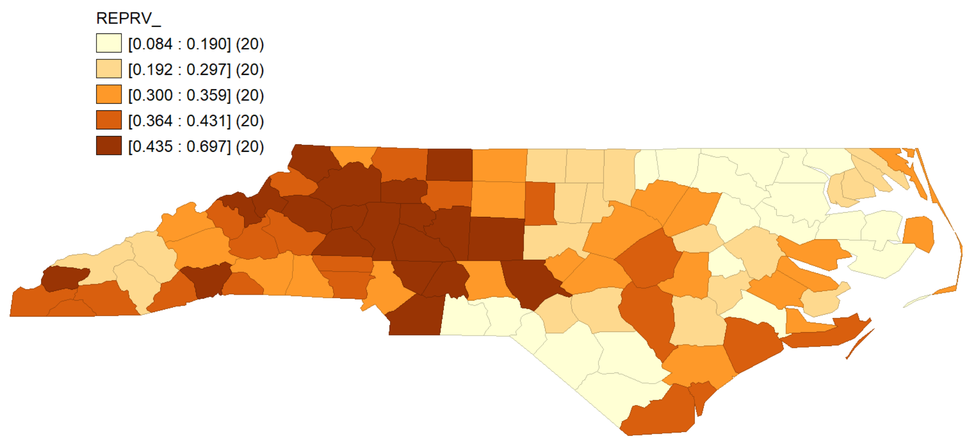
Regression with Spatial Autocorrelation

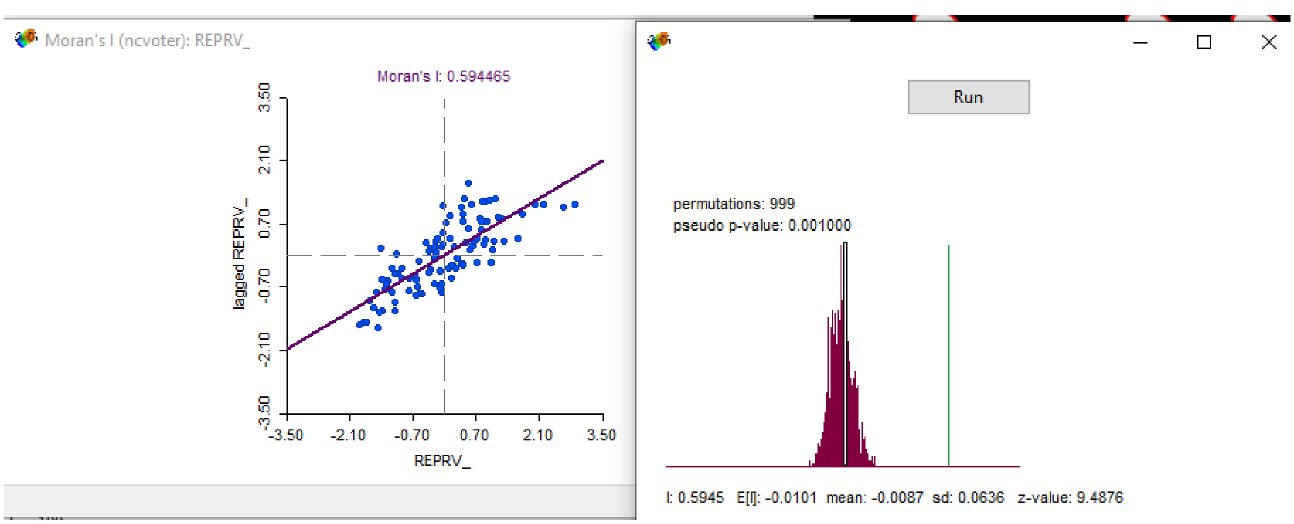
1. In the Datasets folder listed under Module 5 there is a zipped folder "ncvoter". Download and unzip the contents of this folder into your working directory.

2. Load GeoDa and open the ncvoter.shp file. Go to the W (weights) menu item and generate a set of rook contiguity spatial weights for the North Carolina counties.
In the Weights Manager, choose create and choose FIPS for your ID variable.

