

Exam 2 solutions

Philip Dixon

4/25/2020

Short answer questions:

Each is worth 5 points; partial credit given.

1) Moran's I

- a) Averaged across the study region, there is no evidence of spatial correlation.

Notes: I was looking for no "evidence" and was especially pleased if you noted that this averaged over the study region.

- b) Local spatial correlation varies across the study region. The R and L ends have positive correlation; the middle has negative.

Note: The variation across the region (+ on ends, - in middle) is why the global statistic has a small coefficient.

2) Model selection

- a) The CAR model is the most appropriate. It has the smallest AIC.
- b) Each model implies a variance for each cell. The spatial pattern of those variances is different for the three models. And, the details of the spatial correlation (how quickly it dies out as you move from a focal region) are also slightly different.
- c) Yes, it is appropriate because the models have been fit by ML (not by REML). Each model is fit to the same data (unlike when you use REML with different trend models).

3) Interpretation of point pattern plots

- a) Evidence of clustering (more points than expected within a distance of a focal point) at distances from nearly 0 up to ca 0.4. The observed $K(x)$ function is outside the 95% simulation bounds for those distances.
- b) Evidence of clustering at distances from nearly 0 up to ca 0.2 and evidence of inhibition at distances from ca 0.2 to 0.5. The $\log g(x)$ function is above the simulation envelope at distances less than 0.2 and below the simulation envelope for most distances from 0.2 to 0.5.
- c) $G(x)$ should show evidence of clustering. Nearest neighbors tend to be close to a point, so $G(x)$ will show the pattern for small distances.

Note: $G(x)$, the nearest-neighbor distribution, is a completely different aspect of a point pattern than $g(x)$, the pair correlation function. Distressing that the two functions have similar names, especially when $f(x)$ and $F(x)$ in non-spatial statistics are very closely related.

- d) Yes, there are issues. Both measures assume constant variance to work as intended. Both plots show non-constant variance.

4) Estimating intensity

- a) approximately 0.06. This has the smallest value of leave-one-out MSE, i.e. is the location of the minimum of the curve.
- b) left plot. 0.06 is a very small bandwidth relative to the size of the study area, so the smoothed intensity plot should have lots of fine-scale detail.

Note: The smoothing parameters used for each intensity plot were 0.065 (the actual Diggle-Berman value), 0.32, and 6.5

5) Data analysis

I expected an answer in three parts:

- The stat methods paragraph - these were much better than last time. Thanks!
- A list of choices and what you looked at to choose.
- Answers to the what the investigators want items.

If I wrote “Item” something, that is a comment about your answer to one of the investigators want list, numbered from top to bottom.

My code and answers focus on the investigators want items.

