Spatial Regression Models for Geog 210B Winter 2018

Kostas Goulias

2/19/2018

This is the laboratory document we will use for spatial regression. We will follow the usual process of reading and selecting the regional data we used last week (Riverside County).

We do this four parts:

1. We estimate Ordinary Least Squares models with and without variables that classify the type of Census block group (BG) in center, suburb, exurb, and rural.
2. We define the neighborhood of each BG using queen contiguity with row standardized weights and check for spatial (auto)correlation.
3. We define a few different types of spatial regression models and perform different tests to check if they are a reasonable description of the data we have. To do this we follow the Anselin-Rey flow chart reviewed in lecture notes.
4. We compare different models and discuss in class.

Preliminary house keeping tasks

First make sure we have these libraries. If others are needed will be added later.

```
library(spdep)

## Warning: package 'spdep' was built under R version 3.4.3

## Loading required package: sp

## Loading required package: Matrix

## Loading required package: spData

## Warning: package 'spData' was built under R version 3.4.3

## To access larger datasets in this package, install the spDataLarge
## package with: `install.packages('spDataLarge')`

library(maptools)

## Checking rgeos availability: TRUE

library(RColorBrewer)
library(stargazer)

## Warning: package 'stargazer' was built under R version 3.4.3
```

```
##
## Please cite as:

##  Hlavac, Marek (2018). stargazer: Well-Formatted Regression and Summary St
atistics Tables.

##  R package version 5.2.1. https://CRAN.R-project.org/package=stargazer
```

This is a local requirement that I used for my laptop. You modify your accordingly.

```
setwd("~/Documents/COURSES UCSB/Course Winter 2018/California")
CA.poly <- readShapePoly('LPA_Pop_Char_bg.shp')

## Warning: use rgdal::readOGR or sf::st_read
```

Define indicator variables to use later - BG classification

```
CA.poly@data$center = CA.poly@data$LPAgrp == 4
CA.poly@data$suburb = CA.poly@data$LPAgrp == 3
CA.poly@data$exurb = CA.poly@data$LPAgrp == 2
CA.poly@data$rural = CA.poly@data$LPAgrp == 1
CA.poly@data$none = CA.poly@data$LPAgrp == 0
```

Select the Riverside county data

```
YCOUNTY <- CA.poly[CA.poly@data$countyname== c("Riverside"), ]
```

Create the dependent variable we use (Y= Vehicle Miles per person in each block group) Replace NA (one BG) with zero and check we have 1030 BG with data on the variable we will use

```
YCOUNTY@data$VMTpr = YCOUNTY@data$VMT/YCOUNTY@data$n_pr
YCOUNTY@data$VMTpr[is.na(YCOUNTY@data$VMTpr)] <- 0
```

You can also run summary (YCOUNTY@data) to check the contents

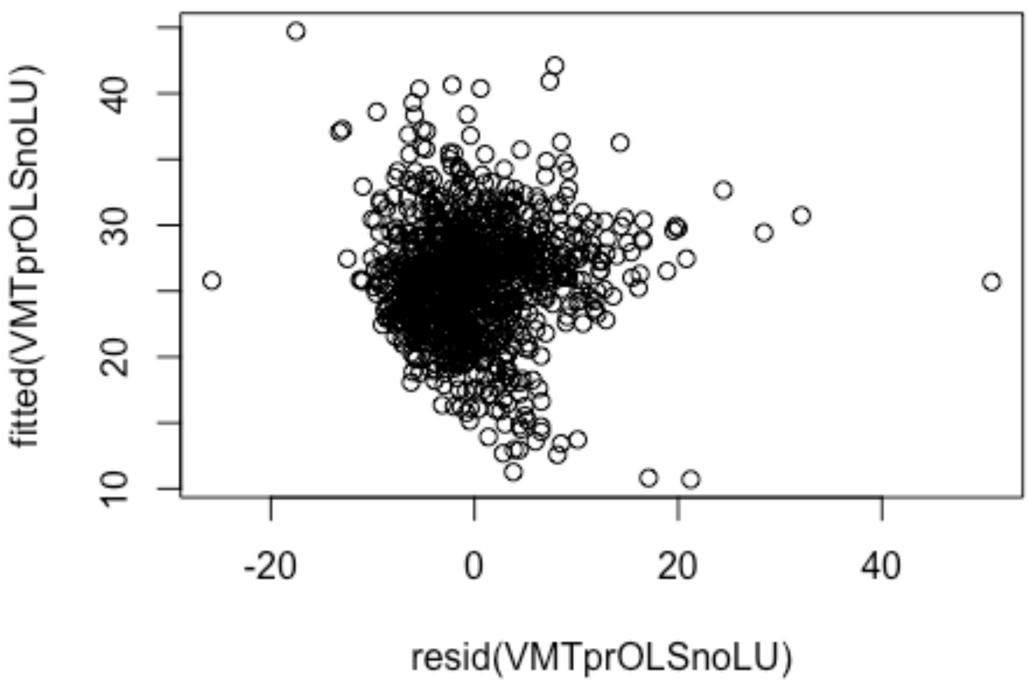
Part 1 Ordinary Least Squares Regression of Two Variants one without BG classification (noLU) and a second with BG classification (yesLU)

BG = US Census Block Group

```
VMTprOLSnoLU<-lm(VMTpr~HHVEH0+HHVEH1+HHVEH2+HHVEH3+HHVEH4+HHVEH5+HHVEH6+HHAGE
7, data=YCOUNTY@data)
summary(VMTprOLSnoLU)

##
## Call:
## lm(formula = VMTpr ~ HHVEH0 + HHVEH1 + HHVEH2 + HHVEH3 + HHVEH4 +
##     HHVEH5 + HHVEH6 + HHAGE7, data = YCOUNTY@data)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -25.791  -3.859  -0.813   3.092  50.775
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) 25.791074   0.430696  59.882  < 2e-16 ***
## HHVEH0      -0.044313   0.009743  -4.548 6.05e-06 ***
## HHVEH1      -0.022634   0.003663  -6.179 9.28e-10 ***
## HHVEH2       0.005652   0.003683   1.535  0.12517
## HHVEH3       0.021955   0.007723   2.843  0.00456 **
## HHVEH4       0.097393   0.015591   6.247 6.15e-10 ***
## HHVEH5      -0.366058   0.037928  -9.651  < 2e-16 ***
## HHVEH6       0.116903   0.054222   2.156  0.03132 *
## HHAGE7       0.019793   0.003828   5.171 2.80e-07 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 5.98 on 1021 degrees of freedom
## Multiple R-squared:  0.3552, Adjusted R-squared:  0.3501
## F-statistic: 70.29 on 8 and 1021 DF,  p-value: < 2.2e-16
```