3. You can map your dependent variable from the map menu on the toolbar. Choose map, select quantile and n=5. The dependent variable is REPRV_ (Republican registered voter). Be careful with the underscore at the end! The variable seems to show strong positive spatial autocorrelation.



4. Check the autocorrelation for the dependent variable. Choose the Space menu option, select Univariate Moran's I, select the dependent variable again and specify your weights. Here I have used Rook contiguity weights.



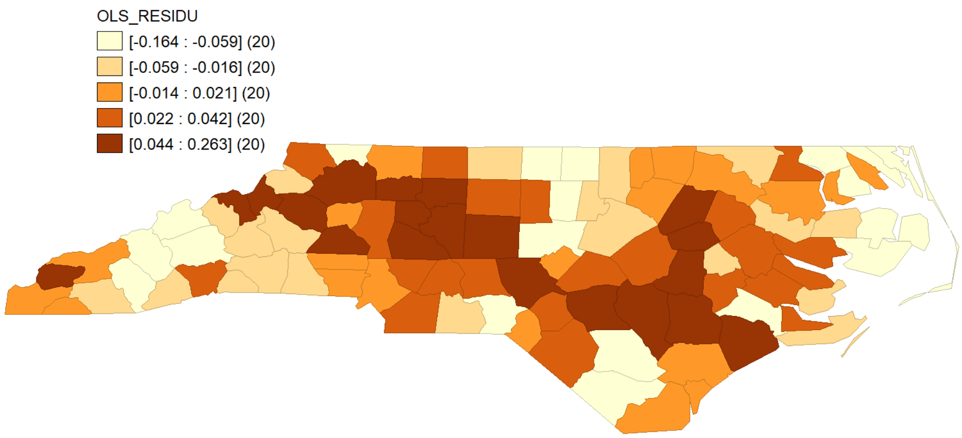
The Z score associated with this plot is over 9, so clear evidence of a non-random pattern. To get the z-score, right click on the Moran plot, choose Randomization and 999 permutations. The resulting plot is a distribution that shows the z-score for the Moran coefficient and how far it is from the value 0. The further from 0, the more likely you are to reject the null hypothesis that the spatial distribution of the values of the variable of interest is random.

5. You can run a standard (non-spatial) regression model in GeoDa. Let's try this. Choose Regression from the menu bar, then select your dependent variable (REPRV_) and then select your two independent variables NHWHITE (non-Hispanic white population share) and MEDIANHI (median household income). So here I have a very crude model that suggests republican voters in North Carolina are more likely to be non-Hispanic white with higher median household incomes. You are not going to run a spatial model for the moment, so you do not need to select spatial weights. Choose the Classic model and click the box that is titled Pred. Val. And Res. This stands for predicted values (fits) and residuals from the regression (the difference between observed and predicted values). Run the model and then close the output and go back to the Regression manager again and click on the save to table button. You only need the residuals, so choose a name for your OLS-residuals (OLS_RESID).

