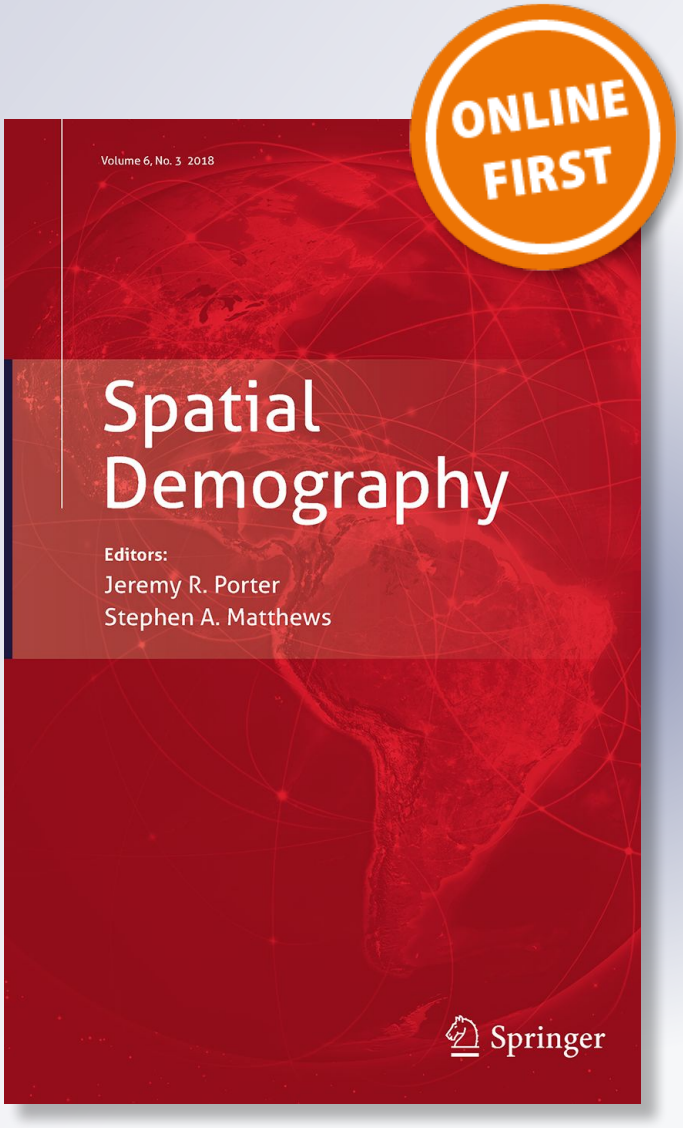
Spatial Regression Analysis of Poverty in R

Maria Kamenetsky, Guangqing Chi, Donghui Wang & Jun Zhu

Spatial Demography

ISSN 2364-2289

Spat Demogr
DOI 10.1007/s40980-019-00048-0



Springer

Spatial Demography
https://doi.org/10.1007/s40980-019-00048-0

RESEARCH NOTE

Spatial Regression Analysis of Poverty in R

Maria Kamenetsky[1] · Guangqing Chi[2] · Donghui Wang[3] · Jun Zhu[4]

Abstract

Poverty has been studied across many social science disciplines, resulting in a large body of literature. Scholars of poverty research have long recognized that the poor are not uniformly distributed across space. Understanding the spatial aspect of poverty is important because it helps us understand place-based structural inequalities. There are many spatial regression models, but there is a learning curve to learn and apply them to poverty research. This manuscript aims to introduce the concepts of spatial regression modeling and walk the reader through the steps of conducting poverty research using R: standard exploratory data analysis, standard linear regression, neighborhood structure and spatial weight matrix, exploratory spatial data analysis, and spatial linear regression. We also discuss the spatial heterogeneity and spatial panel aspects of poverty. We provide code for data analysis in the R environment and readers can modify it for their own data analyses. We also present results in their raw format to help readers become familiar with the R environment.

Keywords Poverty · Exploratory spatial data analysis · Spatial regression · R

✉ Guangqing Chi
gchi@psu.edu

[1] Department of Population Health Sciences, University of Wisconsin-Madison, 610 Walnut Street, Madison, WI 53726, USA

[2] Department of Agricultural Economics, Sociology, and Education, Population Research Institute, and Social Science Research Institute, The Pennsylvania State University, 112E Armsby, University Park, PA 16802, USA

[3] Center on Contemporary China, Princeton University, 359 Wallace Hall, Princeton, NJ 08544, USA

[4] Department of Statistics, University of Wisconsin-Madison, 1300 University Avenue, Madison, WI 53706, USA

Published online: 04 March 2019

1 Introduction

Throughout human history, poverty has been associated with many social problems and historic events, including inequality, wars, and revolutions. Even in developed countries, poverty persists. Thus, it is not a surprise that poverty is a topic that has been studied across many social science disciplines, generating a large body of literature. Poverty, like many other social phenomena, can be *spatial* and scholars of poverty research have long recognized that the poor are not uniformly distributed across space (Nord et al. 1995; Thiede et al. 2018; Voss et al. 2006; Weber et al. 2005). Understanding the spatial distribution of poverty helps us to understand place-based structural inequalities (Lobao et al. 2008; Tickamyer and Duncan 1990). This school of research is also referred to as the study of "place poverty," in contrast with "people poverty," as the former emphasizes structural and contextual forces while the latter emphasizes individual or family forces (Voss et al. 2006).

The spatial aspect of poverty has been increasingly studied using formal spatial analysis and spatial regression models (e.g., Curtis et al. 2018; Curtis, Voss, and Long 2012). While the methods are many, there is a learning curve to learn and apply them to poverty research. This manuscript aims to introduce the concepts of spatial regression modeling and walk the reader through the steps of conducting poverty research using R, an open-source statistical software that is gaining increasing popularity among social scientists (R Development Core Team 2008). Similar to other teaching notes (Sparks 2013a, b), we provide code for data analysis in the R environment and readers can modify it for their own data analyses. We also present results in their raw format to help readers become familiar with the R environment. Poverty research has been reviewed in several articles and books (e.g., Iceland 2013; Jennings 1999; Sandoval et al. 2009; Weber et al. 2005) and thus is not reviewed in this manuscript. Readers are suggested to refer to these review articles for the status of poverty research.

In the following sections, we introduce our data (Sect. 2), walk the readers through the steps of exploratory data analysis (Sect. 3), standard linear regression (Sect. 4), neighborhood structure and spatial weight matrix (Sect. 5), exploratory spatial data analysis (Sect. 6), and spatial linear regression (Sect. 7), all in R, and discuss other spatial aspects of poverty that could be addressed (Sect. 8).

2 Data and Units of Analysis

This manuscript provides an example in R of quantifying the spatial relationship between county-level poverty rates and several socioeconomic factors in the 48 states of the contiguous United States. Among the existing place poverty studies, county is often used as the unit of analysis; counties are salient units in policy making and planning perspectives such that many policy decisions potentially relevant to poverty rates are made at the county level (Greenlee and Howe 2009;

Lichter and Johnson 2007; Thiede et al. 2018; Voss et al. 2006). Moreover, compared with the boundaries of other administrative units, county boundaries are subject to little boundary change, therefore facilitating scholarly study of poverty trends over time. Studies have found that county-level poverty is associated with many other structural disadvantages, especially when it comes to some major health indicators, such as cancer stage (Greenlee and Howe 2009), obesity (Bennett et al. 2011), and HIV prevalence (Vaughan et al. 2014). Commonly identified factors associated with the spatial concentration of county-level poverty rates include economic structure (Goetz and Swaminathan 2006; Lobao et al. 2008), racial composition (Thiede et al. 2018; Wimberley and Morris 2002), and human capital stock (Levernier et al. 2000).

The variable of interest is poverty (*povty*), measured as the percentage of individuals age 18–64 living in poverty in a county in year 2000. We include a set of economic, social, and demographic factors that may relate with county-level poverty rates. Specifically, three variables, agriculture (*ag*), manufacturing (*manu*), and retail (*retail*), are the percentages of workers in the agricultural sector, the manufacturing sector, and the retail sector, respectively. For socially-related factors, *foreign* is the percentage of the foreign-born population in a county and female employment (*feemp*) is the percentage of female employment in the total population. Human capital stock is captured by high school (*hsch*), which indicates the percentage of the population that completed a high school education. Lastly, we use percentage Blacks (*black*) and percentage Hispanics (*hisp*) to capture racial and ethnic compositions. These variables are measured in 2000 based on the Decennial Census data. We download the data using SocialExplorer, an online demographic research tool that pre-processed the raw Census data and reports calculated percentages by county. While other variables have been used in poverty research, in this manuscript we focus on these selected variables for the purpose of illustrating spatial regression modeling for poverty research in R. We have made the dataset (uspoverty2000.csv) and supplementary tutorials available at https://mkamenet3.github.io/SpatialRegPovertyR/.