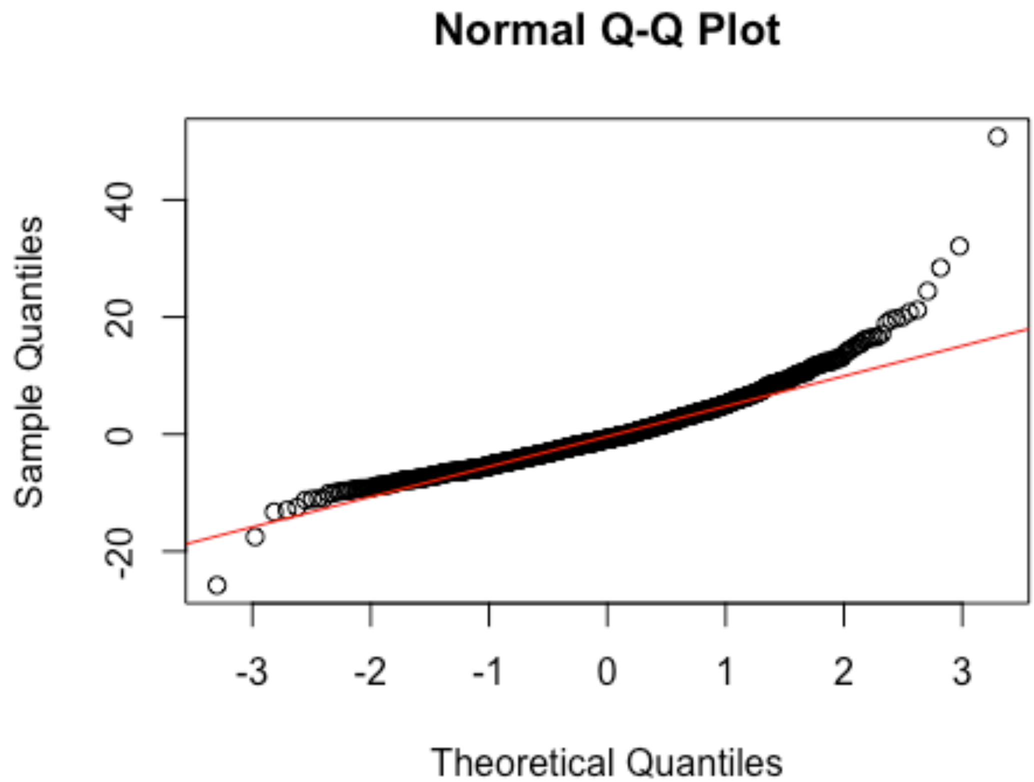
Check residuals vs fitted y values

```
VMTprOLSresnoLU <- resid(VMTprOLSnoLU) # save the residuals

plot(resid(VMTprOLSnoLU), fitted(VMTprOLSnoLU)) # Tukey-Anscombe's plot
```



Check if residuals look like normally distributed

```
qqnorm(VMTprOLSresnoLU)
qqline(VMTprOLSresnoLU,col="red")
```



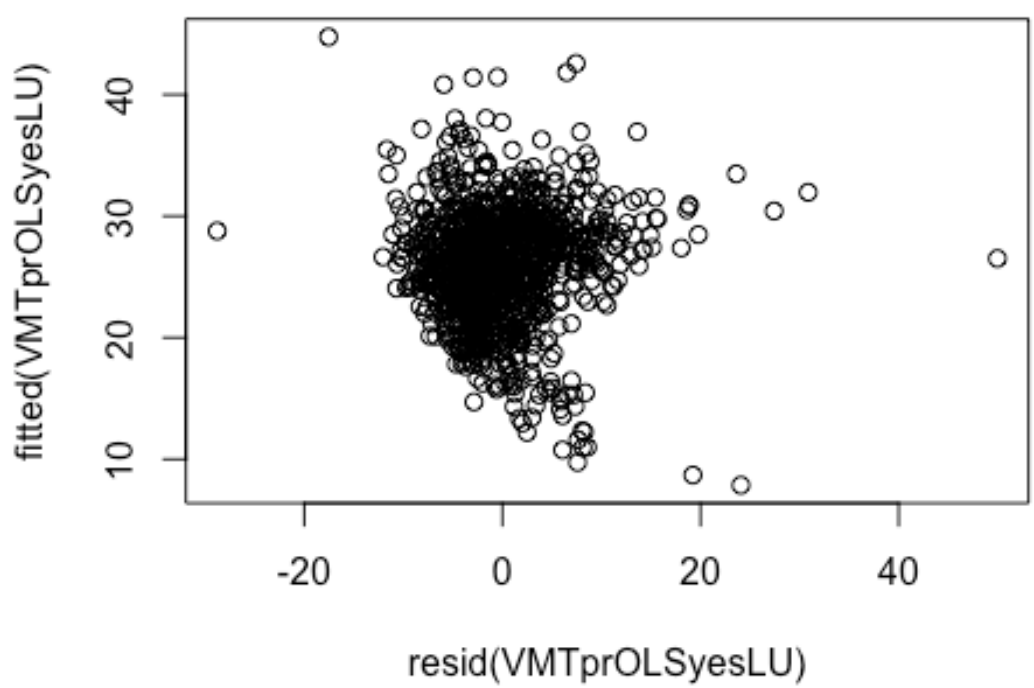
Next model adding the BG classification into center, suburb, exurb, rural

```
VMTprOLSyesLU<-lm(VMTpr~suburb+exurb+rural+HHVEH0+HHVEH1+HHVEH2+HHVEH3+HHVEH4
+HHVEH5+HHVEH6+HHAGE7, data=YCOUNTY@data)
summary(VMTprOLSyesLU)

##
## Call:
## lm(formula = VMTpr ~ suburb + exurb + rural + HHVEH0 + HHVEH1 +
##     HHVEH2 + HHVEH3 + HHVEH4 + HHVEH5 + HHVEH6 + HHAGE7, data = YCOUNTY@da
ta)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -28.781  -3.742  -0.633   3.031  49.963
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept) 21.5608253  0.7433166  29.006  < 2e-16 ***
## suburbTRUE   5.0324931  0.6974305   7.216 1.05e-12 ***
## exurbTRUE    2.7945616  0.8729604   3.201  0.00141 **
## ruralTRUE    7.2197938  1.1558705   6.246 6.17e-10 ***
## HHVEH0      -0.0411067  0.0100889  -4.074 4.97e-05 ***
```

```
## HHVEH1      -0.0182446  0.0036618  -4.982 7.38e-07 ***
## HHVEH2       0.0005669  0.0036996   0.153  0.87825
## HHVEH3       0.0206943  0.0074417   2.781  0.00552 **
## HHVEH4       0.1141963  0.0155432   7.347 4.15e-13 ***
## HHVEH5      -0.3092137  0.0384675  -8.038 2.51e-15 ***
## HHVEH6       0.1192213  0.0572565   2.082  0.03757 *
## HHAGE7       0.0176199  0.0038043   4.632 4.10e-06 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 5.756 on 1018 degrees of freedom
## Multiple R-squared:  0.4044, Adjusted R-squared:  0.398
## F-statistic: 62.84 on 11 and 1018 DF,  p-value: < 2.2e-16
```