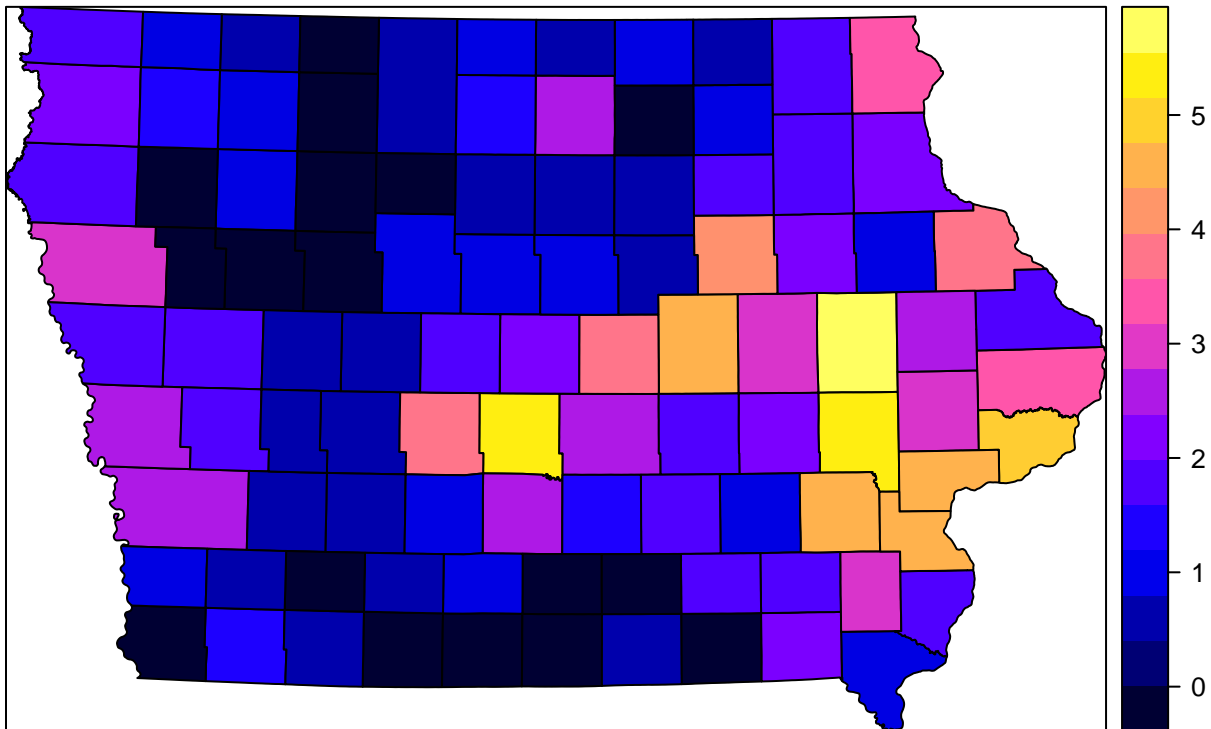
version 1, areal data

```
load('iowa.Rdata')

iowa.nb <- poly2nb(iowa, queen=F)
# binary weights
iowa.w1 <- nb2listw(iowa.nb, style='B')
# row standardized weights
iowa.w2 <- nb2listw(iowa.nb, style='W')
```

Map of log(N+1) counts on 13 Apr:

```
spplot(iowa, 'logapr13')
```



Spatial correlation:

```
# using binary weights
```

```
moran.test(iowa$logapr13, iowa.w1)
```

```
##
## Moran I test under randomisation
##
## data: iowa$logapr13
## weights: iowa.w1
##
## Moran I statistic standard deviate = 8.3407, p-value < 2.2e-16
## alternative hypothesis: greater
## sample estimates:
## Moran I statistic      Expectation      Variance
##      0.532456698      -0.010204082      0.004233084
```

```
# using row standardized weights
```

```
moran.test(iowa$logapr13, iowa.w2)
```

```
##
## Moran I test under randomisation
##
## data: iowa$logapr13
## weights: iowa.w2
##
## Moran I statistic standard deviate = 7.7354, p-value = 5.153e-15
## alternative hypothesis: greater
```

```
## sample estimates:
## Moran I statistic      Expectation      Variance
##      0.509808210      -0.010204082      0.004519171
```

p-values using randomization are similar

Association between 22 Mar and 13 Apr.

Note: This can be assessed by regression. Because the data are observational, the relationship is best described as association. You could use correlation instead of regression, but this is difficult to do when values are spatially correlated. Remember there is a very close relationship between correlation coefficients and regression slopes, e.g. for a one-variable regression, $\text{slope} = r * \text{sd_Y} / \text{sd_X}$. Since one variable is later in time than the other, best to use the later (logapr13) as Y.

What model should we use? Consider 4: independence, SAR binary weights, SAR row-std weights, CAR

```
apr13.m10 <- lm(logapr13 ~ logmar22, data=iowa) # independence
apr13.m1a <- spautolm(logapr13 ~ logmar22, data=iowa, iowa.w1) # SAR, binary
apr13.m1b <- spautolm(logapr13 ~ logmar22, data=iowa, iowa.w2) # SAR, row-std
apr13.m1c <- spautolm(logapr13 ~ logmar22, data=iowa, iowa.w1, family='CAR') #CAR
c(indep=AIC(apr13.m10), SARbin = AIC(apr13.m1a), SARrow=AIC(apr13.m1b), CAR=AIC(apr13.m1c) )
```

```
## indep SARbin SARrow CAR
## 287.1441 251.2744 250.5904 250.4003
```

CAR has the lowest AIC, but SARrow is very close and SARbinary is within 2.

