Spatial data analysis (2)

Yue Jiang

Duke University

A disclaimer

The following material was used during a live lecture. Without the accompanying oral comments and discussion, the text is incomplete as a record of the presentation. A full recording may be found via Zoom on the course Sakai site.

Motivating example

```
##          name uninsured   mhhi rural
## 1   ALAMANCE        18 50.480  28.6
## 2  ALEXANDER        17 49.138  72.8
## 3  ALLEGHANY        22 39.735 100.0
## 4      ANSON        16 38.023  78.5
## 5       ASHE        19 41.864  84.9
## 6      AVERY        24 41.701  88.8
## 7   BEAUFORT        17 46.411  65.6
## 8     BERTIE        16 35.433  83.2
## 9     BLADEN        20 36.976  91.2
## 10 BRUNSWICK        16 60.163  43.0
## 11  BUNCOMBE        16 53.960  24.1
## 12     BURKE        18 44.946  42.7
## 13  CABARRUS        13 69.297  19.3
## 14  CALDWELL        17 43.328  34.4
## 15    CAMDEN        14 65.955  99.5
```

Is there an association between the adult uninsured % in each county and the rurality of a county, adjusting for median household income?

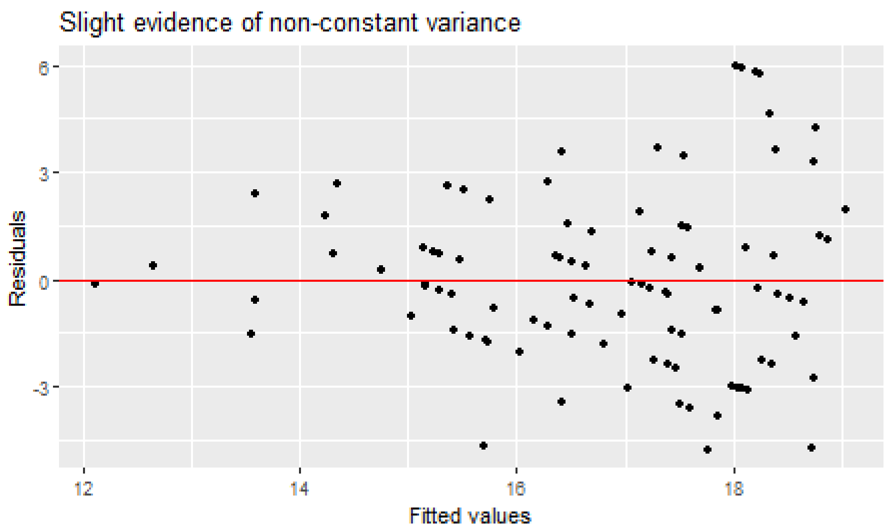
Linear regression model

```
m1 <- lm(uninsured ~ rural + mhhi, data = nc)
summary(m1)
```

```
##
## Call:
## lm(formula = uninsured ~ rural + mhhi, data = nc)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -4.7702 -1.5317 -0.1947  1.2996  5.9702
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 20.18807    1.69199  11.932  < 2e-16
## rural        0.02630    0.00942   2.792 0.006316
## mhhi        -0.10264    0.02761  -3.718 0.000336
##
## Residual standard error: 2.384 on 97 degrees of freedom
## Multiple R-squared:  0.2787,    Adjusted R-squared:  0.2638
## F-statistic: 18.74 on 2 and 97 DF,  p-value: 1.316e-07
```