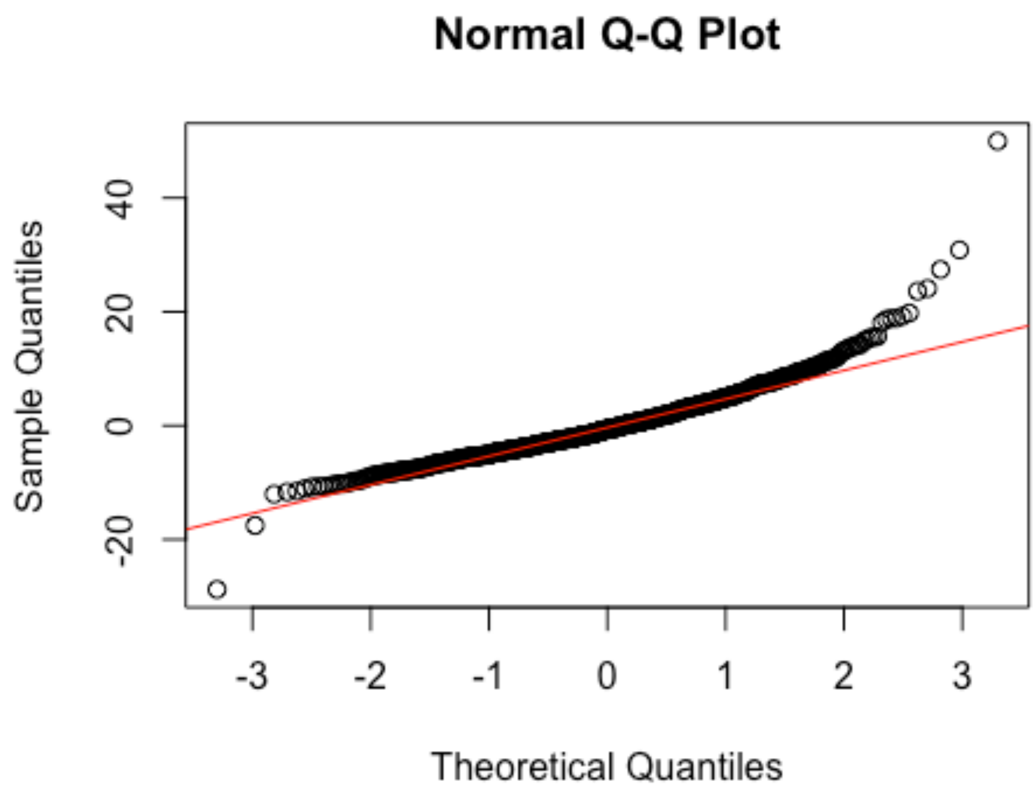
The model above shows we improved its specification by adding the categories of each BG

```r
VMTprOLSresyesLU <- resid(VMTprOLSyesLU) # save the residuals

plot(resid(VMTprOLSyesLU), fitted(VMTprOLSyesLU)) # Tukey-Anscombe's plot
```



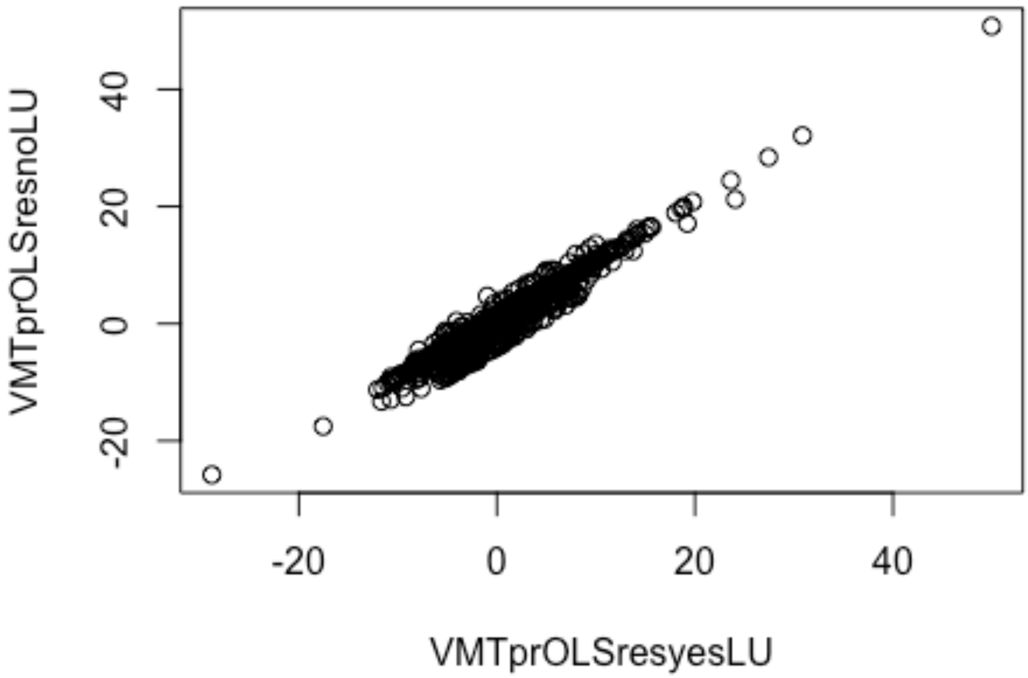
The residuals don't look very diffrent than the noLU

```r
qqnorm(VMTprOLSresyesLU)
qqline(VMTprOLSresyesLU,col="red")
```



Let's compare residduals from model with spatial land use variables (yesLU) and without (noLU)

```
plot(VMTprOLSresyesLU, VMTprOLSresnoLU) # comparison of residuals from two OL
S models
```


VMTprOLSresnoLU
-20
0
20
40
-20
0
20
40
VMTprOLSresyesLU

Part 2 Spatial weights creation and checking the data

```
list.queenY<-poly2nb(YCOUNTY, queen=T) # from polygons create neighborhoods
coordsY<-coordinates(YCOUNTY) # assign coordinate system
plot(YCOUNTY) # plot the polygons
plot(list.queenY, coordsY, add=T) # plot the links on top of polygons
```

```
summary(list.queenY) #  summarize the connectivity

## Neighbour list object:
## Number of regions: 1030
## Number of nonzero links: 6508
## Percentage nonzero weights: 0.6134414
## Average number of links: 6.318447
## Link number distribution:
##
##   1   2   3   4   5   6   7   8   9  10  11  12  13  14  15  16  22
##   1  12  56 102 215 217 186 113  62  28  16   9   3   3   3   2   2
## 1 least connected region:
## 9993 with 1 link
## 2 most connected regions:
## 20444 20592 with 22 links
```

```r
queen_w <- nb2listw(list.queenY, style="W")  # create the spatial weights tha
t are row standradized
summary(queen_w)  # check the weights matrix
```

```
## Characteristics of weights list object:
## Neighbour list object:
## Number of regions: 1030
## Number of nonzero links: 6508
## Percentage nonzero weights: 0.6134414
## Average number of links: 6.318447
## Link number distribution:
##
##   1   2   3   4   5   6   7   8   9  10  11  12  13  14  15  16  22
##   1  12  56 102 215 217 186 113  62  28  16   9   3   3   3   2   2
## 1 least connected region:
## 9993 with 1 link
## 2 most connected regions:
## 20444 20592 with 22 links
##
## Weights style: W
## Weights constants summary:
##      n      nn   S0       S1     S2
## W 1030 1060900 1030 345.3927 4291.3
```