3 Standard Exploratory Data Analysis

After initial installation, we first load the R libraries that the analysis needs.

```
library(sp)
library(spdep)
library(ggplot2)
library(tidyverse)
library(tigris)
library(dplyr)
library(sf)
library(readr)
library(car)
```

After the census data are downloaded and the variables of interest are extracted, we streamline the data and save them to a proper data file to be read into R as a data frame before performing statistical data analysis. We use the `read_csv()` function from the `readr` package (Wickham et al. 2018a) to import the csv file into R and to specify that the variable `FIPS_N` is imported as a character variable (`col_types = cols(FIPS_N = col_character())`) and that the import does not omit leading zeros from FIPS_N variable (`trim_ws = FALSE`). We immediately pipe (`%>%`) to the `arrange()` function from the `dplyr` package (Wickham et al. 2018b) to sort by `FIPS_N`.

```
povdf = read_csv("uspoverty2000.csv",
                 col_types = cols(FIPS_N = col_character()),
                 trim_ws = FALSE) %>%
    arrange(as.numeric(as.character(FIPS_N)))
nrow(povdf)
```

```
## [1] 3070
```

The counties are identified by their five-digit federal information processing (FIP) standards code, with the first two digits corresponding to a state code and the last three digits corresponding to a county code. The data frame `povdf` features the poverty rate and the socioeconomic variables from the census year 2000 in the 3070 counties of the 48 contiguous states. Using the R function `head()`, we view the first six counties in the data frame `povdf`:

```
head(povdf)
```

```
## # A tibble: 6 x 12
##   FIPS_N    ag black feemp foreign  hisp  hsch  manu povty retail STATEFP
##   <chr>  <dbl> <dbl> <dbl>   <dbl> <dbl> <dbl> <dbl> <dbl>  <dbl>   <dbl>
## 1 01001   2.13 17.1   53.2   1.17   1.40  33.8  16.5  9.06  12.9        1
## 2 01003   1.57 10.3   50.5   2.11   1.76  29.6  12.5  9.13  14.2        1
## 3 01005   3.97 46.3   43.5   1.5    1.65  32.4  31.4 21.9   10.5        1
## 4 01007   3.30 22.2   44.3   0.430  1.01  35.7  23.8 17.7    9.93       1
## 5 01009   2.18  1.19  48.1   3.10   5.33  36.0  19.5  9.89  11.6        1
## 6 01011  10.5  73.1   36.5   3.06   2.75  35.2  23.2 28.6    7.53       1
## # ... with 1 more variable: COUNTYFP <chr>
```

To perform exploratory data analysis, we use summary statistics and graphical methods. Most commonly used summary statistics and graphical methods for exploratory data analysis are applied to either one variable at a time (i.e., univariate) or two variables at a time (i.e., bivariate). Using the R function `summary()`, we obtain several univariate summary statistics for each of the variables in the data frame `povdf`:

```
summary(povdf)
```

```
##     FIPS_N                ag              black            feemp
##  Length:3070        Min.   : 0.030   Min.   : 0.000   Min.   :23.21
##  Class :character   1st Qu.: 1.680   1st Qu.: 0.290   1st Qu.:46.94
##  Mode  :character   Median : 3.790   Median : 1.675   Median :51.88
##                     Mean   : 6.129   Mean   : 8.710   Mean   :51.64
##                     3rd Qu.: 7.810   3rd Qu.: 9.865   3rd Qu.:56.30
##                     Max.   :55.600   Max.   :86.490   Max.   :78.47
##    foreign            hisp             hsch            manu
##  Min.   : 0.000   Min.   : 0.080   Min.   :10.93   Min.   : 0.000
##  1st Qu.: 0.890   1st Qu.: 0.910   1st Qu.:30.86   1st Qu.: 8.945
##  Median : 1.710   Median : 1.780   Median :34.91   Median :14.965
##  Mean   : 3.419   Mean   : 6.230   Mean   :34.80   Mean   :15.938
##  3rd Qu.: 3.900   3rd Qu.: 5.107   3rd Qu.:38.99   3rd Qu.:21.997
##  Max.   :50.940   Max.   :97.540   Max.   :53.25   Max.   :48.550
##      povty           retail         STATEFP        COUNTYFP
##  Min.   : 2.04   Min.   : 1.72   Min.   : 1.00   Length:3070
##  1st Qu.: 8.51   1st Qu.:10.30   1st Qu.:19.00   Class :character
##  Median :11.61   Median :11.56   Median :29.00   Mode  :character
##  Mean   :12.69   Mean   :11.48   Mean   :30.36
##  3rd Qu.:15.70   3rd Qu.:12.72   3rd Qu.:45.00
##  Max.   :53.83   Max.   :26.90   Max.   :56.00
```

For a continuous variable, the summary statistics are its minimum, first quartile, median, mean, third quartile, and maximum. The summary statistics of the response variable of poverty rate show that the lowest county-level poverty rate was 2.04% and the highest county-level poverty rate was 53.83% among the 3070 counties in the contiguous United States in 2000. The center of the county-level poverty rates is 11.61% measured by median and 12.69% measured by mean (i.e., average). In addition, the interquartile range is between the first quartile of 8.51% and the third quartile of 15.70%. That is, half of the counties had poverty rates below 11.61% and the other half had poverty rates above 11.61%, while the middle half of the counties had poverty rates between 8.51 and 15.70%. Among the explanatory variables, we use female employment rate as an example. The summary statistics of the female employment rate tell us that county-level female employment rates ranged between the lowest at 23.21% and the highest at 78.47% among the 3070 counties in the contiguous United States in 2000. Half of the counties had female employment rates below the median of 51.88% and the other half had female employment rates above 51.88%, while the middle half of the counties had female employment rates between 46.94 and 56.30%.

Using the R function `cor()`, we obtain the sample correlation, a bivariate summary statistic, between poverty rate and any given socioeconomic factor. For example, the correlation between poverty rate and female employment rate is:

```
cor(povdf$povty, povdf$feemp)
```

```
## [1] -0.6900635
```

The sign of this sample correlation is negative, indicating a negative correlation between the poverty and female employment rates. That is, higher female employment rates are associated with lower poverty rates, whereas lower female employment rates are associated with higher poverty rates. The magnitude of this correlation is 0.69 (rounded from 0.6900635) and reflects a moderate amount of association between the poverty and female employment rates for such an observational study in the social sciences.

Among the many graphical methods used for exploratory data analysis, in this manuscript we demonstrate two often-used graphs, one univariate and the other bivariate. In particular, we draw a histogram and a scatter plot by the `ggplot2` R package (Wickham 2016), using the geometries, `geom_histogram()` and `geom_point()`, respectively:

```
ggplot(povdf, aes(x=povty)) +
    geom_histogram(aes(y=..density..),fill="grey", col="black") +
    theme_bw() +
    xlab("Poverty Rate") + ylab("Density")

ggplot(povdf, aes(x=feemp, y=povty)) +
    geom_point(size = 0.5) +
    theme_bw() +
    xlab("Female Employment Rate") + ylab("Poverty Rate")
```

Figure 1 shows the histogram of *povty* and the scatter plot for *povty* by *feemp*. The histogram shows that the range of poverty rates is between 0 and 55% with a center around 10% (Fig. 1a). The histogram is also right skewed, revealing counties with high poverty rates in the right tail. The histogram is based on density such that the areas of the vertical bars at 5% increments add up to a total probability of 1. Alternatively, we could plot the histogram by the number of counties at the 5% increments and the shape of the histogram would be the same.

The scatter plot shows a negative trend (Fig. 1b). As the female employment rate increases from 20% to nearly 80%, the poverty rate declines. This finding is consistent with the negative sample correlation, indicating a negative association between female employment rate and poverty rate.

4 Standard Linear Regression

4.1 Model Fitting

To quantify the relationships between the poverty rate and the socioeconomic variables, we perform standard linear regression such that the response variable is poverty rate and the eight explanatory variables are percentage of agricultural workers (*ag*), percentage of manufacturing workers (*manu*), percentage of retail workers (*retail*), percentage of foreign-born residents (*foreign*), percentage of female employment (*feemp*), percentage of high school graduates (*hsch*), percentage of Blacks (*black*), and percentage of Hispanics (*hisp*).

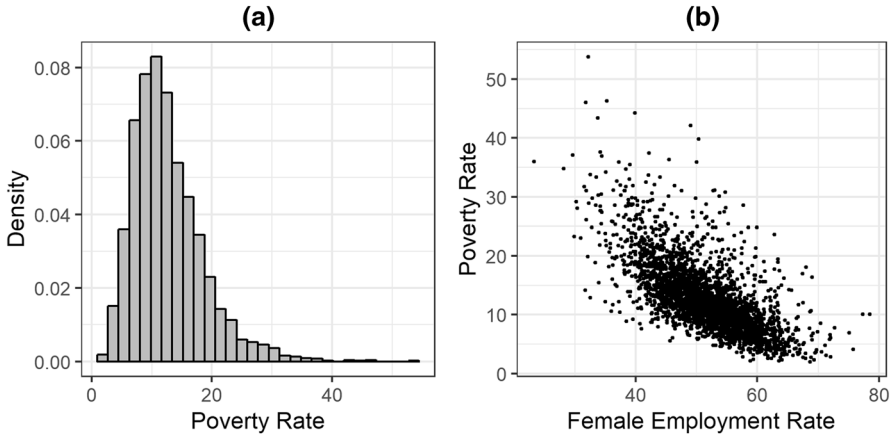


Fig. 1 Histogram of poverty (**a**) and scatter plot between the poverty rate and the female employment rate (**b**)

The analysis in the remainder of this note focuses on interpretation and relies on the large sample size of exploring all counties in the continental United States (3070 counties). Given that our response ranges from 0 to 100% and that the models in this analysis are linear models, it is possible that fitted values may lie outside of the range leading to predictions of negative poverty or poverty over 100%. In the interest of interpretability and because consideration of appropriate transformations depends on a case by case basis, we do not use any transformations to address this. For smaller area studies, we encourage the reader to consider linear transformation of the response (ex: logit, log, square root, arcsine) in order to meet the normality assumptions imposed by these models. Weighting by the population size may also be explored. Supplementary tutorials can be found at https://mkamenet3.github.io/SpatialRegPovertyR/.