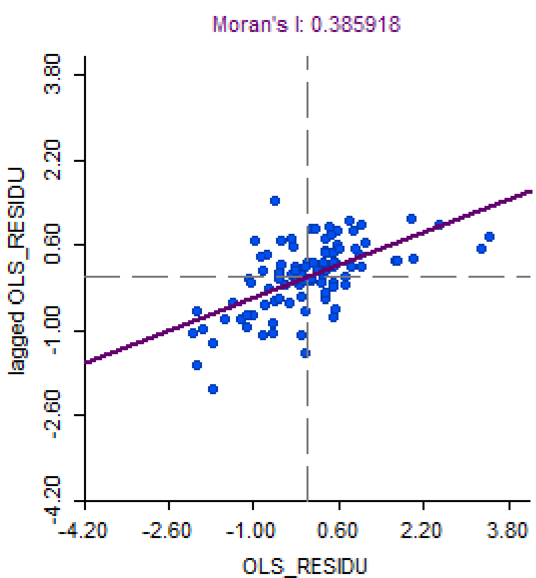
6. Here is the output from the regression. You can check the overall goodness of fit of the regression using the R-squared = 0.67 value. So the two independent variables explain about 67% of the variance in the dependent variable. And you can examine the regression coefficients, the t-test statistics and p-values (Probability) from the regression table. Both independent variables are positively related to the dependent variable and both are statistically significant.

```
REGRESSION
----------
SUMMARY OF OUTPUT: ORDINARY LEAST SQUARES ESTIMATION
Data set            :  ncvoter
Dependent Variable  :      REPRV_  Number of Observations:  100
Mean dependent var  :    0.325838  Number of Variables   :    3
S.D. dependent var  :    0.127914  Degrees of Freedom    :   97

R-squared           :    0.670540  F-statistic           :     98.7106
Adjusted R-squared  :    0.663747  Prob(F-statistic)     :4.10604e-024
Sum squared residual:    0.539058  Log likelihood        :     119.261
Sigma-square        :   0.0055573  Akaike info criterion :    -232.523
S.E. of regression  :   0.0745473  Schwarz criterion     :    -224.707
Sigma-square ML     :  0.00539058
S.E of regression ML:   0.0734206

-----------------------------------------------------------------------------
       Variable      Coefficient      Std.Error    t-Statistic   Probability
-----------------------------------------------------------------------------
          CONSTANT     -0.248731      0.0505245      -4.92297     0.00000
         WHREGVTR_      0.561849      0.0462511       12.1478     0.00000
          MEDIANHI  4.02123e-006   1.36905e-006       2.93724     0.00414
-----------------------------------------------------------------------------

============================== END OF REPORT ================================
```

7. Now, in task 5 above, you saved the residuals from your standard (OLS) regression. Go back to the map menu, choose another quantile map (or something different if you prefer), select the number of classes for your map and plot it.



There is clear evidence of spatial dependence here. Again, back to the Univariate Moran and plot these OLS-residuals



So these are the plots from Module 5 Lecture 4.

8. So now we can think about running some spatial regressions – the spatial lag and spatial error models. Return to the regression manager. Choose your dependent and independent variables again (same as above). Click on the Weights File box and choose your weights file. Then select the spatial lag regression and run it. The top of the output window gives you the standard regression model coefficients. The output is captured below. If you read carefully, you will see that this model

is not estimated with ordinary least squares (OLS) anymore. Indeed, the spatial models cannot be estimated with OLS. Here they are estimated by another technique called Maximum Likelihood Estimation. You don't have to worry about what this is. Note also that as well as partial regression coefficients on the two original independent variables, you also have a new variable W_REPREV_. This is the spatial lag of the dependent variable. Note that the regression coefficient on this variable is significant. At the bottom of the output there are some diagnostics for the spatial lag model. Notice that the Likelihood ratio test at the very bottom gives us a value for a test-statistic that tells us how important the spatial lag form of spatial dependence is in these data. The value is 22.3226.