You will need this for your assignment. Do not run in the class lab

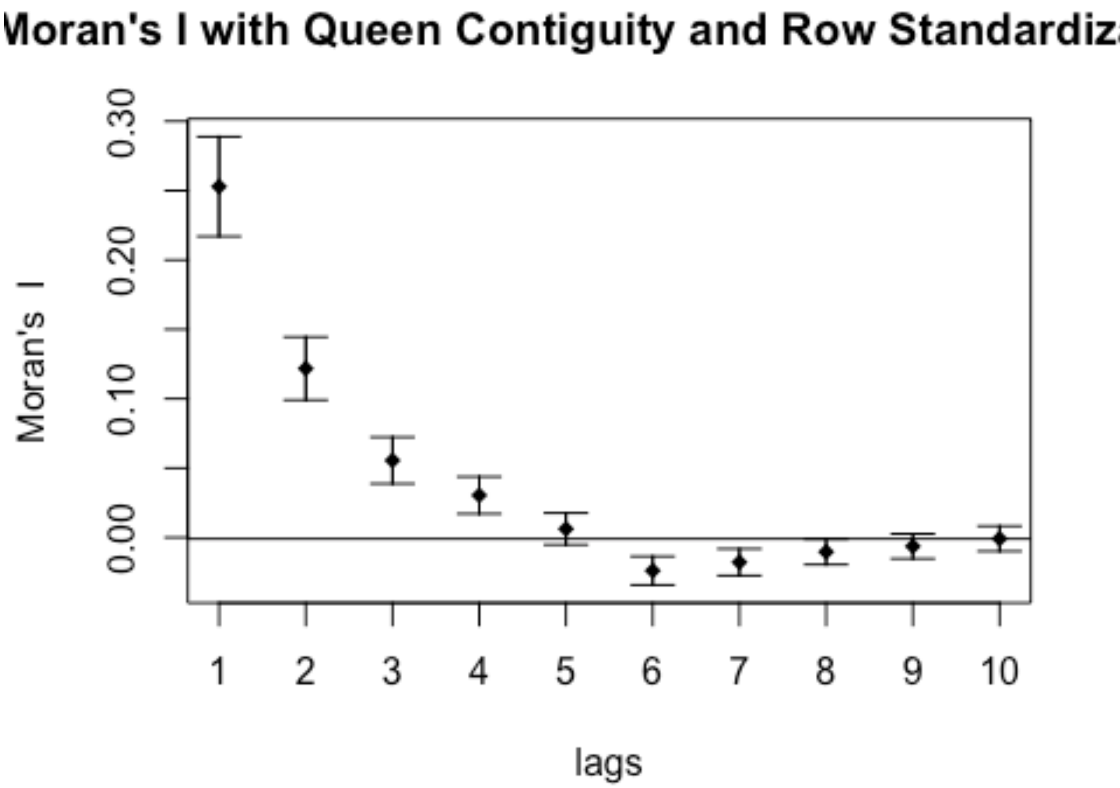
K = 10 nearest neighborhood

coords<-coordinates(YCOUNTY) IDs<-row.names(as(YCOUNTY, "data.frame")) list_kn10<-knn2nb(knearneigh(coords, k=10), row.names=IDs) plot(list_kn10, coordsY, add=T) summary (list_kn10) kn10_w <- nb2listw(list_kn10, style="W") summary(kn10_w) structure(kn10_w)

You will need the code above for your assignment. Do not run in the class lab

Moran's I to get an idea of possible spatial correlation with lags > 1

```r
mor10q <- sp.correlogram(list.queenY, var=YCOUNTY@data$VMTpr, order=10, metho
d="I")
plot(mor10q, main = "Moran's I with Queen Contiguity and Row Standardization"
)
```



The first lag is the highest as expected. This tells me there is spatial autocorrelation but does not tell me of which type.

Part 3 Spatial Regression Models

First we define some basic items and then move to specifying and estimating models.

The spatial lag model (Y of a BG is a function of the neibhorhood's Ys)

$$Y = X\beta + \rho WY + e$$

This model is estimated using the function lagsarlm

The spatial error model (the error of a BG is a function of the neighborhood's errors)

$$Y = X\beta + u$$

$$u = \lambda Wu + e$$

The SAC/SARAR model is

$$Y = X\beta + \rho WY + u$$

$$u = \lambda Wu + e$$

Anselin derived statistical tests that can be used to give us an idea of the possible model.

In R this is done with the function lm.LMtests

From the vignette:

The function reports the estimates of tests chosen among five statistics for testing for spatial dependence in linear models. The statistics are the simple LM test for error dependence (LMerr), the simple LM test for a missing spatially lagged dependent variable (LMlag), variants of these robust to the presence of the other (RLMerr, RLMlag - RLMerr tests for error dependence in the possible presence of a missing lagged dependent variable, RLMlag the other way round), and a portmanteau test (SARMA, in fact LMerr + RLMlag). Note: from spdep 0.3-32, the value of the weights matrix trace term is returned correctly for both underlying symmetric and asymmetric neighbour lists, before 0.3-32, the value was wrong for listw objects based on asymmetric neighbour lists, such as k-nearest neighbours (thanks to Luc Anselin for finding the bug).

From the LM vignette:

The two types of dependence are for spatial lag rho and spatial error lambda:

y = X beta + rho W1 y + u

u = lambda W2 u + e

where e is a well-behaved, uncorrelated error term. Tests for a missing spatially lagged dependent variable test that rho = 0, tests for spatial autocorrelation of the error u test whether lambda = 0. W is a spatial weights matrix; for the tests used here they are identical.

```
LM<-lm.LMtests(VMTprOLSyesLU, queen_w, test="all")
print(LM)

##
##  Lagrange multiplier diagnostics for spatial dependence
##
## data:
## model: lm(formula = VMTpr ~ suburb + exurb + rural + HHVEH0 +
## HHVEH1 + HHVEH2 + HHVEH3 + HHVEH4 + HHVEH5 + HHVEH6 + HHAGE7, data
## = YCOUNTY@data)
## weights: queen_w
##
## LMerr = 9.5722, df = 1, p-value = 0.001975
##
##
##  Lagrange multiplier diagnostics for spatial dependence
##
## data:
## model: lm(formula = VMTpr ~ suburb + exurb + rural + HHVEH0 +
## HHVEH1 + HHVEH2 + HHVEH3 + HHVEH4 + HHVEH5 + HHVEH6 + HHAGE7, data
## = YCOUNTY@data)
## weights: queen_w
##
## LMlag = 17.869, df = 1, p-value = 2.367e-05
##
##
##  Lagrange multiplier diagnostics for spatial dependence
##
## data:
## model: lm(formula = VMTpr ~ suburb + exurb + rural + HHVEH0 +
## HHVEH1 + HHVEH2 + HHVEH3 + HHVEH4 + HHVEH5 + HHVEH6 + HHAGE7, data
## = YCOUNTY@data)
## weights: queen_w
##
## RLMerr = 0.4513, df = 1, p-value = 0.5017
##
##
##  Lagrange multiplier diagnostics for spatial dependence
##
## data:
## model: lm(formula = VMTpr ~ suburb + exurb + rural + HHVEH0 +
## HHVEH1 + HHVEH2 + HHVEH3 + HHVEH4 + HHVEH5 + HHVEH6 + HHAGE7, data
## = YCOUNTY@data)
## weights: queen_w
##
## RLMlag = 8.7479, df = 1, p-value = 0.0031
##
##
##  Lagrange multiplier diagnostics for spatial dependence
##
```

