

OPEN ACCESS
via Creative Commons 3.0

COLUMN

Spatial Analysis in R: Part 2

Performing spatial regression modeling in R with ACS data

Corey Sparks
Editor, Software & Code

In this second installment of the “Software and Code” posting, I explore the use of the statistical programming environment R for spatial regression modeling. I introduced R in the [previous post](#), including how to get data from a Factfinder 2 query into R, merging it with a TIGER shapefile creating thematic maps and doing some basic exploratory spatial data analysis (ESDA).

In this installment, I show how to use R to estimate commonly used spatial regression models.

Specifically, the simultaneous autoregressive (SAR) model and conditionally autoregressive (CAR) model for a continuous outcome are covered, in addition to the use of specification tests for an ordinary least squares (OLS) model. These models are applied to data from the American Community Survey that were used in the [previous column](#).

Part 1) Read data from the ACS extracts
For this exercise, I’m using the American Factfinder DP2 (Social Characteristics) and DP3 (Economic Characteristics) summary tables for Bexar County, TX census tracts derived from the 2005-2009 5 year ACS summary file. Here is a

screen capture from AFF.

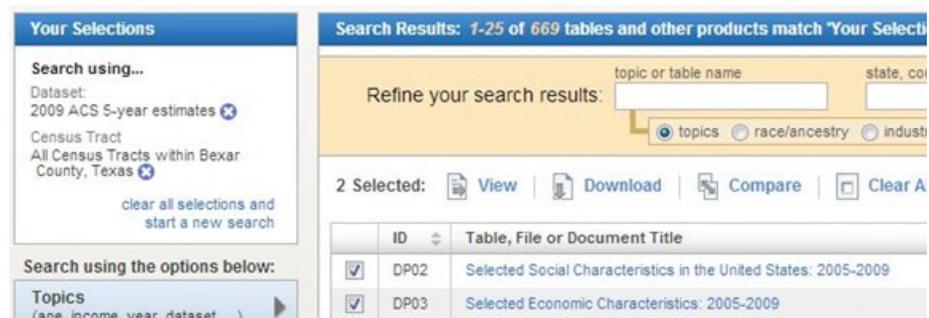
First, I load some libraries that I need, again, this was covered in the [previous column](#).

```
install.packages(c("acs", "spdep",  
"RColorBrewer", "car"), dependencies=T)  
library(acs)  
library(spdep)  
library(RColorBrewer)
```

I had to look at the csv file to find which columns were which, since factfinder doesn't download table numbers.

Here I read in the Social and Economic summary files.

```
social<-read.acs("C:/Users/ozd504/Dropbox/  
spatialDemography/column2data/ACS_09_5Y
```



ID	Table, File or Document Title
<input checked="" type="checkbox"/>	DP02 Selected Social Characteristics in the United States: 2005-2009
<input checked="" type="checkbox"/>	DP03 Selected Economic Characteristics: 2005-2009

```
R_DP5YR2_with_ann.csv", geocols=3:1,
skip=0)
economic<-read.acs("C:/Users/ozd504/
Dropbox/spatialDemography/column2data/AC
S_09_5YR_DP5YR3_with_ann.csv",
geocols=3:1, skip=0)
```

And I extract several variables from each of the files:

```
popsiz<-social@estimate[,171]
fertrate<-social@estimate[,77]
pnohs<-
social@estimate[,118]+social@estimate[,120]
pfornborn<-social@estimate[,184]
pspanspkhh<-social@estimate[,228]
pmarriedwomen<-social@estimate[,64]
popsiz<-social@estimate[,171]
tract<-social@geography$GEO.id2
unemp<-economic@estimate[,17]
medhhinc<-economic@estimate[,125]/1000
poverty<-economic@estimate[,205]
```

And I assemble an R dataframe from these various measures

```
dat<-data.frame(popsiz=popsiz,
fertrate=fertrate, pnohs=pnohs,
pfornborn=pfornborn,
pspanspkhh=pspanspkhh,
pmarriedwomen=pmarriedwomen,
unemp=unemp, medhhinc=medhhinc,
poverty=poverty, tract=tract)
```

and I examine the first few cases using head()

```
head(dat)
I then read the Census tract shapefile
```

```
geodat<-
readShapePoly("C:/Users/ozd504/Dropbox/spa
tialDemography//48_TEXAS/48029_Bexar_Co
unty/tl_2009_48029_tract00.shp",
proj4string=CRS('+proj=longlat
+datum=NAD83')
```