Note: If this were my research, I would check coefficients for the other reasonably well supported models to make sure my conclusions don't change substantially.

What is the association? Will use the CAR model.

Note: Default output includes the slope and se. Have to calculate the confidence interval by hand. Since these are ml estimates, use z quantiles (i.e. assume normality with known variance). `confint()` is not able to extract the appropriate stuff from the `spautolm` object (It should, perhaps in the future).

```
summary(apr13.m1c)
```

```
##
## Call: spautolm(formula = logapr13 ~ logmar22, data = iowa, listw = iowa.w1,
##      family = "CAR")
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -2.059849 -0.519583 -0.052106  0.497481  2.066737
##
## Coefficients:
##              Estimate Std. Error z value Pr(>|z|)
## (Intercept)  0.98043    0.19578   5.0078 5.507e-07
## logmar22     1.09457    0.12763   8.5761 < 2.2e-16
##
## Lambda: 0.18762 LR test value: 38.744 p-value: 4.8325e-10
## Numerical Hessian standard error of lambda: 0.0041638
##
## Log likelihood: -121.2002
## ML residual variance (sigma squared): 0.58427, (sigma: 0.76437)
## Number of observations: 99
## Number of parameters estimated: 4
## AIC: 250.4
```

```
1.0945 + c(-1,1)*1.96*0.1276
```

```
## [1] 0.844404 1.344596
```

Very strong evidence of an association. Estimated slope = 1.09

95% confidence interval is (0.84, 1.34)

Notes: If you used other models, the estimated coefficients are similar: 1.04 for SAR binary weights, 1.07 for SAR row std. If you assumed independence, the coefficient is substantially different (1.55, which is outside the CI for any spatial model).

You can't use row-std weights with the CAR model because the C matrix (connections between areas) has to be symmetric. As of now, the spautolm() function does not check this.

Adding population to the model

```
apr13.m2a <- spautolm(logapr13 ~ logmar22 + log10pop, data=iowa, iowa.w1)
apr13.m2b <- spautolm(logapr13 ~ logmar22 + log10pop, data=iowa, iowa.w2)
apr13.m2c <- spautolm(logapr13 ~ logmar22 + log10pop, data=iowa, iowa.w1, family='CAR')
c(SARbin = AIC(apr13.m2a), SARrow=AIC(apr13.m2b), CAR=AIC(apr13.m2c) )
```

```
## SARbin SARrow CAR
## 217.9360 220.7465 218.1551
```

With a different set of covariates, the best model is now the SAR binary model. Results from either the CAR model initially considered or the SAR binary model were accepted with full credit.

Four ways to assess whether adding log10pop improved the model (and there are others):

- Do a Z test of log10pop coefficient (results in the summary output)
- Compare AIC statistics for the model with and the model without
- Do a likelihood ratio test comparing the two models, since the 1st is nested in the second.
- Compare residual sd for each model and choose the model with the smaller. Note: this is doing model selection using MSE. In linear Gaussian models, this is the same as using adjusted R-squared.