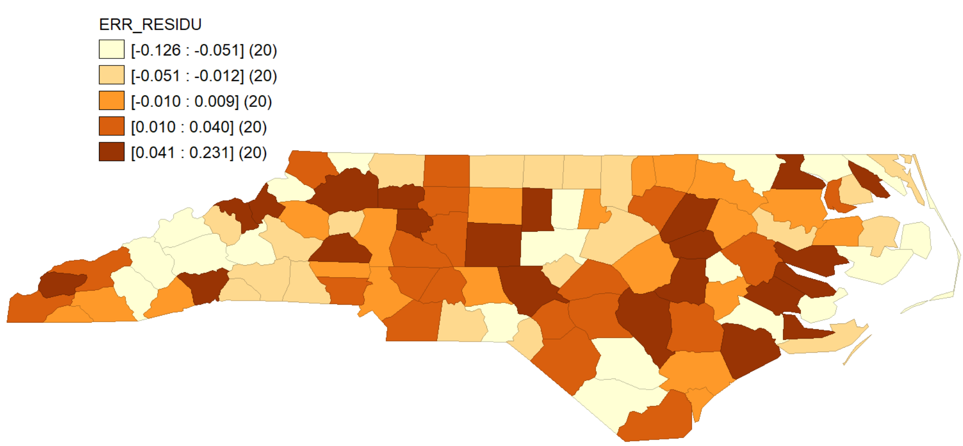
```
>>02/26/22 11:51:17
REGRESSION
----------
SUMMARY OF OUTPUT: SPATIAL LAG MODEL - MAXIMUM LIKELIHOOD ESTIMATION
Data set            : sids
Spatial Weight      : sids_rook
Dependent Variable  :      REPRV_  Number of Observations:  100
Mean dependent var  :    0.325838  Number of Variables   :    4
S.D. dependent var  :    0.127914  Degrees of Freedom    :   96
Lag coeff.   (Rho)  :    0.450193

R-squared           :    0.728093  Log likelihood        :     126.255
Sq. Correlation     : -            Akaike info criterion :     -244.51
Sigma-square        :  0.00444891  Schwarz criterion     :    -234.089
S.E of regression   :   0.0667001

-----------------------------------------------------------------------------
       Variable       Coefficient     Std.Error       z-value    Probability
-----------------------------------------------------------------------------
          W_REPRV_      0.450193      0.0860662        5.23077     0.00000
          CONSTANT     -0.207236      0.0450624       -4.59886     0.00000
           NHWHITE    0.00353052    0.000542464         6.5083     0.00000
          MEDIANHI  3.88696e-006    1.2275e-006        3.16656     0.00154
-----------------------------------------------------------------------------

REGRESSION DIAGNOSTICS
DIAGNOSTICS FOR HETEROSKEDASTICITY
RANDOM COEFFICIENTS
TEST                                     DF      VALUE        PROB
Breusch-Pagan test                       2        17.7793     0.00014

DIAGNOSTICS FOR SPATIAL DEPENDENCE
SPATIAL LAG DEPENDENCE FOR WEIGHT MATRIX : sids_rook
TEST                                     DF      VALUE        PROB
Likelihood Ratio Test                    1        22.3226     0.00000
============================== END OF REPORT ================================
```