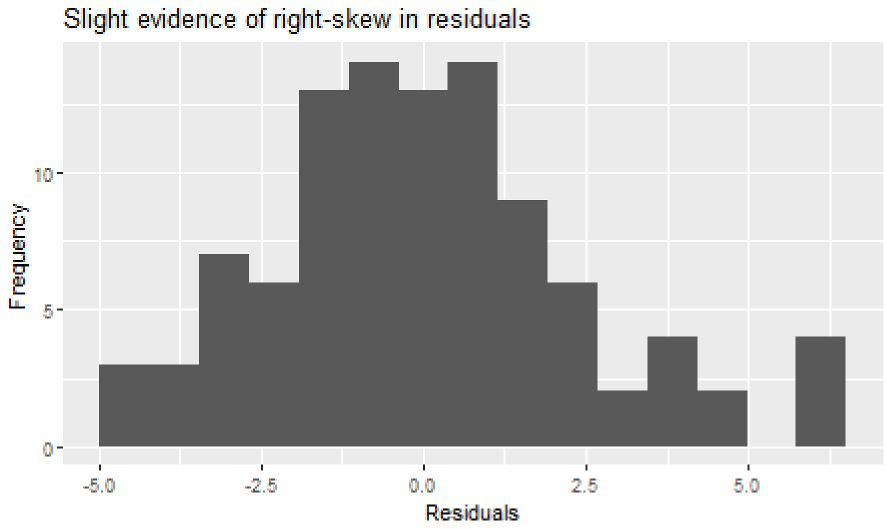
Linear regression model

```
temp <- tibble(res = m1$residuals,
               fitted = m1$fitted.values)
ggplot(data = temp, aes(x = fitted, y = res)) +
  geom_point() +
  labs(x = "Fitted values", y = "Residuals",
       title = "Slight evidence of non-constant variance") +
  geom_hline(yintercept = 0, color = "red")
```