The R function `lm()` is applied to the data frame *povty* and the output `m1` is an `lm` object. We then apply the R function `summary()` to the `m1` object and obtain the results of the standard linear regression:

```
m1 = lm(povty ~ ag + manu + retail + foreign + feemp + hsch + black + hisp,
data=povdf)
summary(m1)

##
## Call:
## lm(formula = povty ~ ag + manu + retail + foreign + feemp + hsch +
##     black + hisp, data = povdf)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -12.9991  -2.2388  -0.3421   1.7029  29.0205
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) 46.462367   1.019860  45.558  < 2e-16 ***
## ag           0.104110   0.012701   8.197 3.58e-16 ***
## manu        -0.028851   0.009740  -2.962  0.00308 **
## retail      -0.064335   0.039627  -1.624  0.10458
## foreign     -0.171651   0.021274  -8.069 1.01e-15 ***
## feemp       -0.526357   0.011166 -47.141  < 2e-16 ***
## hsch        -0.188037   0.014749 -12.749  < 2e-16 ***
## black        0.081289   0.005466  14.872  < 2e-16 ***
## hisp         0.061865   0.008176   7.567 5.03e-14 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 3.809 on 3061 degrees of freedom
## Multiple R-squared:  0.5861, Adjusted R-squared:  0.585
## F-statistic: 541.7 on 8 and 3061 DF,  p-value: < 2.2e-16
```

There are four parts to the summary of this standard linear regression. The initial function call is echoed in the first part. The second part reports the summary statistics of residuals (minimum, first quartile, median, third quartile, and maximum). Here *the residual is defined as the difference between an observed response (poverty rate) and its fitted value by the linear regression*. We use residuals for model diagnostics near the end of this subsection. The third part reports, for each explanatory variable, a fitted regression coefficient (under `Estimate`), its standard errors (under `Std. Error`), the ratio of the two values as a T test statistic (under `t value`), and a p value for testing whether the true regression coefficient is zero or not (under `Pr(>|t|)`). For example, the estimated regression coefficient for `feemp` is -0.526 (rounded from -0.52635676) with standard error 0.011166. The T test statistic is -47.141 and the p value is $<2.2\text{E-}16$. This tells us that a 1% increase in the female employment rate is associated with a 0.526% decrease in the poverty rate when all other explanatory variables are held constant. How significant is this result? Relative to the standard error of about 1.12%, the T statistic is very large and the p value is extremely small. There is very strong evidence that the true regression coefficient for female employment rate is not zero. The last part of the output provides a residual standard 3.809 on 3061 degrees of freedom, which estimates the standard deviation of the

error term in the standard linear regression model. In addition, a multiple and an adjusted R-squared are reported, indicating that about 58% of the variation in the response variable of poverty rate is explained by the relationship with the socioeconomic variables considered herein. Finally, an F-test is carried out for testing whether all of the regression coefficients are zero. The p value is $<2.2\text{E}{-}16$ and there is very strong evidence that the not all the regression coefficients are zero.

In addition, we may extract the estimated regression coefficients and the corresponding 95% confidence intervals by applying the R functions `coef()` and `confint()` to the `m1` object. By the R function `cbind()` below, we combine these results into a table of three columns, one for the estimated regression coefficients (named `coefest`) and the other two for the lower and upper limits of the 95% confidence intervals.

```
cbind(coefest = coef(m1), confint(m1))
```

```
##                 coefest       2.5 %       97.5 %
## (Intercept) 46.46236736 44.46268723 48.462047496
## ag           0.10410987  0.07920740  0.129012335
## manu        -0.02885059 -0.04794830 -0.009752886
## retail      -0.06433542 -0.14203379  0.013362963
## foreign     -0.17165122 -0.21336404 -0.129938403
## feemp       -0.52635676 -0.54824943 -0.504464081
## hsch        -0.18803678 -0.21695614 -0.159117430
## black        0.08128912  0.07057202  0.092006219
## hisp         0.06186454  0.04583405  0.077895030
```

For example, the estimated regression coefficient for `feemp` is -0.526 with a 95% confidence interval of $[-0.548, -0.504]$. That is, there is a 95% confidence of between 0.504 and 0.548% decrease in the poverty rate associated with a 1% increase in the female employment rate when all other explanatory variables are held constant.

4.2 Model Selection

Among the socioeconomic explanatory variables, one of them, `retail`, is not significant (p value $=0.10458$). It is common practice to perform model selection in search of a more parsimonious model that has possibly fewer explanatory variables. We use the R function `step()` to perform a backward elimination based on Akaike's information criterion (AIC) and save the result to an `lm` object `m2`. That is, we start with the full model `m1`, which has all the socioeconomic explanatory variables, and we drop the explanatory variable that results in the largest decrease in AIC iteratively until there is no further decrease in AIC. Recall that a smaller AIC indicates a better model fit balanced with model parsimony. It should be noted that the stepwise exercise could result in biased p values and thus should be avoided when possible.

```
#backward elimination based on AIC
m2 = step(m1)

## Start:  AIC=8219.82
## povty ~ ag + manu + retail + foreign + feemp + hsch + black + hisp
##
##           Df Sum of Sq   RSS    AIC
## <none>                  44402 8219.8
## - retail   1        38 44440 8220.5
## - manu     1       127 44529 8226.6
## - hisp     1       831 45232 8274.7
## - foreign  1       944 45346 8282.4
## - ag       1       975 45376 8284.5
## - hsch     1      2358 46759 8376.6
## - black    1      3208 47610 8432.0
## - feemp    1     32236 76638 9893.4
```

When the explanatory variable `retail` is dropped, the AIC value increases from 8219.8 to 8220.5. This indicates that the model without `retail` is not as good a model as the full model with all the explanatory variables `m1`. This holds for all the other explanatory variables as well, and thus none of the explanatory variables are dropped from the full model based on AIC. The final model `m2` is the same as the full model `m1`.

Alternatively, we could use the R function `step()` to perform backward elimination based on Schwartz's Bayesian information criterion (BIC) by setting the penalty coefficient to the log of the sample size ($k = \log(n)$) and save the result to an `lm` object `m3`. Like AIC, a smaller BIC indicates a good model fit balanced with model parsimony. We thus start with the full model `m1`, which has all the socioeconomic explanatory variables, and we drop the explanatory variable that results in the largest decrease in BIC iteratively until there is no further decrease in BIC:

```
#backward elimination based on BIC
n = nrow(povdf) #n is the sample size
m3 = step(m1, k=log(n))

## Start:  AIC=8274.08
## povty ~ ag + manu + retail + foreign + feemp + hsch + black + hisp
##
##           Df Sum of Sq   RSS    AIC
## - retail   1        38 44440 8268.7
## <none>                  44402 8274.1
## - manu     1       127 44529 8274.8
## - hisp     1       831 45232 8322.9
## - foreign  1       944 45346 8330.7
## - ag       1       975 45376 8332.7
## - hsch     1      2358 46759 8424.9
## - black    1      3208 47610 8480.2
## - feemp    1     32236 76638 9941.7
##
## Step:  AIC=8268.69
## povty ~ ag + manu + foreign + feemp + hsch + black + hisp
##
##           Df Sum of Sq   RSS    AIC
## - manu     1       102 44542 8267.7
## <none>                  44440 8268.7
## - hisp     1       845 45285 8318.5
## - foreign  1       933 45373 8324.5
## - ag       1      1576 46016 8367.7
## - hsch     1      2404 46844 8422.4
## - black    1      3332 47772 8482.6
## - feemp    1     32207 76647 9934.0
##
## Step:  AIC=8267.73
## povty ~ ag + foreign + feemp + hsch + black + hisp
##
##           Df Sum of Sq   RSS    AIC
## <none>                  44542 8267.7
## - hisp     1       893 45436 8320.7
## - foreign  1       987 45529 8327.0
## - ag       1      2211 46754 8408.4
## - black    1      3233 47775 8474.8
## - hsch     1      3572 48115 8496.5
## - feemp    1     32912 77454 9958.2
```