I then need to merge the summary file data to shapefile by the tract identifier. To do this, I set up a temporary data file in R, mdat to store the shapefile attribute table

```
mdat<-geodat@data #temp file for shapefile
attributes
mdat<-merge(x=mdat, y=dat,
by.x="CTIDFP00", by.y="tract", all.x=T, sort=F)
#temp file for merged data
```

In this step is it imperative to not sort the data, otherwise it will make the data file not match the order of the geographies. That's what sort=F does, tells R not to sort the data by "CTIDFP00", or tract ID.

```
geodat@data<-mdat
```

which attaches the merged data to the shapefile again

```
rm(mdat) #remove temp file
```

Here I see which tracts have nonmissing observations for the ACS estimate of the fertility rate, and remove any tracts with missing cases.

```
keep<-!is.na(geodat$fertrate)
geodat<-geodat[keep,]
```

Next, I make a choropleth map of the proportion of the population over age 25 without a high school education and save it as a jpeg image

```
jpeg(filename="C:/Users/ozd504/Dropbox/spa
tialDemography/Col2nohs.jpg", quality=100)
```

```
splot(geodat, "pnohs",
at=quantile(geodat$pnohs, p=c(0,.25, .5, .75,
1), na.rm=T),
col.regions=brewer.pal(5, "Reds"), main=
"Choropleth map of Bexar County", sub="%
Without High School Education")
```

```
dev.off()
```

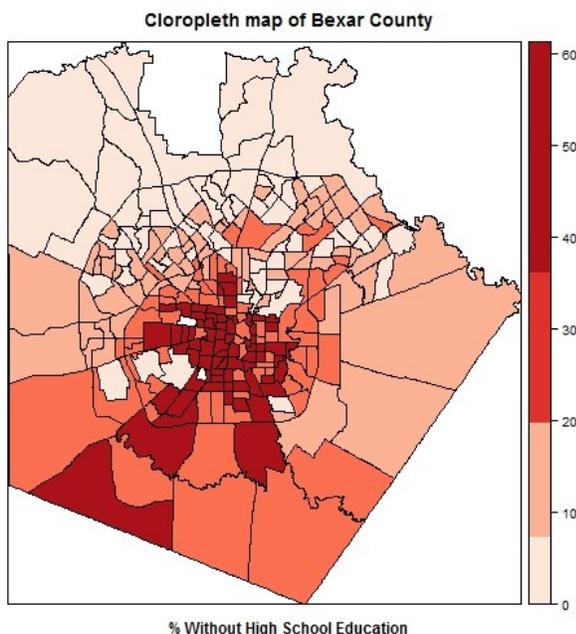
The proportion of the population age 25+ without a high school education shows a strong spatial trend, with areas to the west and south of the central area of the county having lower levels of education (higher proportions).

In the next section, I will fit some Ordinary Least Squares regression models to this education

outcome, and examine the model assumptions.

Part 2) Regression analysis

In this section, I consider the OLS regression model for the education variable shown above, using some social and economic variables as



predictors. This is not meant to be an analysis grounded in theory, but just a simple example of how to fit the models and examine their assumptions. The goal here, is to move from exploratory analysis, to simple regression modeling to more complicated spatial regression models in the final section.