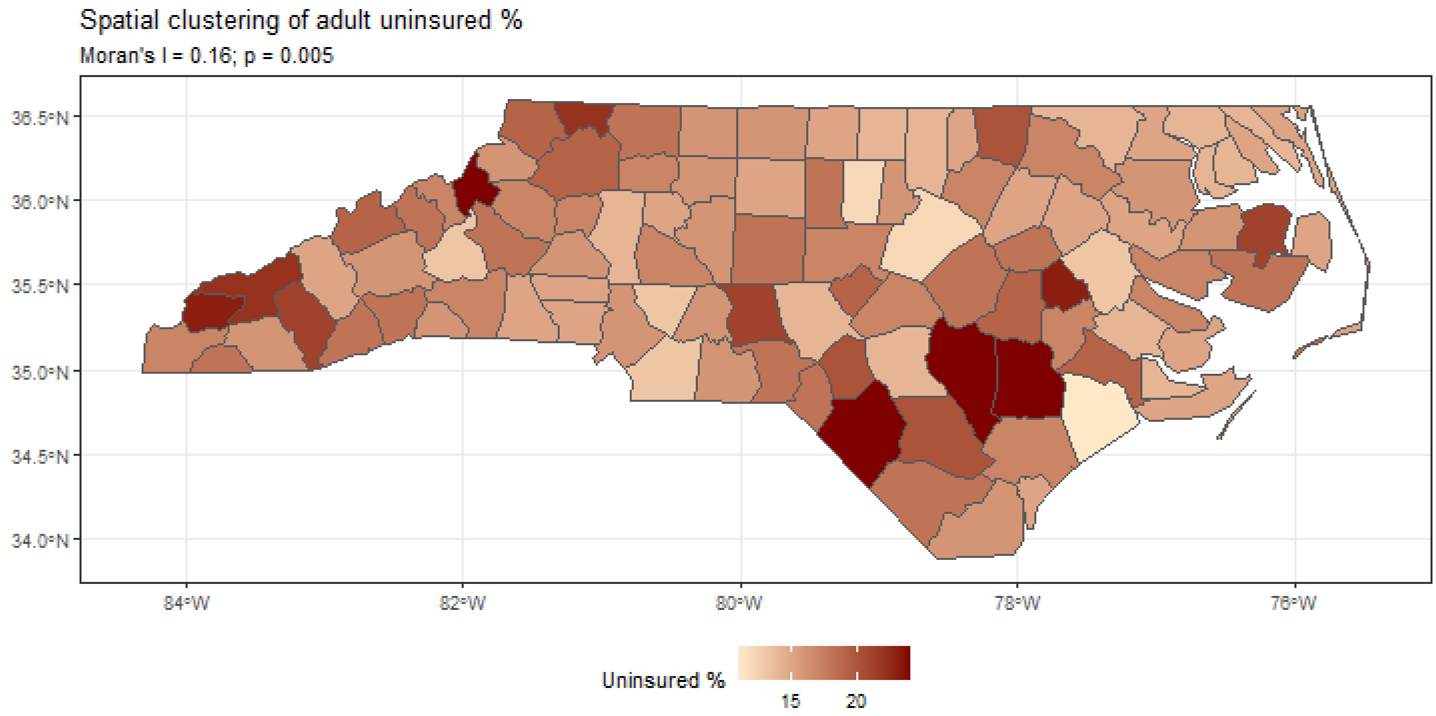


Linear regression model

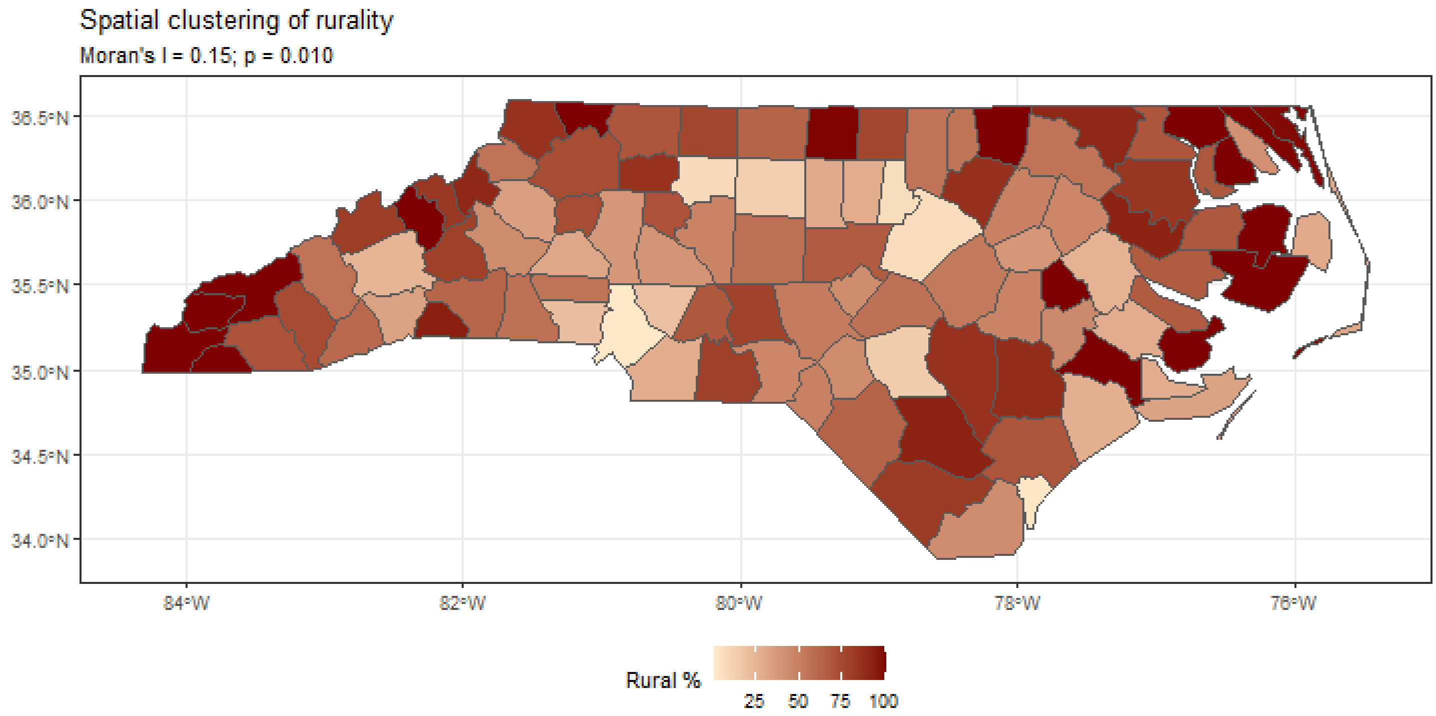
```
ggplot(data = temp, aes(x = res)) +
  geom_histogram(bins = 15) +
  labs(x = "Residuals", y = "Frequency",
       title = "Slight evidence of right-skew in residuals")
```