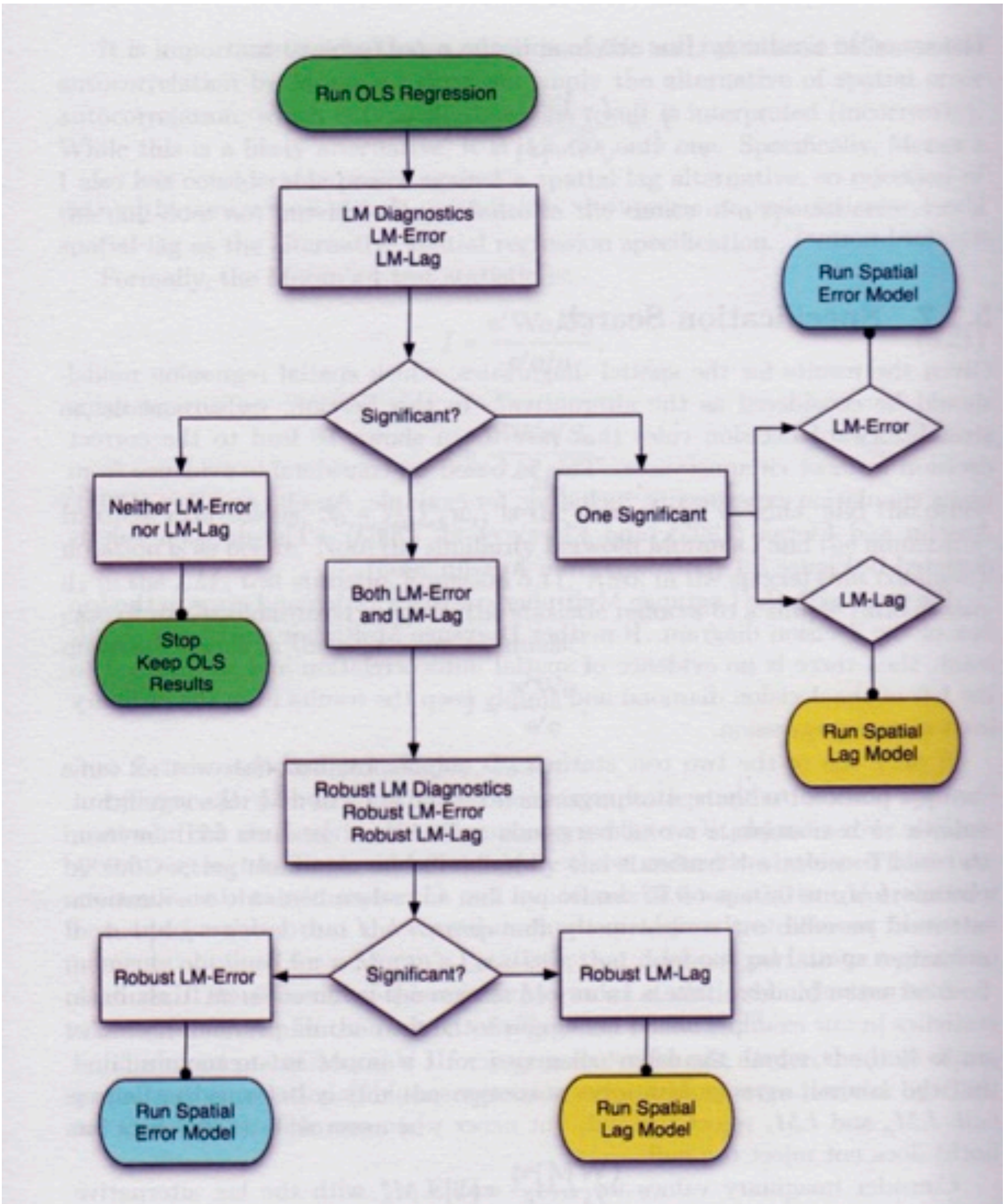
```
## data:
## model: lm(formula = VMTpr ~ suburb + exurb + rural + HHVEH0 +
## HHVEH1 + HHVEH2 + HHVEH3 + HHVEH4 + HHVEH5 + HHVEH6 + HHAGE7, data
## = YCOUNTY@data)
## weights: queen_w
##
## SARMA = 18.32, df = 2, p-value = 0.0001052
```

LMerr > critical value of a chi-square distribution and points to possible spatial error correlation (u depends on neighbors) LMlag > critical value of a chi-square distribution and points to possible spatial lag (Y depends on neighbors)

BUT the more robust statistics (see Anselin's paper on Gauchospace):

RLMerr < critical value of a chi-square distribution and points to spurious results RLMlag > critical value of a chi-square distribution and points to possible spatial lag (Y depends on neighbors)

SARMA > critical value of a chi-square distribution (with 2 df) and points to possible spatial lag (Y depends on neighbors) and possible spatial error correlation (u depends on neighbors)



Anselin-Rey Flowchart

Estimation of Spatial Regression Models

The spatial lag model (Y of a BG is a function of the neibhorhood's Ys)

$$Y = X\beta + \rho WY + e$$

This model is estimated using the function lagsarlm

```
SpaLag <- lagsarlm(VMTpr~suburb+exurb+rural+HHVEH0+HHVEH1+HHVEH2+HHVEH3+HHVEH
4+HHVEH5+HHVEH6+HHAGE7, data=YCOUNTY, queen_w)
summary(SpaLag)

##
## Call:lagsarlm(formula = VMTpr ~ suburb + exurb + rural + HHVEH0 +
##     HHVEH1 + HHVEH2 + HHVEH3 + HHVEH4 + HHVEH5 + HHVEH6 + HHAGE7,
##     data = YCOUNTY, listw = queen_w)
##
## Residuals:
##       Min        1Q    Median        3Q       Max
## -29.85777  -3.65292  -0.67104   2.96769  50.12344
##
## Type: lag
## Coefficients: (asymptotic standard errors)
##                Estimate  Std. Error z value  Pr(>|z|)
## (Intercept)  1.7735e+01  1.1715e+00 15.1380 < 2.2e-16
## suburbTRUE   4.4683e+00  7.0762e-01  6.3146 2.709e-10
## exurbTRUE    2.1527e+00  8.7654e-01  2.4559   0.01405
## ruralTRUE    6.1908e+00  1.1616e+00  5.3294 9.852e-08
## HHVEH0      -4.0378e-02  9.9355e-03 -4.0640 4.823e-05
## HHVEH1      -1.6191e-02  3.6212e-03 -4.4711 7.781e-06
## HHVEH2       1.9491e-05  3.6402e-03  0.0054   0.99573
## HHVEH3       1.8240e-02  7.3472e-03  2.4826   0.01304
## HHVEH4       1.1072e-01  1.5317e-02  7.2288 4.874e-13
## HHVEH5      -2.9237e-01  3.7935e-02 -7.7071 1.288e-14
## HHVEH6       1.1247e-01  5.6518e-02  1.9899   0.04660
## HHAGE7       1.6838e-02  3.7465e-03  4.4942 6.983e-06
##
## Rho: 0.16958, LR test value: 16.441, p-value: 5.0186e-05
## Asymptotic standard error: 0.041528
##     z-value: 4.0835, p-value: 4.4371e-05
## Wald statistic: 16.675, p-value: 4.4371e-05
##
## Log likelihood: -3249.915 for lag model
## ML residual variance (sigma squared): 32.07, (sigma: 5.663)
## Number of observations: 1030
## Number of parameters estimated: 14
## AIC: 6527.8, (AIC for lm: 6542.3)
## LM test for residual autocorrelation
## test value: 1.2262, p-value: 0.26815
```

This output shows the Rho is significantly different than zero. This means Y for each BG depends on the Ys of its neighbors We will compare the betas after we finsh the other models.

Although the Anselin LM test show the spatial erros may not be correlated, I still want to sestimate a model and check.