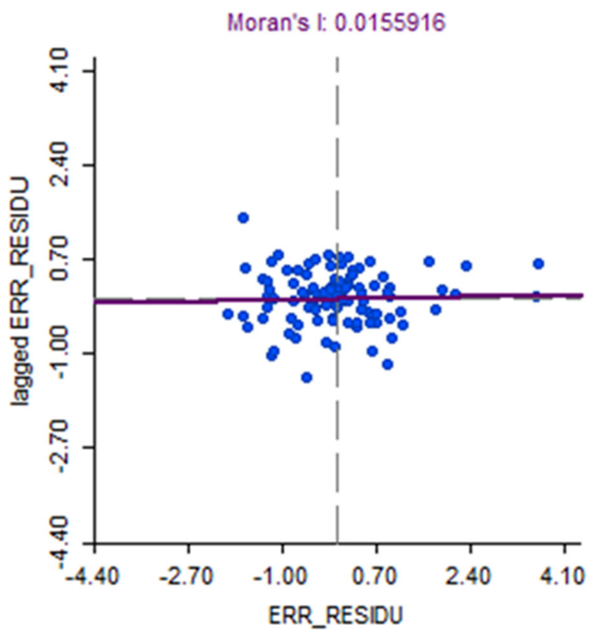
9. Go back to the Regression manager window and run the spatial error model. The output is listed below. Notice again, the model is fit with the Maximum Likelihood Estimator. Instead of the spatial lag, now the extra term in the regression model (LAMBDA) represents the spatial error. Look at the spatial autocorrelation diagnostics and the likelihood ratio test at the bottom of the output again. Notice that the test statistic is 27.6294. This test statistic is a little higher than that for the spatial lag model which suggests the spatial error model is capturing more of the spatial dependence in the data. Thus, you will select to use the spatial error rather than the spatial lag model.

```
REGRESSION
----------
SUMMARY OF OUTPUT: SPATIAL ERROR MODEL - MAXIMUM LIKELIHOOD ESTIMATION

Data set            : sids
Spatial Weight      : sids_rook
Dependent Variable  :      REPRV_  Number of Observations:  100
```

```
Mean dependent var  :    0.325838  Number of Variables   :    3
S.D. dependent var  :    0.127914  Degrees of Freedom    :   97
Lag coeff. (Lambda) :    0.601403

R-squared           :    0.754663  R-squared (BUSE)      : -
Sq. Correlation     : -            Log likelihood        :  128.939738
Sigma-square        :  0.00401417  Akaike info criterion :    -251.879
S.E of regression   :   0.0633575  Schwarz criterion     :    -244.064

-----------------------------------------------------------------------------
       Variable       Coefficient      Std.Error       z-value    Probability
-----------------------------------------------------------------------------
          CONSTANT       -0.18677      0.0622197       -3.00179     0.00268
           NHWHITE     0.00497423    0.000588429         8.4534     0.00000
          MEDIANHI   4.53066e-006   1.62795e-006        2.78304     0.00539
            LAMBDA       0.601403      0.0921991        6.52288     0.00000
-----------------------------------------------------------------------------

REGRESSION DIAGNOSTICS
DIAGNOSTICS FOR HETEROSKEDASTICITY
RANDOM COEFFICIENTS
TEST                                     DF       VALUE         PROB
Breusch-Pagan test                       2         21.8598      0.00002

DIAGNOSTICS FOR SPATIAL DEPENDENCE
SPATIAL ERROR DEPENDENCE FOR WEIGHT MATRIX : sids_rook
TEST                                     DF       VALUE         PROB
Likelihood Ratio Test                    1         27.6924      0.00000
============================== END OF REPORT ================================
```

10. Run the spatial error model again and save the residuals from that model run. I plot the residuals below along with the Univariate Moran for the new residuals from the spatial error model. Again, these reflect the images I showed in lecture. Now it appears that you have removed most of the spatial autocorrelation from your regression model, so you have more faith interpreting the usual coefficients.





PART B – if you really want to do this in R it is quite straightforward…..

1. Fire up RStudio and install the package "spatialreg".

2. Then load the following libraries: spdep, sf, sp & spatialreg

3. Set your current working directory to the folder where the ncvoter.shp file is located.

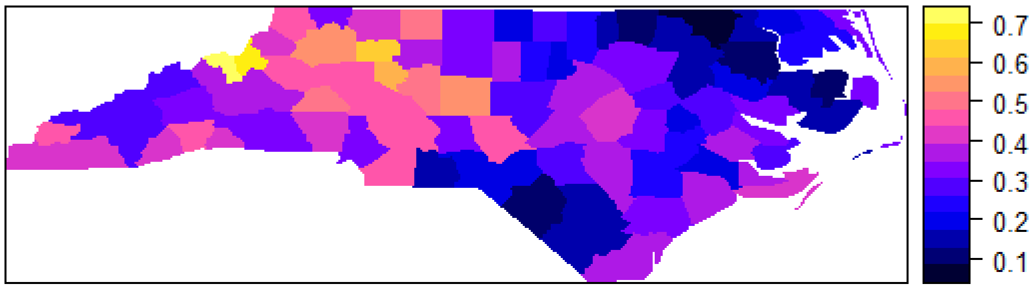
4. Read the shapefile into R
> ncvoter <- st_read("ncvoter.shp")

*Hint >sf_use_s2(FALSE)

And make sure your R object is recognized as spatial

>ncvoter2 <- as(ncvoter, "Spatial")

5. You could load tmap and do a quick plot of your variable, or you could do a quick and dirty spplot (colors are a little weird, I haven't played much with this)
> spplot(ncvoter2, 'REPRV_', col='transparent')



Higher values around the counties shaded yellow. This is consistent with GeoDa.