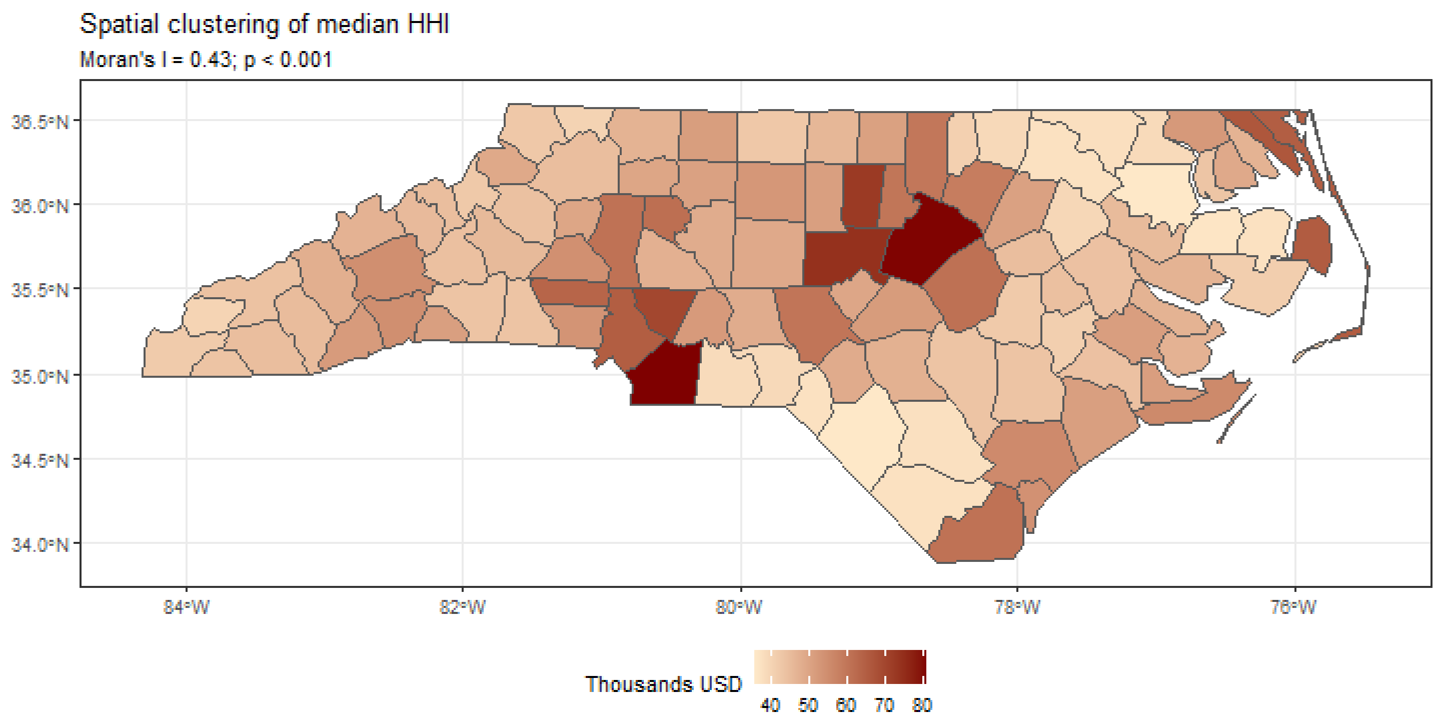
Exploratory data analysis



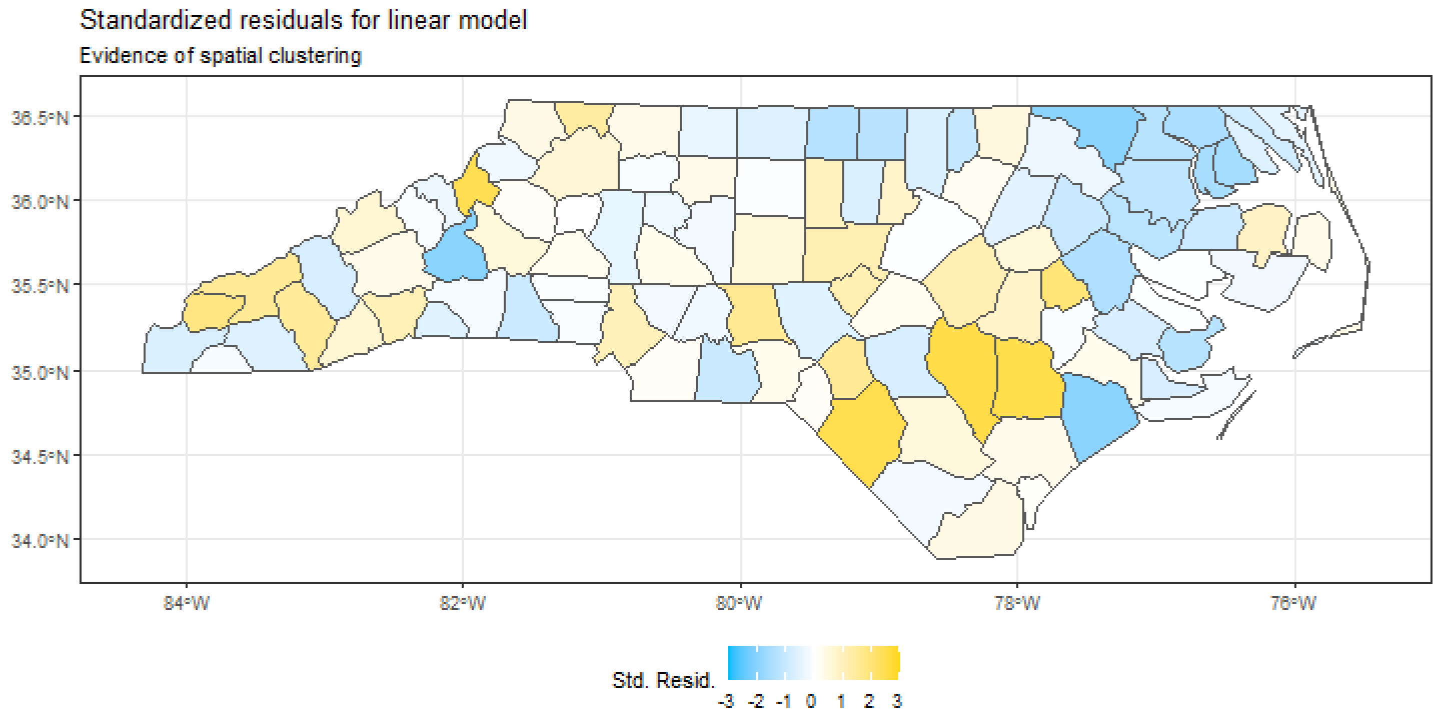
Exploratory data analysis



Exploratory data analysis



Independence assumption of model is violated!