There are three steps in this model selection by BIC. In the first step, the `retail` explanatory variable is dropped from the reference model with all the explanatory variables because the BIC value of 8268.7 without `retail` is smaller than the BIC value of 8274.1 for the reference model with `retail`, whereas dropping any of the other explanatory variables would result in a BIC value larger than 8274.1 for the reference model. In the second step, the reference model without `retail` has a BIC value of 8268.7 and the `manu` explanatory variable is dropped because the BIC value of 8267.7 without `manu` is smaller than the BIC value of 8268.7 for the reference model with `manu`, whereas dropping any of the other explanatory variables would result in a BIC value larger than 8268.7 for the reference model. In the third

and last step, the reference model without `retail` and `manu` has a BIC value of 8267.7. Because leaving out any of the remaining explanatory variables would result in an increase in the BIC value, the model selection is finished. The final best model is `m3` without `retail` and `manu`, with the following summary of the model fit:

```
summary(m3)
```

```
##
## Call:
## lm(formula = povty ~ ag + foreign + feemp + hsch + black + hisp,
##     data = povdf)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -13.1475  -2.2578  -0.3776   1.6626  29.3469
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) 45.911167   0.858997  53.447  < 2e-16 ***
## ag           0.125505   0.010178  12.331  < 2e-16 ***
## foreign     -0.174881   0.021227  -8.239 2.55e-16 ***
## feemp       -0.528419   0.011108 -47.573  < 2e-16 ***
## hsch        -0.206914   0.013202 -15.673  < 2e-16 ***
## black        0.079324   0.005320  14.910  < 2e-16 ***
## hisp         0.063941   0.008159   7.837 6.31e-15 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 3.813 on 3063 degrees of freedom
## Multiple R-squared:  0.5847, Adjusted R-squared:  0.5839
## F-statistic: 718.8 on 6 and 3063 DF,  p-value: < 2.2e-16
```

This standard linear regression summary for `m3` provides estimates for the regression coefficients that are similar to those of `m1`. For example, in `m3` the estimated regression coefficient for `feemp` is -0.528 with standard error 0.011108, compared with an estimate of -0.526 with standard error 0.011166 in `m1`. There is a very slight decrease of the multiple and the adjusted R-squared values, but the amount of variation in the response variable explained by this final best model remains about 58%. We proceed with this set of six explanatory variables (`ag`, `foreign`, `feemp`, `hsch`, `black`, and `hisp`) in the remainder of this manuscript.

4.3 Model Diagnostics

Now that we have fitted standard linear regression models and selected a final model `m3`, we perform model diagnostics for the purpose of evaluating the model assumptions. There are four model assumptions to evaluate: linearity, independence, equal variance, and normality. *For linearity and equal variance, it is common to use the plot of residuals versus fitted responses.* This is the first option in the R function `plot()` applied to `m3` `(plot(m3, which=1))`. For normality, it is common to use the normal quantile–quantile (QQ) plot of the standardized residuals,

which is the second option in the R function `plot()` applied to `m3` (`plot(m3, which=2)`). These two plots can also be created using `ggplot2`. For the residuals versus fitted response, we extract the fitted and residual values from `m3` and use `geom_smooth()` layered over `geom_point()` to create the plot. For the QQ plot, we extract the standardized residuals from `m3` and use `stat_qq_line()` layered over `geom_qq()` to create the plot.

```
#QQ Plot

ggplot(m3,aes(sample=rstandard(m3))) +
    geom_qq(size=0.5) +
    stat_qq_line() +
    theme_bw() +
    xlab("Theoretical Quantiles") + ylab("Standardized Residuals")

#Residuals vs. Fitted Plot

ggplot(m3,aes(x=.fitted, y=.resid)) +
    geom_point(size=0.5) +
    geom_smooth() +
    theme_bw() +
    xlab("Fitted Values") + ylab("Residuals")
```

The normal Q–Q plot (Fig. 2a) shows a departure from the straight line at the upper end, indicating right skewness in the residuals (i.e., more large positive residuals than a normal distribution would typically have). The right skewness is also reflected in the plot of the residuals versus the fitted responses (Fig. 2b), while the remaining residuals appear to be scattered fairly randomly.

When model diagnostics like those above indicate any departure from the standard linear regression assumptions, remedial measures are not always needed due to the robustness of the regression. When remedial measures are needed, a commonly

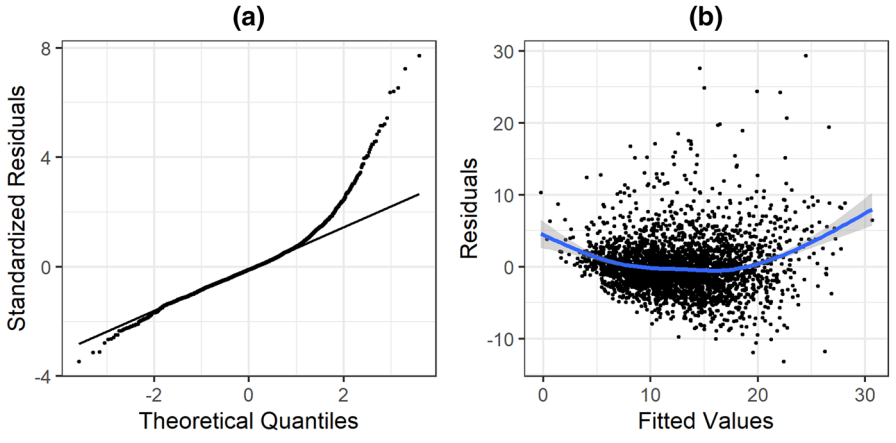


Fig. 2 Quantile–quantile (Q–Q) plot of the standard linear regression model residuals

used approach is transformation of the response variable and/or the explanatory variables. For illustration, we take the logit transformation of the response variable $\left(\log\left(\frac{p}{1-p}\right)\right)$ and fit a linear regression model. We use the `logit()` function from the `car` package (Fox and Weisberg 2011) as it adjusts proportions that are perfectly 0 or 1. Then we perform model diagnostics as before by a residual versus fitted response plot and a normal Q–Q plot (Fig. 3):

```
m3.logit = lm(car::logit(povty, percents = TRUE) ~ ag + foreign + feemp + hsc
h + black + hisp, data=povdf)

ggplot(m3.logit, aes(sample = rstandard(m3.logit))) +
    geom_qq(size = 0.5) +
    stat_qq_line() +
    theme_bw() +
    xlab("Theoretical Quantiles") + ylab("Standardized Residuals")

ggplot(m3.logit,aes(x=.fitted, y=.resid)) +
    geom_point(size=0.5) +
    geom_smooth() +
    theme_bw() +
    xlab("Fitted Values") + ylab("Residuals")
```

The normal Q–Q plot (Fig. 3a) shows less departure from the straight line at the upper end but more departure at the lower end, indicating a possible overcorrection of the right skewness in the untransformed data. In the plot of the residuals versus the fitted responses (Fig. 3b), several large negative residuals are marked and there is some indication of smaller variance for larger fitted values (i.e., unequal variance). This example demonstrates some of the challenges in model selection and model diagnostics. Taking a remedial measure to correct the departure from one

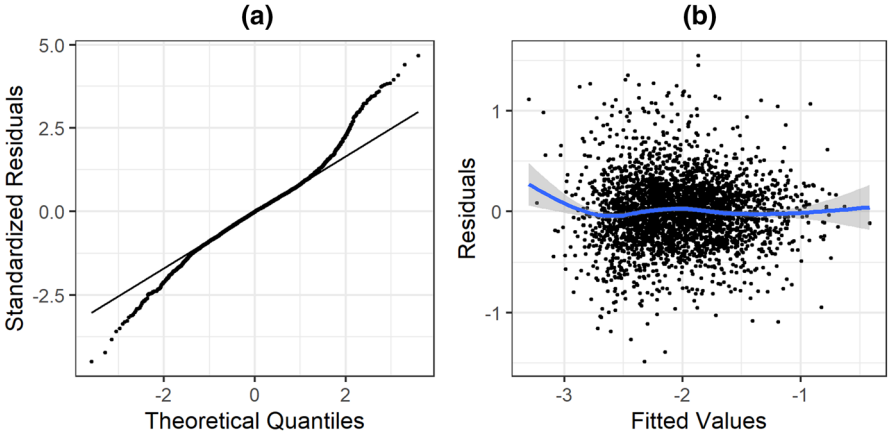


Fig. 3 Quantile–quantile (Q–Q) plot of the standard linear regression model residuals with the response variable transformed

assumption could lead to the departure from another assumption. For the remainder of this manuscript, we model the original poverty rate without any transformation.

Thus far we have evaluated the assumptions of linearity, equal variance, and normality. The independence assumption, however, is not an option in the R function `plot()`; we evaluate the independence assumption in Sect. 7.1.

5 Neighborhood Structure and Spatial Weight Matrix

In Sects. 3 and 4, we performed exploratory data analysis and standard linear regression analysis of the poverty rate data without considering spatial information in the data. In this section, we create a neighborhood structure and a spatial weight matrix in preparation for spatial analysis. In the following two sections, we perform exploratory spatial data analysis (Sect. 6) and carry out spatial regression analysis (Sect. 7).