5. Use spdep to create some spatial weights as before. Let's use rook contiguity again. (zero.policy=T gets around problems of islands or polygons with no neighbors. I don't think there are any here.)

```
> ncwm_r <- poly2nb(ncvoter2, queen=FALSE)
> neighbors <- nb2listw(ncwm_r, zero.policy=T)
```

6. We can generate a quick Moran test of our dependent variable

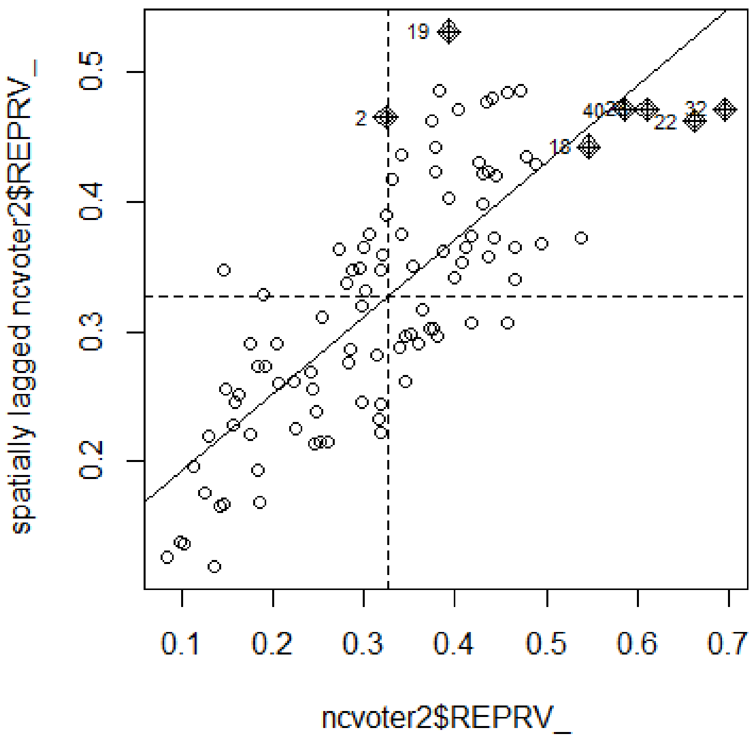
```
> moran.test(ncvoter2$REPRV_, neighbors, zero.policy=T)
```

```
        Moran I test under randomisation

data:  ncvoter2$REPRV_
weights: neighbors

Moran I statistic standard deviate = 9.0383, p-value < 2.2e-16
alternative hypothesis: greater
sample estimates:
Moran I statistic       Expectation          Variance
      0.594464599      -0.010101010       0.004474185
```

This clearly suggests issues of positive spatial autocorrelation.

7. Let's look at the Moran plot

```
> moran.plot(ncvoter2$REPRV_, neighbors, zero.policy=T)
```



8. Now let’s run our OLS regression (non-spatial) again.

> nonspatial = lm(REPRV_ ~ NHWHITE + MEDIANHI, data=ncvoter2)

> summary(nonspatial)

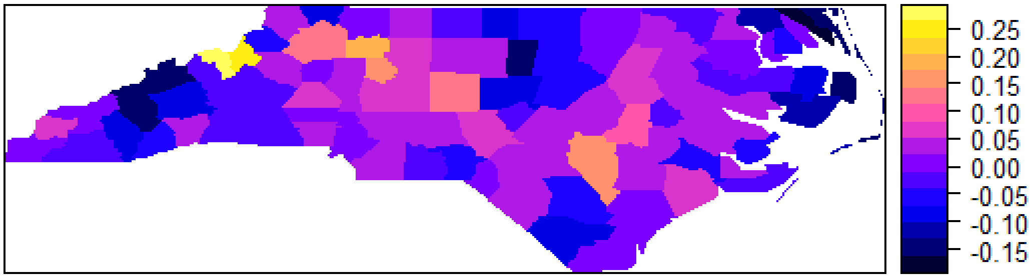
```
Call:
lm(formula = REPRV_ ~ NHWHITE + MEDIANHI, data = ncvoter2)

Residuals:
      Min        1Q    Median        3Q       Max 
-0.163785 -0.045014 -0.000973  0.038320  0.263154 

Coefficients:
              Estimate Std. Error t value Pr(>|t|)    
(Intercept) -2.197e-01  5.191e-02  -4.233 5.25e-05 ***
NHWHITE      5.184e-03  4.582e-04  11.314  < 2e-16 ***
MEDIANHI     5.101e-06  1.404e-06   3.633  0.00045 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.07772 on 97 degrees of freedom
Multiple R-squared:  0.6419,	Adjusted R-squared:  0.6345 
F-statistic: 86.94 on 2 and 97 DF,  p-value: < 2.2e-16
```