Moran's I for model residuals

```
nc_sp <- as(nc, "Spatial")
sp_wts <- poly2nb(nc_sp)
sp_wts_mat <- nb2mat(sp_wts, style='W')
sp_wts_list <- mat2listw(sp_wts_mat, style='W')
```

```
lm.morantest(m1, sp_wts_list, alternative = "two.sided")
```

Why can't we simply calculate Moran's I on the residuals themselves like we did previously (note the different function)?

Moran's I for model residuals

```r
lm.morantest(m1, sp_wts_list, alternative = "two.sided")
```

```
##
##  Global Moran I for regression residuals
##
## data:
## model: lm(formula = uninsured ~ mhhi + rural, data = nc)
## weights: sp_wts_list
##
## Moran I statistic standard deviate = 2.9452, p-value = 0.003227
## alternative hypothesis: two.sided
## sample estimates:
## Observed Moran I      Expectation         Variance
##      0.175500899     -0.015183947      0.004191698
```

What might we conclude? Is there evidence for spatial clustering or dispersion among the residuals in our model? What are the consequences?

Spatial regression models

There are two main ways of dealing with spatial dependence in regression models: spatial error models, and spatial lag models*

Spatial error models: assume that the error terms are correlated; however, independence may still be reasonable - perhaps the residuals are correlated due to an unmeasured confounding variable (and were to measure them, no longer have issues with spatial dependency).

Spatial lag models: independence of observations is violated due to some underlying spatial process - perhaps the *outcome itself* is associated with the outcome in neighboring spatial areas (and must be handled by incorporating spatial lag as a predictor).

*(let's not get into CAR vs. SAR models for now...)

Spatial regression models

Spatial error model:

$$Y = \mathbf{X}\boldsymbol{\beta} + \lambda\mathbf{W}\mathbf{u} + \boldsymbol{\epsilon}$$

Spatial lag model:

$$Y = \rho\mathbf{W}Y + \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\epsilon}$$

We can use Lagrange multiplier tests for *specific* alternatives by comparing each of these models to a constrained model (where λ or ρ equal 0, respectively).

Tests for spatial dependence

```r
lm.LMtests(m1, sp_wts_list, test = c("LMerr", "LMlag"))
```

```
##
##     Lagrange multiplier diagnostics for spatial dependence
##
## data:
## model: lm(formula = uninsured ~ mhhi + rural, data = nc)
## weights: sp_wts_list
##
## LMerr = 6.8982, df = 1, p-value = 0.008628
##
##
##     Lagrange multiplier diagnostics for spatial dependence
##
## data:
## model: lm(formula = uninsured ~ mhhi + rural, data = nc)
## weights: sp_wts_list
##
## LMlag = 4.0042, df = 1, p-value = 0.04539
```

Tests for spatial dependence

There was evidence against both null hypotheses. Unfortunately, the Lagrange multiplier tests also have some power against the other alternative, and so if both are significant, we still don't have a good idea regarding which type(s) of spatial dependence might be present.

We can use robust tests (Anselin et al. 1996) to account for this consideration.

Tests for spatial dependence

```
lm.LMtests(m1, sp_wts_list, test = c("RLMerr", "RLMlag"))
```

```
##
##      Lagrange multiplier diagnostics for spatial dependence
##
## data:
## model: lm(formula = uninsured ~ mhhi + rural, data = nc)
## weights: sp_wts_list
##
## RLMerr = 4.1472, df = 1, p-value = 0.0417
##
##
##      Lagrange multiplier diagnostics for spatial dependence
##
## data:
## model: lm(formula = uninsured ~ mhhi + rural, data = nc)
## weights: sp_wts_list
##
## RLMlag = 1.2532, df = 1, p-value = 0.2629
```

Aside: SARMA models

$$Y = \rho \mathbf{W} Y + \mathbf{X} \boldsymbol{\beta} + \lambda \mathbf{W} \mathbf{u} + \boldsymbol{\epsilon}$$