First, I simply fit an OLS model to the education variable using the % foreign born (pfornborn), % of married women (pmarriedwomen), proportion of Spanish speaking households (pspanspkhh), the unemployment rate (unemp), poverty rate (poverty) and the log of the median household income (medhhinc). I store the model fit in an object called "fit.1"

```
fit.1<-
lm(pnohs~pfornborn+pmarriedwomen+pspans
pkhh+unemp+poverty+log(medhhinc),
data=geodat)
```

The `summary()` function will report model fit statistics and tests of the regression coefficients.

```
summary(fit.1)
```

```
Call:
lm(formula = pnohs ~ pfornborn + pmarriedwomen
+ pspanspkhh + unemp + poverty + log(medhhinc),
data = geodat)
```

```
Residuals:
Min      1Q  Median      3Q      Max
-18.779 -3.850  0.218  3.676  24.004
```

```
Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept) 19.80358    6.65991  2.974 0.00321 **
pfornborn   -0.01824    0.07331 -0.249 0.80366
pmarriedwomen 0.08460    0.04023  2.103 0.03640 *
pspanspkhh   0.49778    0.02858 17.419 < 2e-16 ***
unemp        0.31807    0.10458  3.041 0.00259 **
poverty      0.16450    0.05123  3.211 0.00148 **
log(medhhinc) -7.27924    1.67950 -4.334 2.07e-05 ***
```

```
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
Residual standard error: 5.727 on 269 degrees of freedom
Multiple R-squared:  0.8746
Adjusted R-squared:  0.8718
F-statistic: 312.7 on 6 and 269 DF, p-value: < 2.2e-16
```

The summary function shows some statistics on the model residuals, but most importantly we see tests for the model coefficients. In this case, there is a significant positive associations between education level and pmarried women, pspanspkhh, unemp and poverty. This suggests that higher values of the these variables are associated with higher proportions of the population without a high school education. The log of income also shows a negative relationship, which suggests in areas with higher income level, the outcome is lower. Overall the model is fitting the data fairly well, with an adjusted model R² of 87%.

Next, I display some model graphical diagnostics for the model. I plot the histogram of the outcome to examine it for normality, the histogram of the model studentized residuals and several default plots R produces regarding constancy of residudual variance

(homoscedasticity) and normality of model residuals. I put all of this into a single plot with 6 sub-plots and save it as a jpeg image.

```
jpeg(filename="C:/Users/ozd504/Dropbox/spatialDemography/Col2Mod1.jpg", quality=100)
par(mfrow=c(3,2)) #divide the plot into 3 rows and 2 columns
hist(dat$pnohs, main="Distribution of Outcome")
hist(rstudent(fit.1), main="Model Residuals")
plot(fit.1)
dev.off()
```

The outcome is certainly not normally distributed, but that's ok, the model doesn't assume it is, only that the residuals are normal, which they resemble very closely, given the residual Q-Q plot in the second row. There is some concern about heteroskedasticity, since the residual vs. fitted plots show a slight trend, so a formal test is probably warranted.

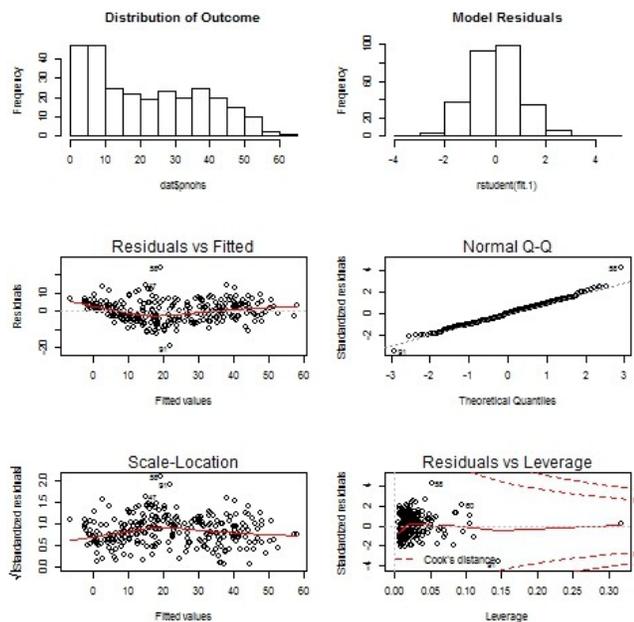
Breusch-Pagan test for constant variance

```
bptest(fit.1)
```

studentized Breusch-Pagan test

data: fit.1

BP = 34.1131, df = 6, p-value = 6.398e-06



Which suggests that we have non-constant error variance, we may consider a weighted least squares model now, with population size as the weighting factor. I call this fit.1wt

```
fit.1wt<-
lm(pnohs~pfornborn+pmarriedwomen+pspanspkhh+unemp+poverty+log(medhhinc),weights=popsize, data=geodat)
summary(fit.1wt)
```