9. Residuals are generated automatically. Just get them into the object you want.

> ncvoter2$residuals <- nonspatial$residuals

And then plot

> spplot(ncvoter2, 'residuals', col="transparent")



The larger positive residuals are where they were in the GeoDa plots.

10. So we know that we have some spatial autocorrelation in our residuals, so let's run the spatial lag model

> lag <- lagsarlm(REPRV_ ~ NHWHITE + MEDIANHI, data=ncvoter2, listw=neighbors, zero.policy=T, tol.solve=1e-30)

> summary(lag)

```
Call:lagsarlm(formula = REPRV_ ~ NHWHITE + MEDIANHI, data = ncvoter2,
    listw = neighbors, zero.policy = T, tol.solve = 1e-30)

Residuals:
       Min         1Q     Median         3Q        Max
-0.1157554 -0.0465393 -0.0016921  0.0366461  0.2339880

Type: lag
Coefficients: (asymptotic standard errors)
               Estimate  Std. Error z value  Pr(>|z|)
(Intercept) -2.0724e-01  4.5062e-02 -4.5989 4.248e-06
NHWHITE      3.5305e-03  5.4246e-04  6.5083 7.600e-11
MEDIANHI     3.8870e-06  1.2275e-06  3.1666  0.001543

Rho: 0.45019, LR test value: 22.323, p-value: 2.3048e-06
Asymptotic standard error: 0.086066
    z-value: 5.2308, p-value: 1.688e-07
Wald statistic: 27.361, p-value: 1.688e-07

Log likelihood: 126.2548 for lag model
ML residual variance (sigma squared): 0.0044489, (sigma: 0.0667)
Number of observations: 100
Number of parameters estimated: 5
AIC: -242.51, (AIC for lm: -222.19)
LM test for residual autocorrelation
test value: 6.7849, p-value: 0.0091931
```

The output can be interpreted the same way as dicussed for GeoDa above.

11. To generate the spatial error model

```
> error <- errorsarlm(REPRV_ ~ NHWHITE + MEDIANHI, data=ncvoter2, listw=neighbors, zero.policy=T, tol.solve=1e-30)
> summary(error)
```