Recall that the data frame `povdf` comprises of the response variable and the explanatory variables as well as the FIP code that identifies the counties in the 48 states of the contiguous United States.

```
head(povdf)
```

```
## # A tibble: 6 x 12
##   FIPS_N    ag black feemp foreign  hisp  hsch  manu povty retail STATEFP
##   <chr>  <dbl> <dbl> <dbl>   <dbl> <dbl> <dbl> <dbl> <dbl>  <dbl>   <dbl>
## 1 01001   2.13 17.1   53.2   1.17   1.40  33.8  16.5  9.06  12.9        1
## 2 01003   1.57 10.3   50.5   2.11   1.76  29.6  12.5  9.13  14.2        1
## 3 01005   3.97 46.3   43.5   1.5    1.65  32.4  31.4 21.9   10.5        1
## 4 01007   3.30 22.2   44.3   0.430  1.01  35.7  23.8 17.7    9.93       1
## 5 01009   2.18  1.19  48.1   3.10   5.33  36.0  19.5  9.89  11.6        1
## 6 01011  10.5  73.1   36.5   3.06   2.75  35.2  23.2 28.6    7.53       1
## # ... with 1 more variable: COUNTYFP <chr>
```

We now create neighbors and their corresponding spatial weights. First, we import U.S. counties shape file into R using the `counties()` function from the `tigris` package (Walker 2018). In order to only import counties that correspond to the counties in the poverty data set `povdf` (the contiguous United States), we specify (`state = povdf$STATEFP`). We set the resolution of the shape file to be 500 k with (`cb = TRUE`) and the census year for the shape files to be 2000. The county shape files are automatically imported into a `SpatialPolygonsDataFrame`. For faster performance and easier manipulation, we convert `uscounties` to a "simple features" data frame using the `st_as_sf()` function from the `sf` package (Pebesma 2018).

```
uscounties <- counties(state = povdf$STATEFP,
                       cb=TRUE,year=2000)
uscounties_sf <- st_as_sf(uscounties)
```

Now that we have the counties prepared, we need to connect the counties to the `povdf` data frame. To merge the two data frames, we use the `merge()` function from the sp package (Pebesma et al. 2018). Prior to merging, we first create the data vector `FIPS_N` by combining the state and county FIPs codes and set it as a variable in the `uscounties_sf` data frame. This gives us a unique identifier for each county which we can merge with `FIPS_N` from the `povdf` data frame.

```
uscounties_sf$FIPS_N <- paste0(uscounties_sf$STATEFP, uscounties_sf$COUNTYFP)
#merge sf to povdf
povdf_uscounties_sf <- sp::merge(uscounties_sf, povdf, by="FIPS_N")
```

The object `povdf_uscounties_sf` is both a simple features (`sf`) object and a `data.frame`, being associated multi-polygon geometries. In order to create the neighborhood structure of counties in the continental United States, we use the `poly2nb()` function from the spdep package (Bivand et al. 2013). We first coerce `povdf_uscounties_sf` to a `SpatialPolygonsDataFrame` using the `as_Spatial()` function (from the `sf` package) and specify `IDs = povdf_uscounties_sf$FIPS_N`. After setting the row names of `povdf_uscounties_spdf` to be the row names of the `povdf` data frame, we apply `poly2nb()`.

```
povdf_uscounties_spdf <- as_Spatial(povdf_uscounties_sf,
                    IDs = povdf_uscounties_sf$FIPS_N)
row.names(povdf_uscounties_spdf) <- row.names(povdf)
pov_nb <- poly2nb(povdf_uscounties_spdf,
                  row.names = povdf_uscounties_spdf$FIPS_N)
```

The output `pov_nb` is a neighborhood object which explicitly lists who are neighbors with whom and, in our case, which county is a neighbor with which county. We then create spatial weights by the R function `nb2listw()` from the spdep package using the default option of row standardization (`style="W"`) and binary weights (`style="B"`):

```
#W - default row-standardized weights
listw_povW = nb2listw(pov_nb, style="W",zero.policy = TRUE)
#B - binary weights
listw_povB = nb2listw(pov_nb, style="B", zero.policy = TRUE)
```

We specify `zero.policy=TRUE` above because some counties do not have any neighbors, such as Nantucket County in Massachusetts, and the neighborhood object `pov_nb` has entries that are null. The zero policy allows spatial weights to be created for counties with one or more neighbors despite the null entries. The output `listw_povB` is a spatial weight object, which can be visualized by a map as shown in Fig. 4. The `plot()` function applied to a list of

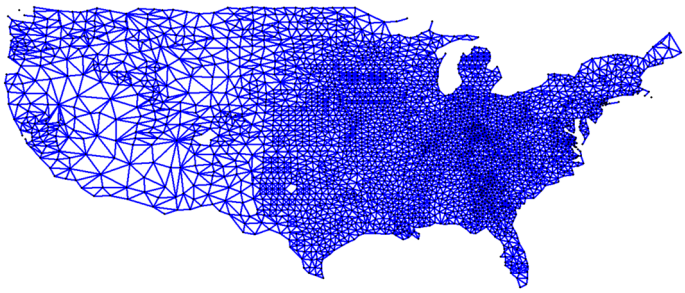


Fig. 4 Neighborhood structure

spatial weights takes the list of spatial weights and coordinates of the centroids as arguments. We extract geometries of the county polygons using `st_geometry()`, the centroids using `st_centroid()`, and finally the coordinates using `st_coordinates()`.

```
county_geoms <- st_geometry(povdf_uscounties_sf)
cntrd <- st_centroid(county_geoms)

coords <- st_coordinates(cntrd)

plot(listw_povB, coords, col="blue", cex=0.1)
```

The centroids of all pairs of neighboring counties are connected by a line and the map can be thought of as a network or a graph such that the county centroids are the vertexes and the connecting lines between neighboring counties are the edges.

6 Exploratory Spatial Data Analysis

We performed exploratory data analysis by summary statistics and graphical methods in Sect. 4. However, standard exploratory data analysis does not take into account the spatial nature of the data. In this section we consider exploratory spatial data analysis by summary statistics and graphical methods that utilize the spatial information in the data. The goal of exploratory spatial data analysis is for the reader to gain insights into the spatial nature of their data as a way to inform them how best to proceed with their choice of model, determination of outliers, and scope of the inferences that can be made from the data via confirmatory analysis. A natural graphical method to use for exploratory spatial data analysis is a heat map, where the levels of a variable are color coded on a map. We can draw the heat map of the poverty rates as well as heat maps of the socioeconomic explanatory variables over the 3070 counties in the contiguous United States.

We use `ggplot2` to plot the response variable, `povty`, by passing the sf data frame `povdf_uscounties_sf` to the `ggplot()` function. We add the

geometry `geom_sf()` and specify the aesthetic `aes(fill=povty)`. Additional features specify the black and white plot theme (`theme_bw()`) and the grey color scale (`(scale_fill_continuous(low="white", high="black"))`).

```
ggplot(povdf_uscounties_sf) +
    geom_sf(aes(fill=povty)) +
    theme_bw() +
    scale_fill_continuous(low="white", high="black")
```

For illustration, we use `ggplot()` to plot one of the explanatory variables, female employment rate `feemp`.

```
ggplot(povdf_uscounties_sf) +
    geom_sf(aes(fill=feemp)) +
    theme_bw() +
    scale_fill_continuous(low="white", high="black")
```

The spatial pattern of `povty` (Fig. 5a) indicates relatively low poverty rates in the northern and western states, whereas the poverty rates are relatively high in the southeastern states. The spatial pattern of `feemp` (Fig. 5b) indicates relatively high female employment rates in eastern states. The relationship between poverty and female employment rates shown in Fig. 5 is not as obvious as in Fig. 1.

It should be noted that the `tmap` R package and extensions can also be used to create high quality thematic maps. We provide a supplementary tutorial online (https://mkamenet3.github.io/SpatialRegPovertyR/) and encourage the reader to further explore spatial visualizations using `tmap` (Lovelace et al. 2019).