Presence of *both* spatial error dependency and spatial lag

```
lm.LMtests(m1, sp_wts_list, test = c("SARMA"))
```

```
##
##      Lagrange multiplier diagnostics for spatial dependence
##
## data:
## model: lm(formula = uninsured ~ mhhi + rural, data = nc)
## weights: sp_wts_list
##
## SARMA = 8.1514, df = 2, p-value = 0.01698
```

Fitting a spatial lag model

$$Y = \rho \mathbf{W} Y + \mathbf{X} \boldsymbol{\beta} + \boldsymbol{\epsilon}$$

```
m2 <- lagsarlm(uninsured ~ mhhi + rural, data = nc,
               listw = sp_wts_list)
```

Remember, spatial lag suggests that the value of the response variable in one area might *depend* on the value of the response(s) of its neighbor(s), *beyond* other potentially unaccounted-for confounders. In these models, we assume that neither the outcomes of the observations are independent, *nor* the errors are independent.

Tests for spatial dependence should not be the only criterion by which you decide what type of spatial model to fit!

Fitting a spatial lag model

```
summary(m2)
```

```
Call:lagsarlm(formula = uninsured ~ mhhi + rural, data = nc, listw = sp_wts_list)

Residuals:
      Min        1Q    Median        3Q       Max
-5.198455 -1.704736 -0.098604  1.279822  5.975005

Type: lag
Coefficients: (asymptotic standard errors)
             Estimate Std. Error z value  Pr(>|z|)
(Intercept) 15.0629131  2.8567001  5.2728 1.343e-07
mhhi        -0.0858604  0.0274800 -3.1245  0.001781
rural        0.0274359  0.0090915  3.0178  0.002546

Rho: 0.25347, LR test value: 3.7863, p-value: 0.051674
Asymptotic standard error: 0.12268
    z-value: 2.066, p-value: 0.038824
Wald statistic: 4.2685, p-value: 0.038824

Log likelihood: -225.3369 for lag model
ML residual variance (sigma squared): 5.2295, (sigma: 2.2868)
Number of observations: 100
Number of parameters estimated: 5
AIC: 460.67, (AIC for lm: 462.46)
LM test for residual autocorrelation
test value: 1.8148, p-value: 0.17794
```

Interpreting a spatial lag model

Can we say that on average, for each additional \$1,000 increase in median household income in a county, we expect to see a decrease of 8.6 percentage points in the adult uninsured population (holding rurality constant)?

No!

Interpreting a spatial lag model

Median household income and rurality in Durham county are associated with the uninsured rate in Durham county.

However, the uninsured rates of neighboring counties are also associated with the uninsured rate in Durham county!

Even worse, the median household incomes and rurality of neighboring counties are associated with the uninsured rate in their respective counties as well!

...and so on.

In short, the covariate effects depend on both the **direct effect** in the associated spatial unit as well as the **indirect effect** due to spatial lag from its neighboring units.

Interpreting a spatial lag model

```r
sp_wts_sparce <- as(sp_wts_list, "CsparseMatrix")
traces <- trW(sp_wts_sparce, type="MC")
m2_decomp <- impacts(m2, tr = traces, R = 1000)
m2_decomp
```

```
## Impact measures (lag, trace):
##            Direct     Indirect       Total
## mhhi  -0.08716950 -0.027843401 -0.11501290
## rural  0.02785422  0.008897107  0.03675133
```

Interpreting a spatial lag model

```
summary(m2_decomp)$direct_sum
```

```
##
## Iterations = 1:1000
## Thinning interval = 1
## Number of chains = 1
## Sample size per chain = 1000
##
## 1. Empirical mean and standard deviation for each variable,
##    plus standard error of the mean:
##
##           Mean       SD  Naive SE Time-series SE
## mhhi  -0.08898 0.027708 0.0008762      0.0008762
## rural  0.02768 0.008887 0.0002810      0.0002810
##
## 2. Quantiles for each variable:
##
##            2.5%      25%      50%      75%    97.5%
## mhhi  -0.143033 -0.10779 -0.08840 -0.06969 -0.03617
## rural  0.009944  0.02139  0.02751  0.03353  0.04553
```