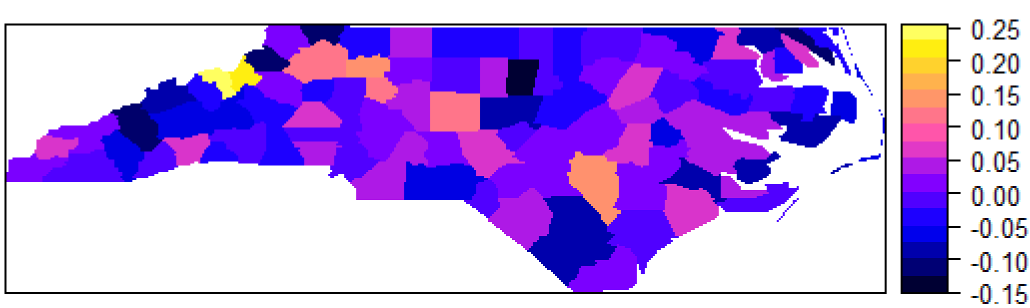
```
Call:errorsarlm(formula = REPRV_ ~ NHWHITE + MEDIANHI, data = ncvoter2,
    listw = neighbors, zero.policy = T, tol.solve = 1e-30)

Residuals:
       Min         1Q     Median         3Q        Max
-0.1257016 -0.0371680 -0.0019118  0.0288118  0.2310203

Type: error
Coefficients: (asymptotic standard errors)
               Estimate  Std. Error z value  Pr(>|z|)
(Intercept) -1.8677e-01  6.2220e-02 -3.0018  0.002684
NHWHITE      4.9742e-03  5.8843e-04  8.4534 < 2.2e-16
MEDIANHI     4.5307e-06  1.6279e-06  2.7830  0.005385

Lambda: 0.6014, LR test value: 27.692, p-value: 1.4222e-07
Asymptotic standard error: 0.092199
    z-value: 6.5229, p-value: 6.8971e-11
Wald statistic: 42.548, p-value: 6.8971e-11

Log likelihood: 128.9397 for error model
ML residual variance (sigma squared): 0.0040142, (sigma: 0.063357)
Number of observations: 100
Number of parameters estimated: 5
AIC: -247.88, (AIC for lm: -222.19)
```

12. If you want the residuals from the spatial error model

```
> ncvoter2$residuals_err <-error$residuals
```

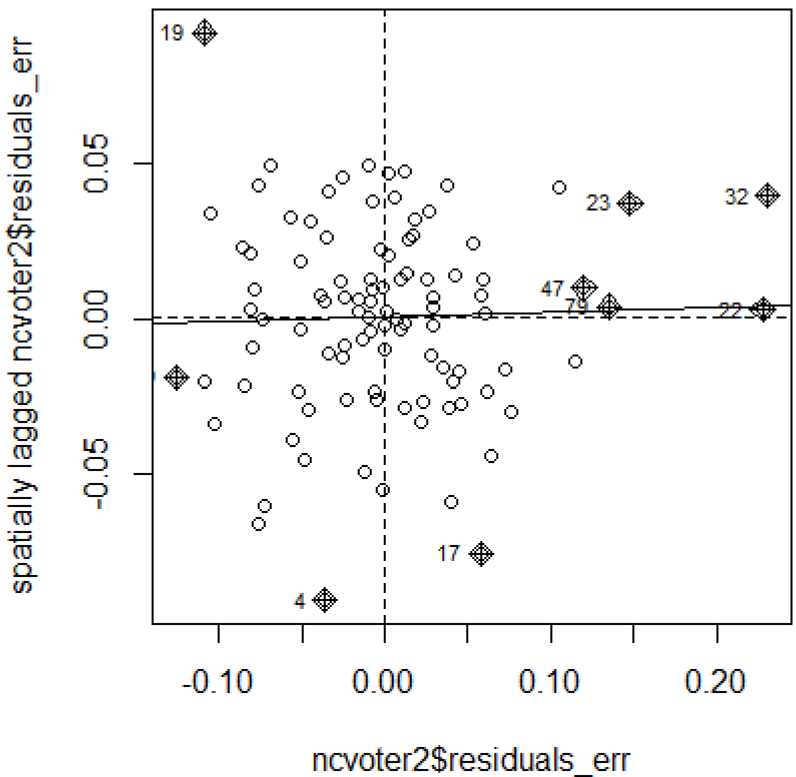
And the plot

```
> spplot(ncvoter2, 'residuals_err', col="transparent")
```



13. Finally, here is the Moran plot and test statistic for these residuals

> moran.plot(ncvoter2$residuals_err, neighbors, zero.policy=T)



> moran.test(ncvoter2$residuals_err, neighbors, zero.policy=T)

Moran I test under randomisation

```
data:  ncvoter2$residuals_err
weights: neighbors

Moran I statistic standard deviate = 0.38862, p-value = 0.3488
alternative hypothesis: greater
sample estimates:
Moran I statistic       Expectation          Variance
      0.015591620      -0.010101010       0.004370968
```