MGLM <- corSpher(value = 53, form = ~x+y, nugget = FALSE)
corr_MGLM <- Initialize(corr_MGLM, lipcancer)
MGLM <- gnls(observed ~ log.model(aff1, expct, b0, b1),
             data = lipcancer, start = c(b0 = -0.63, b1 = 0.74),
             correlation = corr_MGLM,
             weights = varPower(form = ~fitted(.), fixed = 0.5))
summary(MGLM) # Agrees!

Generalized nonlinear least squares fit
  Model: observed ~ log.model(aff1, expct, b0, b1)
  Data: lipcancer
```

```
       AIC      BIC    logLik
  361.9901 370.0915 -176.9951

Correlation Structure: Spherical spatial correlation
 Formula: ~x + y
 Parameter estimate(s):
   range
53.16267
Variance function:
 Structure: Power of variance covariate
 Formula: ~fitted(.)
 Parameter estimates:
power
  0.5

Coefficients:
        Value Std.Error   t-value p-value
b0 -0.6268765 0.2055992 -3.049022  0.0036
b1  0.6953030 0.1567750  4.435038  0.0000

 Correlation:
   b0
b1 -0.734

Standardized residuals:
       Min         Q1        Med         Q3        Max
-1.3096364 -0.2894963  0.1825809  0.6746601  2.4820862

Residual standard error: 2.663759
Degrees of freedom: 56 total; 54 residual
```