6.1 *Moran's* I *and Geary's* c

For exploratory spatial data analysis, summary statistics such as Moran's I and Geary's c can be used to quantify spatial dependence. With the neighborhoods created in Sect. 5, we use the R function `moran.test()` to estimate the Moran's I statistic and perform a Moran's I test for the null hypothesis that there is no spatial dependence versus a two-sided alternative hypothesis that there is spatial dependence:

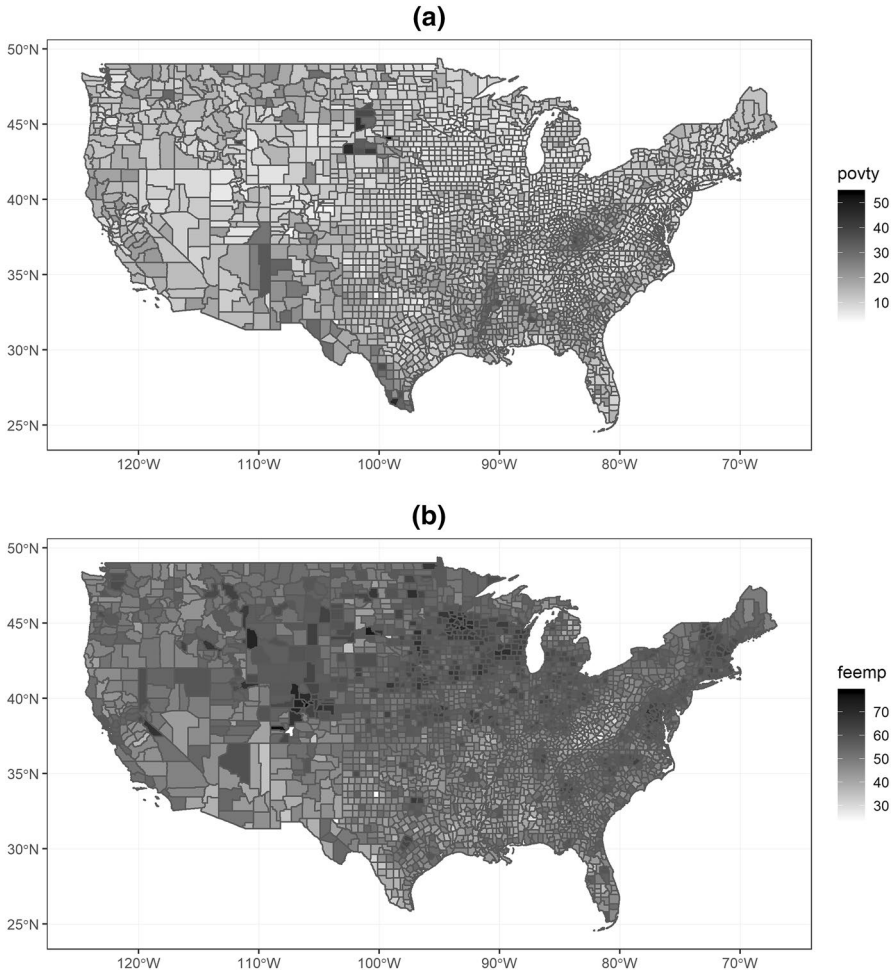


Fig. 5 Poverty rate (**a**) and female employment rate (**b**) in 2000

```
#Moran's I test based on randomization
moran.test(povdf$povty, listw_povB, zero.policy = TRUE,
alternative = "two.sided")

##
##  Moran I test under randomisation
##
## data:  povdf$povty
## weights: listw_povB  n reduced by no-neighbour observations
##
## Moran I statistic standard deviate = 53.934, p-value < 2.2e-16
## alternative hypothesis: two.sided
## sample estimates:
## Moran I statistic       Expectation          Variance
##      0.5668859074     -0.0003260515      0.0001106021
```

Based on the output above, the Moran's I statistic is 0.5669, indicating positive spatial dependence among neighboring counties. If there is no spatial dependence, then the expected value and the variance of the Moran's I statistic are -0.000326 and 0.0001106, respectively. This results in a standard deviation (or Z-value) of 53.934 with a p value less than 2.2E−16 (i.e., 2.2×10^{-16}), which is virtually zero. There is very strong evidence for spatial dependence in the poverty rates across counties in the continental U.S.

The Moran's I test by `moran.test()` is based on a normal approximation and the variance estimation is based on randomization. Alternatively, a Monte Carlo test can be performed by the R function `moran.mc()`, where we pre-specify the number of Monte Carlo simulations (here, `nsim`=999) and compare the Moran's I statistic of 0.5669 against the null distribution of the Moran's I statistic under the assumption of no spatial dependence. We set the randomization seed to 1 so that when the code is re-run we get the same test result:

```
#Moran's I test based on Monte Carlo
set.seed(1)
moran.mc(povdf$povty, listw_povB, zero.policy = TRUE, nsim=999)

##
##  Monte-Carlo simulation of Moran I
##
## data:  povdf$povty
## weights: listw_povB
## number of simulations + 1: 1000
##
## statistic = 0.56641, observed rank = 1000, p-value = 0.001
## alternative hypothesis: greater
```

This Monto Carlo test shows that the rank of the observed Moran's I test statistic of 0.5664 is larger than any of the test statistics from the `nsim`=999 simulations and thus has the highest rank. The p value is 1 out of 1000, which is 0.001 for a null hypothesis that there is no spatial dependence versus a one-sided alternative that

there is positive spatial dependence. That is, there is very strong evidence for positive spatial dependence in the poverty rates among counties. For a two-sided alternative that there is spatial dependence (positive or negative), we double 0.001 and obtain a p value of 0.002. There is still very strong evidence for spatial dependence in the poverty rates among counties.

Besides Moran's I, we may use the R function `geary.test()` to estimate the Geary's c statistic and perform a Geary's c test for the null hypothesis that there is no spatial dependence versus a two-sided alternative hypothesis that there is spatial dependence:

```
geary.test(povdf$povty, listw_povB, zero.policy=TRUE,
alternative = "two.sided")
```

```
##
##  Geary C test under randomisation
##
## data:  povdf$povty
## weights: listw_povB
##
## Geary C statistic standard deviate = 40.714, p-value < 2.2e-16
## alternative hypothesis: two.sided
## sample estimates:
## Geary C statistic       Expectation          Variance
##      0.4287877788      1.0000000000      0.0001968414
```

Based on this output, the Geary's c statistic is 0.4288, indicating a positive spatial dependence among neighboring counties. If there were no spatial dependence, the expected value and the variance of the Geary's c statistic would be 1 and 0.0001968, respectively. This results in a standard deviation (or Z-value) of 40.714 with a p value that is also virtually zero. There is again very strong evidence for spatial dependence in the poverty rates across counties.

The Geary's c test by `geary.test()` is based on a normal approximation and the variance estimation is based on randomization. Alternatively, a Monte Carlo test can be performed by the R function `geary.mc()`, where we pre-specify the number of Monte Carlo simulations (by `nsim`=999) and compare the Geary's c statistic of 0.4288 against its null distribution under the assumption of no spatial dependence:

```
set.seed(1)
geary.mc(povdf$povty, listw_povB, zero.policy = TRUE, nsim=999)

##
##  Monte-Carlo simulation of Geary C
##
## data:  povdf$povty
## weights: listw_povB
## number of simulations + 1: 1000
##
## statistic = 0.42879, observed rank = 1, p-value = 0.001
## alternative hypothesis: greater
```

This Monto Carlo test shows that the rank of the observed Geary's c test statistic of 0.42879 is larger than any of the test statistics from the `nsim`=999 simulations and thus ranks the highest. The p value is thus 0.002 (0.001) for a two-sided (one-sided) alternative and there is very strong evidence for spatial dependence (positive spatial dependence) in the poverty rates across counties.

6.2 *Local Moran's* I

To apply local Moran's I to the data, we use the R function `localmoran()` from the spdep package. The output is extensive, so here we apply the R function `head()` to take a look at the first six counties as an example. We also use the R function `round()` to round the output values to two digits after the decimal point:

```
head(round(localmoran(povdf$povty, listw_povB, zero.policy = TRUE),2))

##          Ii E.Ii Var.Ii  Z.Ii Pr(z > 0)
## 01001 -2.78    0   4.99 -1.24      0.89
## 01003 -2.64    0   5.98 -1.08      0.86
## 01005 15.38    0   7.97  5.45      0.00
## 01007  3.77    0   5.98  1.54      0.06
## 01009  0.08    0   5.98  0.03      0.49
## 01011 18.14    0   4.99  8.12      0.00
```

For example, for the county identified by *FIPS_N* 01001 in the output above (Autauga County, Alabama), the local Moran's I (`Ii`) is -4.99, with virtually zero expectation (`E.Ii`) and a variance of 4.99 (`Var.Ii`). Thus, the standard deviate (`Z.Ii`) is -1.24 and the p value (`1-Pr(z>0)`) is $1-0.89=0.11$ for a one-sided alternative (the p value is 0.22 for a two-sided alternative). There is no evidence of local spatial dependence for this county.

7 Spatial Linear Regression

7.1 Diagnostics for Spatial Dependence

In Sect. 4, we fitted and selected standard linear regression models to quantify the relationship between poverty rates and socioeconomic variables. We also carried out model diagnostics using residuals plots to evaluate the model assumptions of linearity, equal variance, and normality (but not yet independence). We are now ready to evaluate the independence assumption.

Here we use row-standardized spatial weights and recall the spatial weights object `listw_povW`. We apply the R function `lm.morantest()` to the residuals of the fitted standard linear regression models. The null hypothesis is that there is no spatial dependence in the error term of the standard linear regression model.

```
lm.morantest(m3, listw_povW, zero.policy = TRUE, alternative = "two.sided")
```

```
##
##  Global Moran I for regression residuals
##
## data:
## model: lm(formula = povty ~ ag + foreign + feemp + hsch + black +
## hisp, data = povdf)
## weights: listw_povW
##
## Moran I statistic standard deviate = 25.358, p-value < 2.2e-16
## alternative hypothesis: two.sided
## sample estimates:
## Observed Moran I      Expectation         Variance
##     0.2704016625    -0.0015987855     0.0001150541
```

Based on the output above, the Moran's I statistic is 0.2701, indicating a positive spatial dependence among neighboring counties. The standard deviation (or Z-value) is 25.358 with a p value less than 2.2E−16, which is virtually zero. There is very strong evidence for spatial dependence among the errors of the standard linear regression model. This calls for more complex spatial linear regression analysis. The next subsection presents spatial lag models and spatial error models including the simultaneous autoregressive (SAR) and conditional autoregressive (CAR) models.