Call:

```
lm(formula = pnohs ~ pfornborn + pmarriedwomen + pspanspkhh + unemp + poverty + log(medhhinc), data = geodat, weights = popsize)
```

Weighted Residuals:

```
Min      1Q  Median    3Q     Max
-1000.61 -262.98  13.89  294.45 1263.11
```

Coefficients:

```
              Estimate Std. Error t value Pr(>|t|)
(Intercept) 13.45904   7.75056   1.737 0.083616 .
pfornborn    0.02537   0.07481   0.339 0.734777
pmarriedwomen 0.10069   0.04937   2.040 0.042363 *
pspanspkhh   0.47377   0.03057  15.499 < 2e-16 ***
unemp        0.44216   0.11597   3.813 0.000170 ***
poverty      0.21294   0.05705   3.732 0.000231 ***
log(medhhinc) -6.14712   2.07489  -2.963 0.003323 **
```

Signif. Codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 405.6 on 269 degrees of freedom

Multiple R-squared: 0.8845,

Adjusted R-squared: 0.8819

F-statistic: 343.3 on 6 and 269 DF, p-value: < 2.2e-16

Which shows the same associations, but now let's see if we've corrected the heteroskedasticity.

```
bptest(fit.1wt)
```

studentized Breusch-Pagan test

data: fit.1wt

BP = 34.1131, df = 6, p-value = 6.398e-06

Which suggests we haven't. An alternative procedure is to correct the variance-covariance matrix of the OLS model using White's correction. This is implemented easily in the Anova() function in the car package.

```
Anova(fit.1, white.adjust=T)
```

Analysis of Deviance Table (Type II tests)

Response: pnohs

	Df	F	Pr(>F)
pforborn	1	0.0684	0.793935
pmarriedwomen	1	3.6465	0.057250 .
pspanspkhh	1	220.9537	< 2.2e-16 ***
unemp	1	3.3461	0.068472 .
poverty	1	7.8399	0.005481 **
log(medhhinc)	1	15.7775	9.153e-05 ***

Residuals 269

Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Which shows that the pmarriedwomen and unemp variables become only marginally significant after correcting for heteroskedasticity.

Now we examine the model residuals for spatial autocorrelation, first by mapping them then by formal testing using Moran's I.

```
geodat$residfit1<-rstudent(fit.1)
Which adds the studentized residuals to the
shapefile's attribute table
```

```
cols<-brewer.pal(7,"RdBu")
```

makes a set of colors derived from ColorBrewer, corresponding to the Red to Blue diverging scheme. Then I make a jpeg image of the choropleth map of the residuals.

```
jpeg(filename="C:/Users/ozd504/Dropbox/spatialDemography/Col2Res1.jpg", quality=100)
spplot(geodat,"residfit1",
at=quantile(geodat$residfit1), col.regions=cols,
main="Residuals from OLS Fit")
dev.off()
```

Which may have some spatial clustering of residuals, but a test is prudent here.

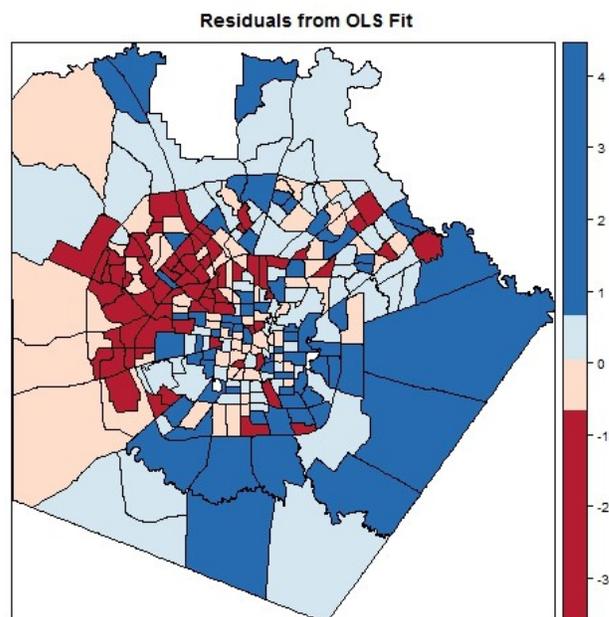
Again, we make our row-standardized spatial weight matrix using a Queen contiguity rule.

```
sa.nb<-poly2nb(geodat, queen=T)
sa.wt<-nb2listw(sa.nb, style="W")
```