Do Z test then look at AIC values (comparing models with same spatial correlation model. I used the Sar/binary model for the Z test and two CAR models for the AIC's. Results will be very similar for other appropriate models.

```
summary(apr13.m2a)
```

```
##
## Call: spautolm(formula = logapr13 ~ logmar22 + log10pop, data = iowa,
## listw = iowa.w1)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -1.244544 -0.454576  0.012199  0.312549  2.415293
##
## Coefficients:
##              Estimate Std. Error z value Pr(>|z|)
## (Intercept) -5.60206     1.03914  -5.3911 7.004e-08
## logmar22     0.49923     0.13331   3.7449 0.0001804
## log10pop     1.63815     0.25115   6.5225 6.914e-11
##
## Lambda: 0.14587 LR test value: 31.42 p-value: 2.078e-08
## Numerical Hessian standard error of lambda: 0.017599
```

```
##
## Log likelihood: -103.968
## ML residual variance (sigma squared): 0.42088, (sigma: 0.64875)
## Number of observations: 99
## Number of parameters estimated: 5
## AIC: 217.94
```

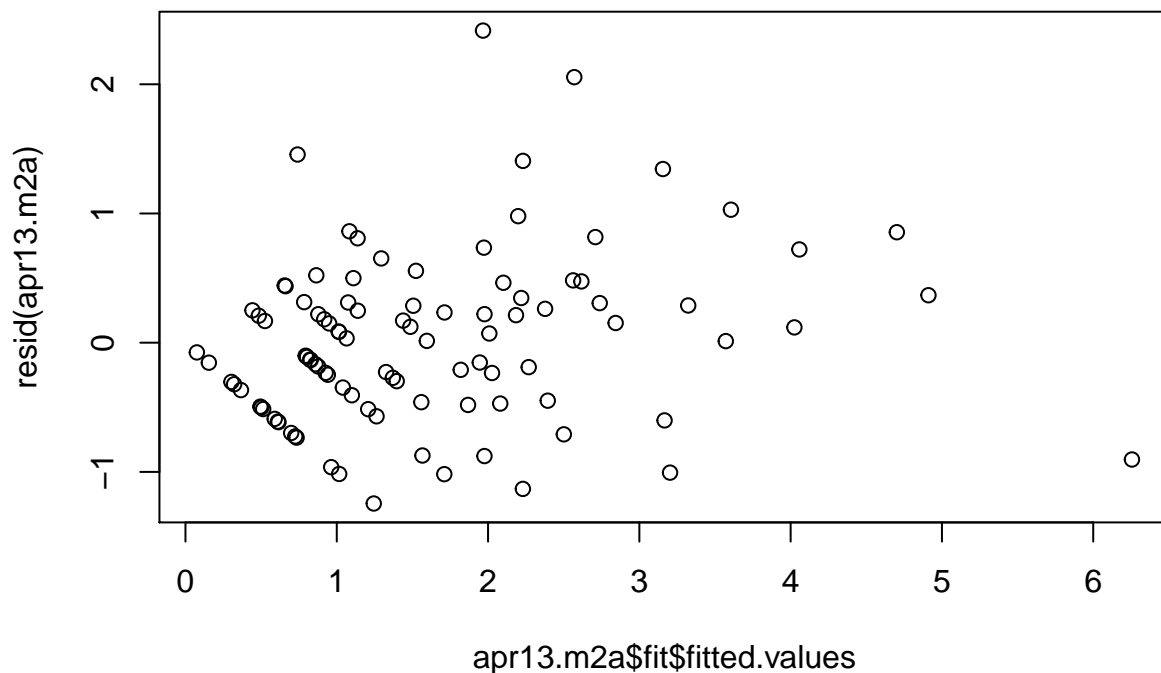
```
c(nopop=AIC(apr13.m1c), withpop=AIC(apr13.m2c))
```

```
##      nopop  withpop
## 250.4003 218.1551
```

Yes adding log10pop improves the model. The p-value for log10pop in a model with logmar22 and logpop is really small. Looking at the AIC statistics for the models with logmar22 and logmar22 + log10pop.

Are assumptions reasonable? Look at a residual vs. predicted value plot.

```
plot(apr13.m2a$fit$fitted.values, resid(apr13.m2a))
```



This is not a great looking residual plot. It isn't really a flat fat sausage. This is not uncommon for data like these because Y is a transformation of counts close to 0. The analysis using a normal distribution for log(# cases + 1) is at best an approximation.