The spatial error model (the error of a BG is a function of the neighborhood's errors)

$$Y = X\beta + u$$

$$u = \lambda W u + e$$

```
SpaErr<-errorsarlm(VMTpr~suburb+exurb+rural+HHVEH0+HHVEH1+HHVEH2+HHVEH3+HHVEH
4+HHVEH5+HHVEH6+HHAGE7, data=YCOUNTY, queen_w)
summary(SpaErr)

##
## Call:errorsarlm(formula = VMTpr ~ suburb + exurb + rural + HHVEH0 +
##     HHVEH1 + HHVEH2 + HHVEH3 + HHVEH4 + HHVEH5 + HHVEH6 + HHAGE7,
##     data = YCOUNTY, listw = queen_w)
##
## Residuals:
##       Min        1Q    Median        3Q       Max
## -29.41623  -3.66937  -0.68831   2.91623  49.79925
##
## Type: error
## Coefficients: (asymptotic standard errors)
##               Estimate  Std. Error z value  Pr(>|z|)
## (Intercept) 21.43716749  0.79547678 26.9488 < 2.2e-16
## suburbTRUE   5.27416883  0.74615025  7.0685 1.566e-12
## exurbTRUE    2.99026386  0.92354089  3.2378  0.001204
## ruralTRUE    7.19476452  1.25446480  5.7353 9.732e-09
## HHVEH0      -0.04191533  0.01010363 -4.1485 3.346e-05
## HHVEH1      -0.01687704  0.00367620 -4.5909 4.414e-06
## HHVEH2      -0.00080136  0.00368589 -0.2174  0.827886
## HHVEH3       0.02056642  0.00738607  2.7845  0.005361
## HHVEH4       0.11230963  0.01543053  7.2784 3.377e-13
## HHVEH5      -0.29355169  0.03852607 -7.6196 2.554e-14
## HHVEH6       0.13777889  0.05758508  2.3926  0.016729
## HHAGE7       0.01702768  0.00388260  4.3856 1.156e-05
##
## Lambda: 0.15944, LR test value: 9.3135, p-value: 0.0022747
## Asymptotic standard error: 0.050331
##     z-value: 3.1679, p-value: 0.0015355
## Wald statistic: 10.035, p-value: 0.0015355
##
## Log likelihood: -3253.479 for error model
## ML residual variance (sigma squared): 32.311, (sigma: 5.6843)
## Number of observations: 1030
## Number of parameters estimated: 14
## AIC: 6535, (AIC for lm: 6542.3)
```

From this output we find the estimated model is contradicting the RLMerr because Lambda is significantly different than zero.

Then let's try the SAC/SARAR model

The SAC/SARAR model is

$$Y = X\beta + \rho WY + u$$

$$u = \lambda Wu + e$$

```
SARAR<-sacsarlm(VMTpr~suburb+exurb+rural+HHVEH0+HHVEH1+HHVEH2+HHVEH3+HHVEH4+H
HVEH5+HHVEH6+HHAGE7, data=YCOUNTY, queen_w)
summary(SARAR)

##
## Call:sacsarlm(formula = VMTpr ~ suburb + exurb + rural + HHVEH0 +
##     HHVEH1 + HHVEH2 + HHVEH3 + HHVEH4 + HHVEH5 + HHVEH6 + HHAGE7,
##     data = YCOUNTY, listw = queen_w)
##
## Residuals:
##       Min        1Q    Median        3Q       Max
## -29.79492  -3.66546  -0.65543   2.94504  50.02582
##
## Type: sac
## Coefficients: (asymptotic standard errors)
##                Estimate  Std. Error z value  Pr(>|z|)
## (Intercept) 15.58920522  1.62614923  9.5866 < 2.2e-16
## suburbTRUE   3.89440336  0.69220564  5.6261 1.844e-08
## exurbTRUE    1.61879375  0.84355809  1.9190   0.05498
## ruralTRUE    5.57320349  1.10336718  5.0511 4.393e-07
## HHVEH0      -0.03810721  0.00972781 -3.9173 8.953e-05
## HHVEH1      -0.01587306  0.00357460 -4.4405 8.975e-06
## HHVEH2       0.00098036  0.00356483  0.2750   0.78331
## HHVEH3       0.01606099  0.00730605  2.1983   0.02793
## HHVEH4       0.10876310  0.01516537  7.1718 7.401e-13
## HHVEH5      -0.28855301  0.03748384 -7.6981 1.377e-14
## HHVEH6       0.08572671  0.05539256  1.5476   0.12171
## HHAGE7       0.01654149  0.00360246  4.5917 4.396e-06
##
## Rho: 0.27033
## Asymptotic standard error: 0.067337
##     z-value: 4.0147, p-value: 5.9533e-05
## Lambda: -0.16084
## Asymptotic standard error: 0.096992
##     z-value: -1.6582, p-value: 0.097267
##
## LR test value: 18.311, p-value: 0.00010563
##
## Log likelihood: -3248.98 for sac model
## ML residual variance (sigma squared): 31.643, (sigma: 5.6252)
## Number of observations: 1030
## Number of parameters estimated: 15
## AIC: 6528, (AIC for lm: 6542.3)
```

Now the picture is clearer. The spatial lagged dependent variables (the neighbors by the queen contiguity) influence the Ys in each BG.
The spatial errors also seem to be correlated with neighbors in a significant way but at about 90% of confidence.

Maybe something is going on. Would it be that the Xs are spatially correlated?

There is a way to check this: Using the "mixed" type in the lagsarlm.

```
SpaLagMix <- lagsarlm(VMTpr~suburb+exurb+rural+HHVEH0+HHVEH1+HHVEH2+HHVEH3+HH
VEH4+HHVEH5+HHVEH6+HHAGE7, data=YCOUNTY, queen_w, type="mixed")
summary(SpaLagMix)

##
## Call:lagsarlm(formula = VMTpr ~ suburb + exurb + rural + HHVEH0 +
##     HHVEH1 + HHVEH2 + HHVEH3 + HHVEH4 + HHVEH5 + HHVEH6 + HHAGE7,
##     data = YCOUNTY, listw = queen_w, type = "mixed")
##
## Residuals:
##       Min        1Q    Median        3Q       Max
## -29.92023  -3.60877  -0.62666   2.87078  50.59810
##
## Type: mixed
## Coefficients: (asymptotic standard errors)
##                  Estimate Std. Error z value  Pr(>|z|)
## (Intercept)    19.8224873  1.5775993 12.5650 < 2.2e-16
## suburbTRUE      5.4115949  1.0938128  4.9475 7.519e-07
## exurbTRUE       2.7536066  1.3278742  2.0737 0.0381076
## ruralTRUE       6.7130718  1.8962521  3.5402 0.0003999
## HHVEH0         -0.0392749  0.0101475 -3.8704 0.0001087
## HHVEH1         -0.0151966  0.0037418 -4.0613 4.879e-05
## HHVEH2         -0.0015679  0.0036900 -0.4249 0.6709079
## HHVEH3          0.0172148  0.0073842  2.3313 0.0197382
## HHVEH4          0.1142135  0.0154808  7.3777 1.610e-13
## HHVEH5         -0.2769408  0.0387478 -7.1473 8.853e-13
## HHVEH6          0.1317582  0.0584244  2.2552 0.0241214
## HHAGE7          0.0154661  0.0041447  3.7315 0.0001903
## lag.suburbTRUE -3.1293489  1.5658723 -1.9985 0.0456657
## lag.exurbTRUE  -3.2543021  1.9715196 -1.6507 0.0988087
## lag.ruralTRUE  -0.5517768  2.4344136 -0.2267 0.8206905
## lag.HHVEH0      0.0121967  0.0210028  0.5807 0.5614296
## lag.HHVEH1     -0.0136500  0.0075922 -1.7979 0.0721924
## lag.HHVEH2      0.0187294  0.0076565  2.4462 0.0144371
## lag.HHVEH3     -0.0178873  0.0162324 -1.1019 0.2704847
## lag.HHVEH4      0.0451336  0.0340232  1.3266 0.1846565
## lag.HHVEH5     -0.0977881  0.0819439 -1.1934 0.2327306
## lag.HHVEH6     -0.2958412  0.1156543 -2.5580 0.0105283
## lag.HHAGE7      0.0033522  0.0077542  0.4323 0.6655163
##
## Rho: 0.12804, LR test value: 6.2727, p-value: 0.012261
```

```
## Asymptotic standard error: 0.050853
##     z-value: 2.5178, p-value: 0.011809
## Wald statistic: 6.3394, p-value: 0.011809
##
## Log likelihood: -3239.985 for mixed model
## ML residual variance (sigma squared): 31.522, (sigma: 5.6145)
## Number of observations: 1030
## Number of parameters estimated: 25
## AIC: 6530, (AIC for lm: 6534.2)
## LM test for residual autocorrelation
## test value: 5.0025, p-value: 0.02531
```

This time we get a smaller Rho (of course because the Xs of the neighbors influence their Ys and we use WY to compute it)

Let's discuss in class what all these findings mean.