Interpreting a spatial lag model

```
summary(m2_decomp)$indirect_sum
```

```
##
## Iterations = 1:1000
## Thinning interval = 1
## Number of chains = 1
## Sample size per chain = 1000
##
## 1. Empirical mean and standard deviation for each variable,
##    plus standard error of the mean:
##
##            Mean       SD  Naive SE Time-series SE
## mhhi  -0.028787 0.020104 0.0006357      0.0006357
## rural  0.009259 0.006955 0.0002199      0.0002199
##
## 2. Quantiles for each variable:
##
##             2.5%       25%       50%      75%     97.5%
## mhhi  -7.593e-02 -0.038968 -0.025986 -0.01508 0.0003603
## rural -8.205e-05  0.004615  0.008046  0.01226 0.0251918
```

Interpreting a spatial lag model

```
summary(m2_decomp)$total_sum
```

```
##
## Iterations = 1:1000
## Thinning interval = 1
## Number of chains = 1
## Sample size per chain = 1000
##
## 1. Empirical mean and standard deviation for each variable,
##    plus standard error of the mean:
##
##           Mean      SD  Naive SE Time-series SE
## mhhi  -0.11776 0.03809 0.0012044      0.0012044
## rural  0.03694 0.01331 0.0004208      0.0004208
##
## 2. Quantiles for each variable:
##
##           2.5%      25%      50%      75%    97.5%
## mhhi  -0.19585 -0.14100 -0.11627 -0.09296 -0.04873
## rural  0.01318  0.02744  0.03592  0.04490  0.06714
```

Fitting a spatial error model

$$Y = \mathbf{X}\beta + \lambda\mathbf{W}\mathbf{u} + \boldsymbol{\epsilon}$$

```
m3 <- errorsarlm(uninsured ~ mhhi + rural, data = nc,
                 listw = sp_wts_list)
```

Remember, spatial error models suggest that the spatial dependency comes through the error term only, and estimates variables treating spatial dependence as a nuisance parameter. In these models, we still assume that the outcomes of the observations are independent, but we do *not* need to assume that the errors are independent.

Once again, tests for spatial dependence should not be the only criterion by which you decide what type of spatial model to fit!

Fitting a spatial error model

```
summary(m3)
```

```
##
## Call:
## errorsarlm(formula = uninsured ~ mhhi + rural, data = nc, listw = sp_wt
##
## Residuals:
##       Min        1Q    Median        3Q       Max
## -5.18649 -1.55960 -0.13406  1.18661  5.93112
##
## Type: error
## Coefficients: (asymptotic standard errors)
##               Estimate Std. Error z value  Pr(>|z|)
## (Intercept) 19.5427226  1.6922752 11.5482 < 2.2e-16
## mhhi        -0.0966632  0.0287121 -3.3666 0.0007609
## rural        0.0321375  0.0088566  3.6286 0.0002849
##
## Lambda: 0.333, LR test value: 6.2254, p-value: 0.012593
## Asymptotic standard error: 0.12598
##     z-value: 2.6434, p-value: 0.0082083
## Wald statistic: 6.9874, p-value: 0.0082083
##
## Log likelihood: -224.1174 for error model
```

What about generalized linear models?

...come see me in office hours.

Essentially, we include specific eigenvectors of the spatial weight matrix as predictors in the model of interest. See `R` function documentation for the appropriate function here (speaking of which, the documentation for the `spdep` pacakge is excellent!).

Resources and references

The Center for Spatial Data Science at the University of Chicago

Additional resources (for R) which may be helpful are available here