Which counties have unusually large residuals?

```
iowa$COUNTY[resid(apr13.m2a) > log(5)]
```

```
## [1] "Louisa" "Tama"
```

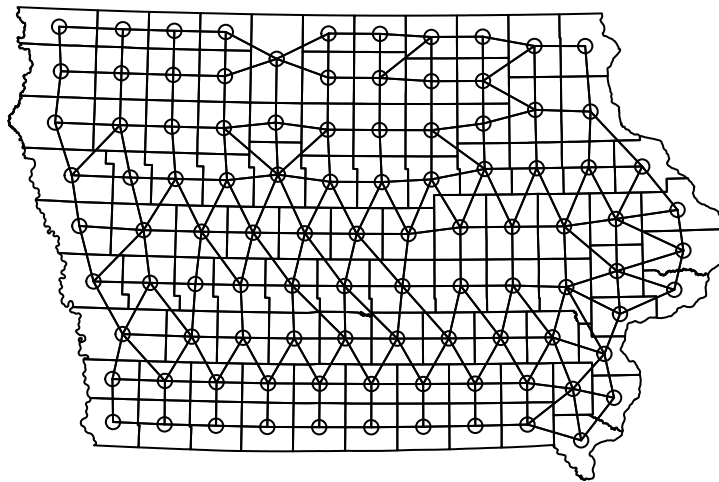
Louisa and Tama counties have observed values more than 5 times their predicted values. That is equivalent

to a residual larger than $\log(5)$.

Note: If you've been following COVID over the last few weeks, you know that both counties are where there have been large clusters of cases, associated with a meatpacking plant in Louisa county and a meatpacking plant and nursing home in Tama county. Other clusters associated with meat-packing plants have since emerged in other counties.

Supplemental material: the residual plot

```
# neighbor plot
plot(iowa)
plot(iowa.nb, coordinates(iowa), add=T )
```



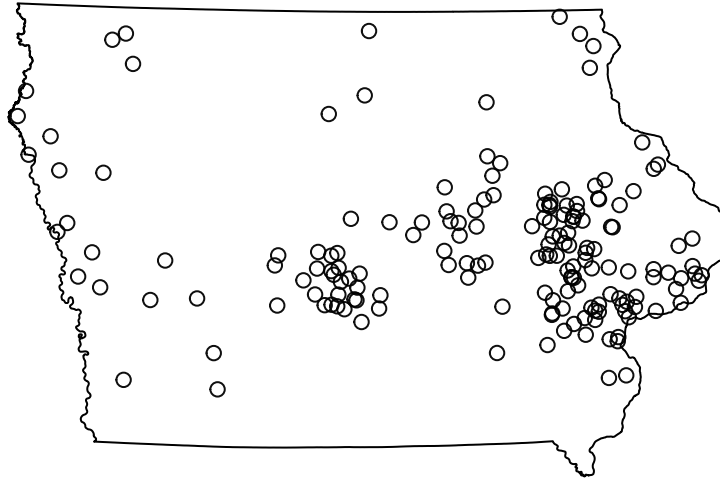
5) Data analysis - version 2, point pattern data

```
load('iowaPPP.Rdata')
load('log10pop.Rdata')
load('logmar22.Rdata')
```