7.2 Spatial Lag Models

We consider fitting a spatial lag model to the poverty rates data. The R function `lagsarlm()` is applied to the data frame `povdf` and the output is `m3_lag`.

```
m3_lag = lagsarlm(povty ~ ag  + foreign + feemp + hsch + black + hisp, data=p
ovdf, listw = listw_povW,type="lag",zero.policy=TRUE)
```

Coefficients estimated from spatial lag models cannot be interpreted directly because of spillovers between the terms in the data generation process. We refer the reader to Golgher and Voss (2016) regarding spatial confounding and spillover effects. We calculate impacts using the `impacts()` function from the spdep package.

```
impacts(m3_lag, listw=listw_povW) # this takes several minutes
```

```
## Impact measures (lag, exact):
##              Direct    Indirect       Total
## ag       0.09941542  0.05776276  0.15717818
## foreign -0.07402554 -0.04301063 -0.11703616
## feemp   -0.38979879 -0.22648252 -0.61628131
## hsch    -0.14686715 -0.08533336 -0.23220051
## black    0.06168646  0.03584133  0.09752779
## hisp     0.03206871  0.01863269  0.05070140
```

The output gives us the direct (or local) effect, indirect (or spillover) effect, and total effect (or sum of the direct and indirect effects). The total effect can be interpreted similarly to our interpretation of regression coefficients in the standard linear model. The total effect of `feemp` is -0.616 (rounded from -0.61628131). A 1% increase in female employment is associated with a 0.616% decrease in poverty.

```
summary(m3_lag, correlation=FALSE)
```

```
##
## Call:lagsarlm(formula = povty ~ ag + foreign + feemp + hsch + black +
##     hisp, data = povdf, listw = listw_povW, type = "lag", zero.policy = TR
UE)
##
## Residuals:
##       Min        1Q    Median        3Q       Max
## -14.28213  -2.03550  -0.34323   1.40795  31.92721
##
## Type: lag
## Regions with no neighbours included:
##  25019 36085 53055
## Coefficients: (asymptotic standard errors)
##               Estimate Std. Error  z value  Pr(>|z|)
## (Intercept) 31.1828042  1.0847285  28.7471 < 2.2e-16
## ag           0.0963964  0.0094252  10.2275 < 2.2e-16
## foreign     -0.0717775  0.0198912  -3.6085  0.000308
## feemp       -0.3779615  0.0126199 -29.9495 < 2.2e-16
## hsch        -0.1424071  0.0128045 -11.1216 < 2.2e-16
## black        0.0598132  0.0051485  11.6175 < 2.2e-16
## hisp         0.0310949  0.0077062   4.0350  5.46e-05
##
## Rho: 0.38694, LR test value: 409.57, p-value: < 2.22e-16
## Asymptotic standard error: 0.018873
##     z-value: 20.502, p-value: < 2.22e-16
## Wald statistic: 420.35, p-value: < 2.22e-16
##
## Log likelihood: -8257.118 for lag model
## ML residual variance (sigma squared): 12.334, (sigma: 3.512)
## Number of observations: 3070
## Number of parameters estimated: 9
## AIC: 16532, (AIC for lm: 16940)
## LM test for residual autocorrelation
## test value: 47.6, p-value: 5.2258e-12
```

There are five parts to this spatial linear regression output. The initial function call is in the first part. The second part reports the summary statistics of residuals (minimum, first quartile, median, third quartile, and maximum). Here the residual is defined as the difference between an observed response and its fitted value by the spatial linear regression. The third part reports, for each explanatory variable, a fitted regression coefficient (under `Estimate`), its standard errors (under `Std. Error`), the ratio of the two values as a Z test statistic (under `z value`), and a p value for testing whether the true regression coefficient is zero or not (under `Pr(>|z|)`). For the spatial lag model, we use the impacts to assess coefficient estimates. For example, the estimated regression coefficient for `feemp` is -0.378 with standard error 0.01262. The Z test statistic is -29.9495 and the p value is less than 2.2E–16. This tells us that a 1% increase in the female employment rate is associated with a 0.378% decrease in the poverty rate when all other explanatory variables are held constant. Relative to the standard error of about 1.262%, the Z statistic -29.9495 is very large and the p value is extremely

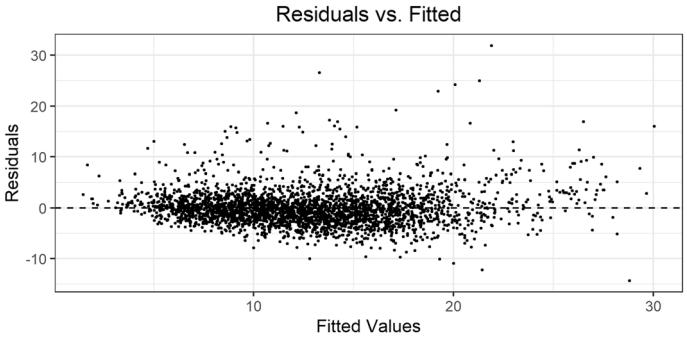


Fig. 6 Fitted values versus residuals of the spatial lag model

small. There is very strong evidence that the true regression coefficient for female employment rate is not zero.

The fourth part of the output provides the estimate of the spatial correlation coefficient `rho`, which is 0.387. Two hypothesis tests—a likelihood ratio (LR) test and a Wald test—are applied for testing whether the true spatial correlation coefficient `rho` is zero or not. The LR test value is 409.57 with a p value of less than 2.2E−16. The Wald test is presented as a Z test with a z-value of 20.502 and a p value of less than 2.2E−16. Equivalently, the Wald test statistic is 420.35, which is the square of the z-value, and the p value of less than 2.2E−16 is the same as the Z test. Both the LR test and the Wald test indicate there is very strong evidence that the spatial correlation coefficient `rho` is not zero.

The fifth and last part of the output provides several useful statistics and tests. The log-likelihood value for the fitted spatial lag model is −8257.118 and the AIC value is 16532, which is smaller than the AIC value of 16,940 for the standard linear regression model. The maximum likelihood estimate of the error variance `sigma squared` is 12.334, with the error standard deviation `sigma` estimated to be its square root, 3.512. Finally, the Lagrange multiplier (LM) test for the spatial dependence in the error term has a test value of 47.6 and a small p value, indicating that there is additional spatial dependence unaccounted for by the spatial lag model.

We have already tested for spatial dependence in the error term and learned that there is strong evidence for spatial dependence. In addition, we may plot the residuals against the fitted responses (Fig. 6) as follows:

```
m3_lag.df <- cbind.data.frame(resids = residuals(m3_lag), fit = fitted(m3_lag
))
ggplot(m3_lag.df, aes(x = fit, y = resids)) +
    geom_point(size = 0.5) +
    geom_hline(yintercept = 0, linetype="dashed") +
    theme_bw() +
    xlab("Fitted Values") + ylab("Residuals")
```

From the residual plot (Fig. 6), we see that the residuals are distributed fairly randomly around the zero horizontal line and the variation tends to be higher for larger

fitted values. This suggests no obvious departure from the linearity assumption, but there is indication of unequal variances among the errors.

We apply the Breusch–Pagan (BP) test for the null hypothesis that the error variance is constant versus the alternative that the error variance is not constant by the R function `bptest.sarlm()`:

```
bptest.sarlm(m3_lag)

##
##  studentized Breusch-Pagan test
##
## data:
## BP = 218.63, df = 6, p-value < 2.2e-16
```

The observed BP test statistic is 218.63 and the p value is extremely small. There is very strong evidence for unequal variance in the error term of the spatial lag model.