Now let's try everything "under the kitchen sink." this is a SARAR model with lagged Xs

The SAC/SARAR model with lagged Xs is

$$Y = X\beta + \rho WY + \gamma WX + u$$

$$u = \lambda Wu + e$$

```
SARARMix<-sacsarlm(VMTpr~suburb+exurb+rural+HHVEH0+HHVEH1+HHVEH2+HHVEH3+HHVEH
4+HHVEH5+HHVEH6+HHAGE7, data=YCOUNTY, queen_w, type= "sacmixed")
summary(SARARMix)
```

```
##
## Call:sacsarlm(formula = VMTpr ~ suburb + exurb + rural + HHVEH0 +
##     HHVEH1 + HHVEH2 + HHVEH3 + HHVEH4 + HHVEH5 + HHVEH6 + HHAGE7,
##     data = YCOUNTY, listw = queen_w, type = "sacmixed")
##
## Residuals:
##       Min        1Q    Median        3Q       Max
## -28.78953  -3.44710  -0.58703   2.70654  48.20969
##
## Type: sacmixed
## Coefficients: (asymptotic standard errors)
##                 Estimate Std. Error z value  Pr(>|z|)
## (Intercept)   10.4867634  1.9920034  5.2644 1.406e-07
## suburbTRUE     5.6571147  1.0972712  5.1556 2.528e-07
## exurbTRUE      3.0875577  1.3360372  2.3110 0.0208339
## ruralTRUE      6.6193244  1.9328259  3.4247 0.0006155
## HHVEH0        -0.0393914  0.0104511 -3.7691 0.0001638
## HHVEH1        -0.0145751  0.0038408 -3.7948 0.0001478
## HHVEH2        -0.0023045  0.0037734 -0.6107 0.5413776
## HHVEH3         0.0173154  0.0074754  2.3163 0.0205413
## HHVEH4         0.1088545  0.0156610  6.9507 3.635e-12
## HHVEH5        -0.2541596  0.0397382 -6.3958 1.597e-10
## HHVEH6         0.1532476  0.0599578  2.5559 0.0105906
## HHAGE7         0.0147813  0.0042007  3.5188 0.0004336
## lag.suburbTRUE -4.8764475  1.4462329 -3.3718 0.0007467
## lag.exurbTRUE  -4.0702935  1.7885831 -2.2757 0.0228635
## lag.ruralTRUE  -3.3463018  2.3661293 -1.4143 0.1572881
## lag.HHVEH0      0.0267375  0.0181038  1.4769 0.1397029
## lag.HHVEH1     -0.0026070  0.0067874 -0.3841 0.7009086
## lag.HHVEH2      0.0141021  0.0066380  2.1245 0.0336314
## lag.HHVEH3     -0.0228213  0.0138016 -1.6535 0.0982241
## lag.HHVEH4     -0.0048585  0.0308955 -0.1573 0.8750428
## lag.HHVEH5      0.0367602  0.0752991  0.4882 0.6254156
## lag.HHVEH6     -0.3102292  0.1005189 -3.0863 0.0020268
## lag.HHAGE7     -0.0042196  0.0068339 -0.6175 0.5369365
##
## Rho: 0.54813
```

```
## Asymptotic standard error: 0.082628
##     z-value: 6.6337, p-value: 3.2731e-11
## Lambda: -0.55431
## Asymptotic standard error: 0.13124
##     z-value: -4.2235, p-value: 2.4051e-05
##
## LR test value: 43.388, p-value: 3.8743e-05
##
## Log likelihood: -3236.442 for sacmixed model
## ML residual variance (sigma squared): 28.341, (sigma: 5.3236)
## Number of observations: 1030
## Number of parameters estimated: 26
## AIC: 6524.9, (AIC for lm: 6542.3)
```

Part 4. Model Comparison and Class Discussion

Let's look at a comparison among these models (see handout in class) – I created this with stargazer but Rmarkdown gave me a hard time and I did this the old fashion way.

Comparison of Regression Models Riverside

	Dependent variable:				
	The First Five Models				
	OLS		*spatial autoregressive*	*spatial error*	*spatial autoregressive*
	(1)	(2)	(3)	(4)	(5)
BG is Suburb		5.032***	4.468***	5.274***	5.412***
		(0.697)	(0.708)	(0.746)	(1.094)
BG is Exurb		2.795***	2.153**	2.990***	2.754**
		(0.873)	(0.877)	(0.924)	(1.328)
BG is Rural		7.220***	6.191***	7.195***	6.713***
		(1.156)	(1.162)	(1.254)	(1.896)
Households with no cars	-0.044***	-0.041***	-0.040***	-0.042***	-0.039***
	(0.010)	(0.010)	(0.010)	(0.010)	(0.010)
Households with 1 car	-0.023***	-0.018***	-0.016***	-0.017***	-0.015***
	(0.004)	(0.004)	(0.004)	(0.004)	(0.004)
Households with 2 cars	0.006	0.001	0.00002	-0.001	-0.002
	(0.004)	(0.004)	(0.004)	(0.004)	(0.004)
Households with 3 cars	0.022***	0.021***	0.018**	0.021***	0.017**
	(0.008)	(0.007)	(0.007)	(0.007)	(0.007)
Households with 4 cars	0.097***	0.114***	0.111***	0.112***	0.114***
	(0.016)	(0.016)	(0.015)	(0.015)	(0.015)
Households with 5 cars	-0.366***	-0.309***	-0.292***	-0.294***	-0.277***
	(0.038)	(0.038)	(0.038)	(0.039)	(0.039)
Households with 6+ cars	0.117**	0.119**	0.112**	0.138**	0.132**
	(0.054)	(0.057)	(0.057)	(0.058)	(0.058)
Older Households	0.020***	0.018***	0.017***	0.017***	0.015***
	(0.004)	(0.004)	(0.004)	(0.004)	(0.004)
lagBG is Suburb					-3.129**
					(1.566)
lagBG is Exurb					-3.254*
					(1.972)
lagBG is Rural					-0.552
					(2.434)
lagHouseholds with no cars					0.012
					(0.021)
lagHouseholds with 1 car					-0.014*
					(0.008)

Model comparison Part 1

lagHouseholds with 2 cars					0.019^{**}
					(0.008)
lagHouseholds with 3 cars					-0.018
					(0.016)
lagHouseholds with 4 cars					0.045
					(0.034)
lagHouseholds with 5 cars					-0.098
					(0.082)
lagHouseholds with 6+ cars					-0.296^{**}
					(0.116)
lagOlder Households					0.003
					(0.008)
Constant	25.791^{***}	21.561^{***}	17.735^{***}	21.437^{***}	19.822^{***}
	(0.431)	(0.743)	(1.172)	(0.795)	(1.578)
Observations	1,030	1,030	1,030	1,030	1,030
R^2	0.355	0.404			
Adjusted R^2	0.350	0.398			
Log Likelihood			-3,249.915	-3,253.479	-3,239.985
$sigma^2$			32.070	32.311	31.522
Akaike Inf. Crit.			6,527.831	6,534.958	6,529.969
Residual Std. Error	5.980 (df = 1021)	5.756 (df = 1018)			
F Statistic	70.294^{***} (df = 8; 1021)	62.841^{***} (df = 11; 1018)			
Wald Test (df = 1)			16.675^{***}	10.035^{***}	6.339^{**}
LR Test (df = 1)			16.441^{***}	9.313^{***}	6.273^{**}
Note:					$^{*}p<0.1$; $^{**}p<0.05$; $^{***}p<0.01$

Model comparison Part 2

The effect of a variable X on the variable Y for models that do not have lagged dependent variables is measured by the regression coefficient (they are the partial derivatives of Y with respect to an x in a linear regerssion).