Plot of case locations

```
plot(covid.ppp)
```

covid.ppp

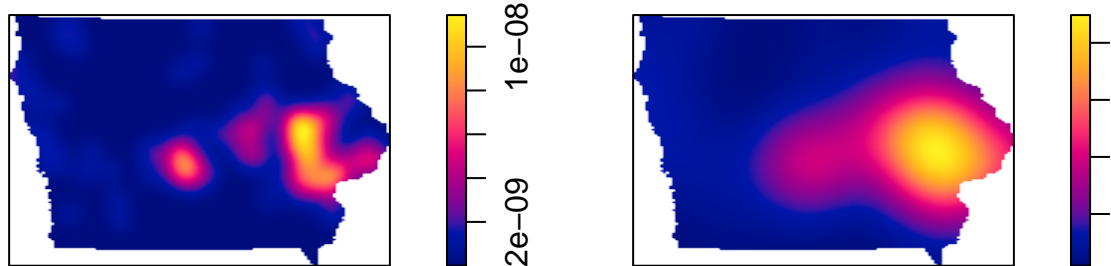


Plot of smoothed intensity: considering both Diggle-Berman and Cronie-van Lieshout estimates of bandwidth.

```
bwd <- bw.diggle(covid.ppp)
bwcvl <- bw.CvL(covid.ppp)
par(mfrow=c(1,2), mar=c(3,3,0,0)+0.2, mgp=c(2,0.8,0))
plot(density(covid.ppp, bwd))
plot(density(covid.ppp, bwcvl))
```


density(covia.ppp, owa)

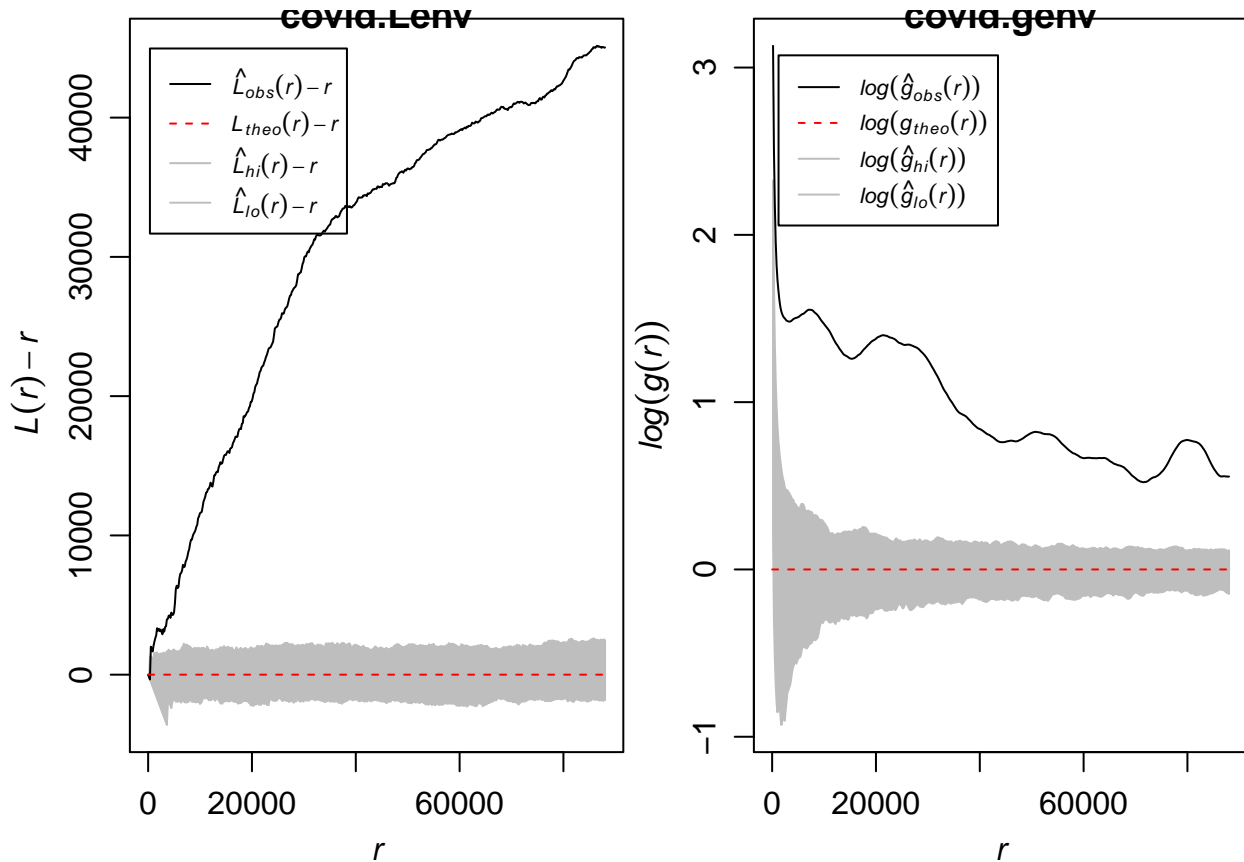
density(covia.ppp, owcvi)



As you would expect the Diggle bandwidth shows finer detail than does the CvL bandwidth. Based on what I know about COVID, I think the Diggle is a better representation, but I accept any reasonable choice.