Then we test for autocorrelation in the residuals using the lm.morantest() function.

```
lm.morantest(fit.1, listw=sa.wt)
```

Global Moran's I for regression residuals



```
data:
model: lm(formula = pnohs ~ pforborn +
pmarriedwomen + pspanspkhh + unemp + poverty
+ log(medhhinc), data = geodat)
weights: sa.wt
```

Moran I statistic standard deviate = 10.9577, p-value < 2.2e-16

alternative hypothesis: greater

sample estimates:

Observed Moran's I	Expectation	Variance
0.355403117	-0.009514177	0.001109054

Which shows an I value of .355, and a significant test statistic based on a z-test.

Based on the previous post, a local Moran map may also be prudent. So I do a local Moran I, then save the values of the local z statistic to the shapefile.

```
geodat$lmfit1<-localmoran(geodat$residfit1,
sa.wt)[,"Z.I"]
```

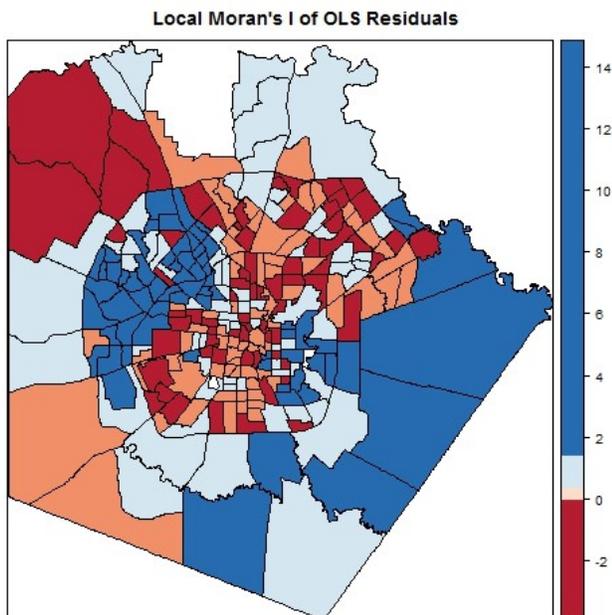
```
jpeg(filename="C:/Users/ozd504/Dropbox/spatialDemography/Col2LocalM.jpg", quality=100)
spplot(geodat, "lmfit1",
  at=quantile(geodat$lmfit1),
  col.regions=brewer.pal(6, "RdBu"),
  main="Local Moran's I of OLS Residuals")
dev.off()
```

Which shows large positive z values toward the northwestern and the southeastern part of the county, suggesting pockets of autocorrelation in the model residuals.

In the next section, I move to full spatially specified regression models.

Part 3) Spatial regression models

In this section, I use some commonly used spatial regression models to extend the analysis presented above. Specifically I consider three model specifications from the spatial econometrics literature. All model specifications follow those presented in LeSage and Pace



(2009), and interested readers should consult that work for these and other model

specifications. The first model considered is called the spatial lag model, specified as:

$$y = \rho W y + X' \beta + \varepsilon$$

where the spatial component ($\rho W y$) is specified on the model intercept. In doing so, the model's intercept is lagged across neighbors. This model specifies the spatial effects as a diffusion process, where neighboring county values of the outcome influence the values of other counties that border them.

The second model considered is the spatial error model. This model is defined by adding a spatial structure term to the OLS model's residuals, ε :

$$y = X' \beta + \varepsilon$$

$$\varepsilon = \rho W \varepsilon + u$$

This model specification essentially says that all autocorrelation is confined to the error term in the model, which can be written in two parts: the spatially structure residual, ε , and the random residual, u , which are random and homoskedastic. The cause of such residual autocorrelation is typically thought to arise from the exclusion of an unobserved endogenous spatially structured covariate that, were it measurable, would explain the spatial autocorrelation in the residuals. The parameter ρ measures the strength of the autoregressive effect on the model residuals amongst neighboring observations (Anselin, 2002; Anselin & Bera, 1998; Chi & Zhu, 2008).