When we have lagged dependent variables the following happens:

Consider two Block Groups BG101 and BG102 that are neighbors). The Xs of BG101 influence the Y of BG101. They also influence the Y of BG102. In lag dependent variable model the Y of BG102 is also in the specification of the model. This means the Xs of BG101 have a direct effect on the Y of BG101 and an indirect effect on the Y of BG101 through the Y of BG102.
We can estimate these using the function impact.

```
impacts(SpaLag,listw=queen_w )

## Impact measures (lag, exact):
##                    Direct      Indirect         Total
## suburbTRUE  4.4899474873  8.908421e-01  5.380790e+00
## exurbTRUE   2.1631379588  4.291841e-01  2.592322e+00
## ruralTRUE   6.2207573218  1.234249e+00  7.455006e+00
## HHVEH0     -0.0405735000 -8.050112e-03 -4.862361e-02
## HHVEH1     -0.0162691503 -3.227932e-03 -1.949708e-02
## HHVEH2      0.0000195852  3.885862e-06  2.347106e-05
## HHVEH3      0.0183282266  3.636469e-03  2.196470e-02
## HHVEH4      0.1112590352  2.207470e-02  1.333337e-01
## HHVEH5     -0.2937820560 -5.828875e-02 -3.520708e-01
## HHVEH6      0.1130111848  2.242234e-02  1.354335e-01
## HHAGE7      0.0169190706  3.356881e-03  2.027595e-02

impacts(SpaLagMix,listw=queen_w )

## Impact measures (mixed, exact):
##                   Direct     Indirect         Total
## suburbTRUE  5.360105636 -2.742737073  2.6173685631
## exurbTRUE   2.692284735 -3.266501757 -0.5742170221
## ruralTRUE   6.719575407  0.346437293  7.0660127004
## HHVEH0     -0.039123436  0.008069120 -0.0310543157
## HHVEH1     -0.015526149 -0.017556175 -0.0330823236
## HHVEH2     -0.001176307  0.020857798  0.0196814911
## HHVEH3      0.016883351 -0.017654585 -0.0007712344
## HHVEH4      0.115476381  0.067269097  0.1827454786
## HHVEH5     -0.279756720 -0.149996966 -0.4297536866
## HHVEH6      0.125862733 -0.314039522 -0.1881767891
## HHAGE7      0.015578769  0.006002769  0.0215815378
```

```r
impacts(SARAR,listw=queen_w )

## Impact measures (sac, exact):
##                   Direct      Indirect       Total
## suburbTRUE   3.9449560612  1.3922894954  5.33724556
## exurbTRUE    1.6398070884  0.5787355165  2.21854260
## ruralTRUE    5.6455484631  1.9924779133  7.63802638
## HHVEH0      -0.0386018701 -0.0136237204 -0.05222559
## HHVEH1      -0.0160791041 -0.0056747825 -0.02175389
## HHVEH2       0.0009930822  0.0003504875  0.00134357
## HHVEH3       0.0162694730  0.0057419692  0.02201144
## HHVEH4       0.1101749377  0.0388839333  0.14905887
## HHVEH5      -0.2922986723 -0.1031606853 -0.39545936
## HHVEH6       0.0868395152  0.0306481854  0.11748770
## HHAGE7       0.0167562100  0.0059137528  0.02266996

impacts(SARARMix,listw=queen_w )

## Impact measures (sacmixed, exact):
##                   Direct     Indirect       Total
## suburbTRUE   5.4415930795 -3.713957702  1.72763538
## exurbTRUE    2.7989302478 -4.973748233 -2.17481799
## ruralTRUE    6.6535466352  0.589731273  7.24327791
## HHVEH0      -0.0387668044  0.010763315 -0.02800349
## HHVEH1      -0.0158612298 -0.022163157 -0.03802439
## HHVEH2      -0.0007461632  0.026854496  0.02610833
## HHVEH3       0.0156973646 -0.027882204 -0.01218484
## HHVEH4       0.1155070009  0.114638613  0.23014561
## HHVEH5      -0.2666072160 -0.214502787 -0.48111000
## HHVEH6       0.1257881375 -0.473192294 -0.34740416
## HHAGE7       0.0152525705  0.008120737  0.02337331
```

Compare this with the betas we get from ordinary least squares (I ma running the model here again)

```
VMTprOLSyesLU<-lm(VMTpr~suburb+exurb+rural+HHVEH0+HHVEH1+HHVEH2+HHVEH3+HHVEH4
+HHVEH5+HHVEH6+HHAGE7, data=YCOUNTY@data)
summary(VMTprOLSyesLU)

##
## Call:
## lm(formula = VMTpr ~ suburb + exurb + rural + HHVEH0 + HHVEH1 +
##     HHVEH2 + HHVEH3 + HHVEH4 + HHVEH5 + HHVEH6 + HHAGE7, data = YCOUNTY@da
ta)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -28.781  -3.742  -0.633   3.031  49.963
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept) 21.5608253  0.7433166  29.006  < 2e-16 ***
## suburbTRUE   5.0324931  0.6974305   7.216 1.05e-12 ***
## exurbTRUE    2.7945616  0.8729604   3.201  0.00141 **
## ruralTRUE    7.2197938  1.1558705   6.246 6.17e-10 ***
## HHVEH0      -0.0411067  0.0100889  -4.074 4.97e-05 ***
## HHVEH1      -0.0182446  0.0036618  -4.982 7.38e-07 ***
## HHVEH2       0.0005669  0.0036996   0.153  0.87825
## HHVEH3       0.0206943  0.0074417   2.781  0.00552 **
## HHVEH4       0.1141963  0.0155432   7.347 4.15e-13 ***
## HHVEH5      -0.3092137  0.0384675  -8.038 2.51e-15 ***
## HHVEH6       0.1192213  0.0572565   2.082  0.03757 *
## HHAGE7       0.0176199  0.0038043   4.632 4.10e-06 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 5.756 on 1018 degrees of freedom
## Multiple R-squared:  0.4044, Adjusted R-squared:  0.398
## F-statistic: 62.84 on 11 and 1018 DF,  p-value: < 2.2e-16
```

You can play with the following plots just to get a feel for the types of values stored by the objects we estimated

SpaLagres <- resid(SpaLag) # save the residuals

plot(resid(SpaLag), fitted(SpaLag)) # Tukey-Anscombe's plot

qqnorm(SpaLagres) qqline(SpaLagres,col="red")

SARARMixres <- resid(SARARMix) # save the residuals

plot(resid(SARARMix), fitted(SARARMix)) # Tukey-Anscombe's plot

qqnorm(SARARMixres) qqline(SARARMixres,col="red")

plot(VMTprOLSresyesLU,SpaLagres) # comparison of residuals from two OLS models

plot(VMTprOLSresyesLU, SARARMixres) # comparison of residuals from two OLS models

plot(SpaLagres, SARARMixres) # comparison of residuals from two OLS models