Clustering? Scale of that clustering?

```
par(mfrow=c(1,2), mar=c(3,3,0,0)+0.2, mgp=c(2,0.8,0))
plot(covid.Lenv, .~r~r, legendargs=list(cex=0.7) )
plot(covid.genv, log(.) ~ r, legendargs=list(cex=0.7) )
```



Very strong evidence of clustering. $L(x)$ is outside its point-wise 95% confidence intervals everywhere (except perhaps very shortist distance) up to 90km. DCLF test has a p-value of 0.005 (1 out of 200). Using $g(x)$, there is evidence of clustering at all distances up to 90km (and probably beyond).

Excess number of cases:

```
covid.K <- Kest(covid.ppp, r=c(0, 20000) )
intensity(covid.ppp) * (covid.K$iso[2] - covid.K$theo[2])
```

```
## [1] 4.339896
```

There are an estimated 4.3 extra cases with 20km of a case.

Notes: This is excess cases, so $\text{intensity} * (\text{K function value} - \text{theoretical expectation})$. You could estimate $K(x)$ at the default collection of distances then find the one closest to 20000, or specify the r vector. For an odd reason, r needs to have at least two values, the first of which is 0.

Association between Mar 22 count and Apr 13 locations: Fit an inhomogeneous Poisson model.

```
covid.m1 <- ppm(covid.ppp ~ logmar22.im)
covid.m1
```

```
## Nonstationary Poisson process
##
## Log intensity: ~logmar22.im
##
## Fitted trend coefficients:
## (Intercept) logmar22.im
## -21.2209544  0.9670526
##
```

```
##           Estimate      S.E.    CI95.lo    CI95.hi Ztest      Zval
## (Intercept) -21.2209544 0.10673966 -21.4301603 -21.011748 *** -198.81039
## logmar22.im  0.9670526 0.06222657  0.8450908  1.089014 ***  15.54083
```

Strong association between the two counts. The estimated slope is 0.967 with a 95% CI of (0.84, 1.09)

Notes: The summary of the fitted model includes the 95% ci by default.

Does including population improve the fit?

```
covid.m2 <- ppm(covid.ppp ~ logmar22.im + log10pop.im)
covid.m2
```

```
## Nonstationary Poisson process
##
## Log intensity: ~logmar22.im + log10pop.im
##
## Fitted trend coefficients:
## (Intercept) logmar22.im log10pop.im
## -26.7081110  0.5608934  1.2787132
##
##           Estimate      S.E.    CI95.lo    CI95.hi Ztest      Zval
## (Intercept) -26.7081110 0.92029236 -28.5118509 -24.9043712 *** -29.021333
## logmar22.im  0.5608934 0.09734143  0.3701077  0.7516791 ***  5.762124
## log10pop.im  1.2787132 0.20981508  0.8674832  1.6899432 ***  6.094477
```