The final model specification is the spatial Durbin model which has a similar structure to the lag model, except that it also includes lagged spatial covariates in the linear predictor for the model. The model is written:

$$y = \rho W y + X' \beta + W X' \theta + \varepsilon$$

Here, the spatial lag model is fit to the outcome:

```
fit.lag<-
lagsarlm(pnohs~pforborn+pmarriedwomen+p
spanpkhh+unemp+poverty+log(medhhinc),
data=geodat, listw=sa.wt, type="lag")
```

And I ask for the model summary, including the model pseudo-R².

```
summary(fit.lag, Nagelkerke=T)
```

```
Call:lagsarlm(formula = pnohs ~ pforborn +
pmarriedwomen + pspanspkhh + unemp + poverty
+ log(medhhinc), data = geodat, listw = sa.wt, type =
"lag")
```

```
Residuals:
  Min      1Q  Median      3Q      Max
-18.06178 -3.55777 -0.82312  3.23268  21.19234
```

```
Type: lag
Coefficients: (asymptotic standard errors)
              Estimate Std. Error z value Pr(>|z|)
(Intercept) 11.494885  6.084607  1.8892 0.0588684
pforborn     0.033713  0.066152  0.5096 0.6103108
pmarriedwomen 0.078430  0.036114  2.1717
0.0298750
pspanspkhh   0.354310  0.034366 10.3099 < 2.2e-16
unemp        0.199106  0.094502  2.1069 0.0351266
poverty      0.169628  0.046001  3.6875 0.0002265
log(medhhinc) -5.201036  1.534324 -3.3898
0.0006995
Rho: 0.31608, LR test value: 47.785, p-value:
4.7552e-12
Asymptotic standard error: 0.045969
z-value: 6.8759, p-value: 6.1606e-12
Wald statistic: 47.278, p-value: 6.1606e-12
```

```
Log likelihood: -845.862 for lag model
ML residual variance (sigma squared): 26.429,
(sigma: 5.1409)
Nagelkerke pseudo-R-squared: 0.89455
Number of observations: 276
Number of parameters estimated: 9
AIC: 1709.7, (AIC for lm: 1755.5)
LM test for residual autocorrelation
test value: 29.146, p-value: 6.713e-08
```

Again, we see the same significant predictors as for the original linear model, but this model fits the data much better, judging by the AIC score of 1709, compared to the AIC from the OLS model of 1755. However, there is still residual spatial autocorrelation in the model residuals, which suggests an error process may still be at work. Next the spatial error model is considered:

```
fit.err<-
errorsarlm(pnohs~pforborn+pmarriedwomen
+pspanspkhh+unemp+poverty+log(medhhinc),
data=geodat, listw=sa.wt)
```

```
summary(fit.err, Nagelkerke=T)
```

```
Call:errorsarlm(formula = pnohs ~ pforborn +
pmarriedwomen + pspanspkhh + unemp + poverty
+ log(medhhinc), data = geodat, listw = sa.wt)
```

```
Residuals:
  Min      1Q  Median      3Q      Max
-11.61767 -3.29728 -0.11333  2.76648  19.31390
```

```
Type: error
Coefficients: (asymptotic standard errors)
              Estimate Std. Error z value Pr(>|z|)
(Intercept) 19.958746  5.811857  3.4341 0.0005944
pforborn     0.028955  0.065744  0.4404 0.6596346
pmarriedwomen 0.075968  0.032967  2.3044
0.0212025
pspanspkhh   0.486704  0.031880 15.2666 < 2.2e-16
unemp        0.120782  0.084026  1.4374 0.1505928
poverty      0.139543  0.041052  3.3992 0.0006759
log(medhhinc) -6.810564  1.400634 -4.8625 1.159e-
06
```

```
Lambda: 0.67642, LR test value: 83.465, p-value: <
2.22e-16
Asymptotic standard error: 0.059939
z-value: 11.285, p-value: < 2.22e-16
Wald statistic: 127.36, p-value: < 2.22e-16
```

```
Log likelihood: -828.0222 for error model
ML residual variance (sigma squared): 21.496,
(sigma: 4.6363)
Nagelkerke pseudo-R-squared: 0.90734
Number of observations: 276
Number of parameters estimated: 9
AIC: 1674, (AIC for lm: 1755.5)
```