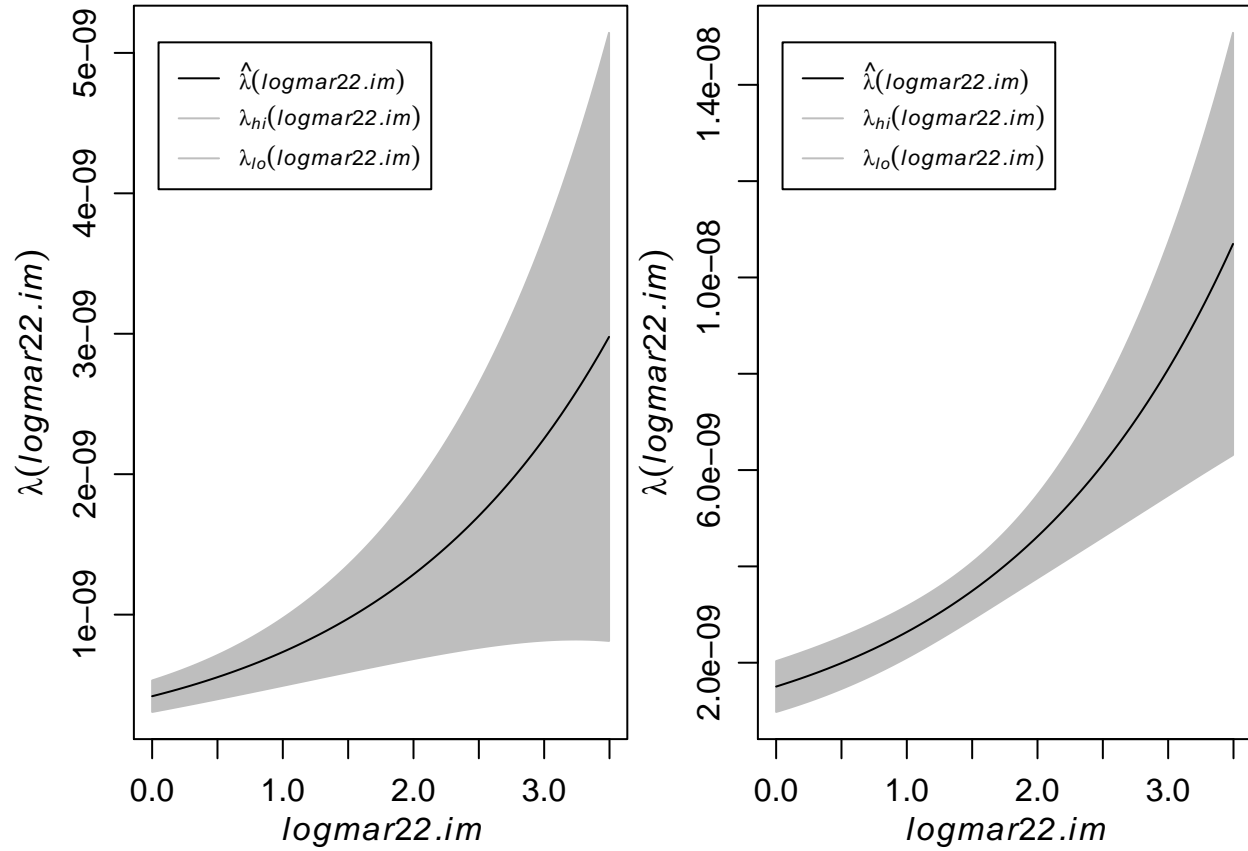
```
c(mar22=AIC(covid.m1), both=AIC(covid.m2))
```

```
## mar22 both
## 7207.775 7173.331
```

Yes, adding log10pop improves the prediction of intensity. You can see this either by Using either the test of log10pop in the model with both variables, or by comparison of AIC values

Relationship between predicted intensity and Mar 22 count for two values of population.

```
par(mfrow=c(1,2), mar=c(3,3,0,0)+0.2, mgp=c(2,0.8,0))
plot( effectfun(covid.m2, 'logmar22.im', log10pop.im=4, se.fit=T),
      main='', legendargs=list(cex=0.7) )
plot( effectfun(covid.m2, 'logmar22.im', log10pop.im=5, se.fit=T),
      main='', legendargs=list(cex=0.7) )
```



Note: The intensity values look really small because they are expected # cases per square m. Values per square km are 10^6 larger, and values per 100 square km are 10^8 larger (e.g. 0.1 to 0.5 for 10000 people and 0.2 to 1.4 for 100000 people).