Which shows an even larger change in the AIC compared to the OLS model. Finally the spatial Durbin model is fit:

```
Call:lagsarlm(formula = pnohs ~ pforborn +
pmarriedwomen + pspanspkhh + unemp + poverty
+ log(medhhinc), data = geodat, listw = sa.wt, type
="mixed")
```

```
Residuals:
  Min      1Q  Median      3Q      Max
-12.21950 -3.15128 -0.23834  2.73831  17.36923
```

```
Type: mixed
Coefficients: (asymptotic standard errors)
              Estimate Std. Error z value Pr(>|z|)
(Intercept) -7.3907803 11.4695259 -0.6444
0.5193264
pforborn     0.0318343  0.0652047  0.4882
0.6253932
pmarriedwomen 0.0704654  0.0338245  2.0833
0.0372271
```

```

pspanspkhh    0.4718467 0.0352451 13.3876 <
2.2e-16
unemp         0.1682672 0.0854116 1.9701
0.0488299
poverty       0.1402400 0.0413830 3.3888
0.0007019
log(medhhinc) -6.0782687 1.4350322 -4.2356
2.279e-05
lag.pfornborn -0.1246679 0.1249638 -0.9976
0.3184577
lag.pmarriedwomen -0.0655288 0.0664247
-0.9865 0.3238817
lag.pspanspkhh -0.2793182 0.0599507 -4.6591
3.175e-06
lag.unemp     0.5146267 0.2052233 2.5076
0.0121540
lag.poverty   0.0092553 0.1014463 0.0912
0.9273072
lag.log(medhhinc) 6.6579107 3.0070977 2.2141
0.0268243

```

Rho: 0.57705, LR test value: 57.508, p-value: 3.364e-14

Asymptotic standard error: 0.068083

z-value: 8.4756, p-value: < 2.22e-16

Wald statistic: 71.837, p-value: < 2.22e-16

Log likelihood: -819.8694 for mixed model

ML residual variance (sigma squared): 20.88,

(sigma: 4.5695)

Nagelkerke pseudo-R-squared: 0.91265

Number of observations: 276

Number of parameters estimated: 15

AIC: 1669.7, (AIC for lm: 1725.2)

LM test for residual autocorrelation

test value: 1.4222, p-value: 0.23304

The lagged predictors all have lag. In front of their names in this output, and we see that the lagged versions of pspanspkhh, unemployment and medhhinc all have significant effects. The AIC value for this model is only slightly different from the lag model, but the residual autocorrelation has been removed by considering the lagged predictors.

Summary

In this column, I have shown how to evaluate the assumptions of the OLS regression model and implement three commonly used spatial regression model specifications to a data set derived from the American Community Survey summary file. While these are certainly not the only spatial regression specifications, they do constitute a core set of models that are used in spatial demographic work. Future contributions will consider Bayesian methods and spatial filtering methods to accommodate spatial heterogeneity and autocorrelation into the regression framework.

References

- Anselin, L. (2002). Under the hood: Issues in the specification and interpretation of spatial regression models. *Agricultural Economics*, 27, 247-267.
- Anselin, L., & Bera, A. K. (1998). Spatial dependence in linear regression models with an introduction to spatial econometrics. In A. Ullah & D. E. A. Giles (Eds.), *Handbook of Applied Economic Statistics* (pp. 237-289). New York: Marcel Dekker.
- Chi, G., & Zhu, J. (2008). Spatial regression models for demographic analysis. *Population Research and Policy Review*, 27(1), 17-42. doi: Doi 10.1007/S11113-007-9051-8.
- LeSage, J., & Pace, R. K. (2009). *Introduction to Spatial Econometrics*: CRC Press.