7.3 Spatial Durbin Error Models

As an alternative to the spatial lag model, we consider fitting a spatial Durbin error model to the poverty rates data. The R function `errorsarlm()` is applied to the data frame `povdf` and the option `Durbin` is to `TRUE` to specify the Durbin error model. The output is `m3_err`. We then apply the R function `summary()` to the `m3_err` object and obtain the results of the spatial error model fit:

```
m3_err = errorsarlm(povty ~ ag + foreign + feemp + hsch + black + hisp, data=
povdf, listw = listw_povW, zero.policy = TRUE, Durbin=TRUE)
summary(m3_err)
```

```
##
## Call:errorsarlm(formula = povty ~ ag + foreign + feemp + hsch + black +
##     hisp, data = povdf, listw = listw_povW, Durbin = TRUE, zero.policy = T
RUE)
##
## Residuals:
##       Min        1Q    Median        3Q       Max
## -12.82154  -1.89410  -0.35771   1.35667  29.29013
##
## Type: error
## Regions with no neighbours included:
##  25019 36085 53055
## Coefficients: (asymptotic standard errors)
##               Estimate Std. Error  z value  Pr(>|z|)
## (Intercept) 44.3359709  1.4348067  30.9003 < 2.2e-16
## ag           0.0807951  0.0125397   6.4431 1.170e-10
## foreign     -0.0485734  0.0286496  -1.6954  0.089995
## feemp       -0.4338157  0.0142404 -30.4636 < 2.2e-16
## hsch        -0.2383383  0.0147707 -16.1358 < 2.2e-16
## black        0.1470144  0.0083749  17.5541 < 2.2e-16
## hisp         0.1070956  0.0146375   7.3165 2.545e-13
## lag.ag       0.0848673  0.0204621   4.1475 3.361e-05
## lag.foreign -0.1326324  0.0448534  -2.9570  0.003106
## lag.feemp   -0.1384157  0.0195142  -7.0931 1.312e-12
## lag.hsch     0.1359980  0.0238282   5.7074 1.147e-08
## lag.black   -0.0747138  0.0113379  -6.5897 4.406e-11
## lag.hisp    -0.0430697  0.0192686  -2.2352  0.025402
##
## Lambda: 0.50771, LR test value: 464.16, p-value: < 2.22e-16
## Asymptotic standard error: 0.021888
##     z-value: 23.195, p-value: < 2.22e-16
## Wald statistic: 538.03, p-value: < 2.22e-16
##
## Log likelihood: -8055.921 for error model
## ML residual variance (sigma squared): 10.564, (sigma: 3.2503)
## Number of observations: 3070
## Number of parameters estimated: 15
## AIC: 16142, (AIC for lm: 16604)
```

There are five parts to this spatial linear regression output. The initial function call is in the first part. The second part reports the summary statistics of residuals (minimum, first quartile, median, third quartile, and maximum. The third part reports each explanatory and fitted coefficients. However, these cannot be directly interpreted due to the spatial lag in the model. As with the spatial lag model, we use the impacts to assess coefficient estimates in this spatial Durbin error model because there is a spatial lag in the model.

```
impacts(m3_err)
```

```
## Impact measures (SDEM, estimable):
##              Direct    Indirect       Total
## ag       0.08079511  0.08486734  0.16566245
## foreign -0.04857337 -0.13263237 -0.18120574
## feemp   -0.43381574 -0.13841569 -0.57223143
## hsch    -0.23833829  0.13599803 -0.10234026
## black    0.14701435 -0.07471384  0.07230051
## hisp     0.10709563 -0.04306973  0.06402590
```

For example, the total effect of `feemp` is -0.572 (rounded from -0.57223143). A 1% increase in female employment is associated with a 0.572% decrease in poverty.

The fourth part of the `summary` output provides the estimate of the spatial correlation coefficient `lambda`, which is 0.508. The LR test value is 464.16 with a p value of virtually zero. The Wald test has a p value that is also virtually zero. Both the LR test and the Wald test indicate there is very strong evidence that the spatial correlation coefficient `lambda` is not zero.

The fifth and last part of the `summary` output provides several useful statistics and tests. The log-likelihood value for the fitted spatial lag model is -8055.921 and the AIC value is 16,142, which is smaller than the AIC value of 16,940 for the standard linear regression model, smaller than the AIC value of 16,604 for the standard linear regression model with demographic lag terms included, and the AIC value of 16,532 for the spatial lag model. The maximum likelihood estimate of the error variance `sigma` squared is 10.564, with the error standard deviation `sigma` estimated to be its square root, 3.2503.

Unlike the spatial lag model fit, there is no test for spatial dependence in the residuals. Thus, we apply the R function `moran.mc()` to test for spatial dependence in the error term of the spatial error model:

```
moran.mc(residuals(m3_err), listw_povW, zero.policy=TRUE, nsim=999)
```

```
##
##  Monte-Carlo simulation of Moran I
##
## data:  residuals(m3_err)
## weights: listw_povW
## number of simulations + 1: 1000
##
## statistic = -0.021028, observed rank = 23, p-value = 0.977
## alternative hypothesis: greater
```

The observed Moran's I test statistic is -0.021028 with an observed rank of 23. Thus, the p value is $(1 - 0.977) \times 2$, which is 0.046. There is weak evidence of additional spatial dependence unaccounted for by the spatial error model.

7.4 Spatial SAR Models

An alternative to `lagsarlm()` is the R function `spautolm()` applied to the data frame `povty`; the output is `m3_sar`. We apply the R function `summary()` to `m3_sar` and obtain the results of the spatial linear regression that assumes a simultaneous autoregressive (SAR) model for the error term:

```
m3_sar = spautolm(povty ~ ag + foreign + feemp + hsch + black + hisp, data=po
vdf, listw = listw_povW, zero.policy = TRUE, family="SAR")
summary(m3_sar)
```

```
##
## Call: spautolm(formula = povty ~ ag + foreign + feemp + hsch + black +
##     hisp, data = povdf, listw = listw_povW, family = "SAR", zero.policy =
TRUE)
##
## Residuals:
##       Min        1Q    Median        3Q       Max
## -14.66286  -1.86254  -0.36497   1.41605  29.96495
##
## Regions with no neighbours included:
##  25019 36085 53055
##
## Coefficients:
##               Estimate Std. Error  z value  Pr(>|z|)
## (Intercept) 41.5028349  1.0719155  38.7184 < 2.2e-16
## ag           0.0771771  0.0126031   6.1236 9.146e-10
## foreign     -0.0719826  0.0270897  -2.6572  0.007879
## feemp       -0.4412998  0.0143189 -30.8193 < 2.2e-16
## hsch        -0.2296389  0.0148251 -15.4898 < 2.2e-16
## black        0.1267021  0.0076672  16.5253 < 2.2e-16
## hisp         0.0901874  0.0123613   7.2960 2.967e-13
##
## Lambda: 0.61632 LR test value: 596.03 p-value: < 2.22e-16
## Numerical Hessian standard error of lambda: 0.020278
##
## Log likelihood: -8163.891
## ML residual variance (sigma squared): 11, (sigma: 3.3166)
## Number of observations: 3070
## Number of parameters estimated: 9
## AIC: 16346
```

Percentages of agricultural workers, black, and Hispanic are positively associated with poverty rates, whereas percentages of foreign born, female employment, and high school graduates are negatively associated with poverty rates. For example, the estimated regression coefficient for feemp is -0.441 (rounded from -0.4412998) with standard error 0.0143189. The Z test statistic is -30.8193 and the p value is less than 2.2×10^{-16}. This suggests that a 1% increase in the female employment rate is associated with a 0.441% decrease in the poverty rate, with all other explanatory variables held constant. Relative to the standard error of about 1.431, the Z statistic -30.8193 is very large and the p value is extremely small. There is very strong evidence that the true regression coefficient for female employment rate is not zero. The output also provides the estimate of the spatial correlation coefficient lambda,

which is 0.61632. The LR test value is 596.03 with a p value of virtually zero. There is very strong evidence that the spatial correlation coefficient lambda is not zero.

7.5 Spatial CAR Models

In the R function `spautolm()`, we may specify a conditional autoregressive (CAR) model for the error term as follows:

```
m3_car = spautolm(povty ~ ag  + foreign + feemp + hsch + black + hisp, data=p
ovdf, listw = listw_povW, zero.policy = TRUE, family="CAR")
summary(m3_car)
```

```
##
## Call: spautolm(formula = povty ~ ag + foreign + feemp + hsch + black +
##     hisp, data = povdf, listw = listw_povW, family = "CAR", zero.policy =
TRUE)
##
## Residuals:
##       Min        1Q    Median        3Q       Max
## -15.22966  -1.86101  -0.28561   1.42590  29.18291
##
## Regions with no neighbours included:
##  25019 36085 53055
##
## Coefficients:
##               Estimate Std. Error  z value  Pr(>|z|)
## (Intercept) 38.8446396  1.1526919  33.6991 < 2.2e-16
## ag           0.0502052  0.0132623   3.7856 0.0001534
## foreign     -0.1126881  0.0288758  -3.9025  9.52e-05
## feemp       -0.4024270  0.0149676 -26.8866 < 2.2e-16
## hsch        -0.2421201  0.0153654 -15.7575 < 2.2e-16
## black        0.1624423  0.0085434  19.0137 < 2.2e-16
## hisp         0.1278227  0.0145159   8.8057 < 2.2e-16
##
## Lambda: 0.971 LR test value: 712.59 p-value: < 2.22e-16
## Numerical Hessian standard error of lambda: 0.016928
##
## Log likelihood: -8105.608
## ML residual variance (sigma squared): 9.8422, (sigma: 3.1372)
## Number of observations: 3070
## Number of parameters estimated: 9
## AIC: 16229
```

Similar to the spatial SAR model fit, the percentages of agricultural workers, Black, and Hispanic are positively associated with poverty rates, whereas the percentages of foreign-born residents, female employment, and high school graduates are negatively associated with poverty rates. For example, the estimated regression coefficient for `feemp` is -0.402 with standard error `0.0149676`. The Z test statistic is `-26.8866` and the p value is less than 2.2E−16. This tells us that a 1% increase in the female employment rate is associated with a 0.402% decrease in the poverty rate when all other explanatory variables are held constant. Relative to the standard error of about 1.450%, the absolute value of the Z statistic -26.8866 is

very large and the p value is extremely small. There is very strong evidence that the true regression coefficient for female employment rate is not zero.

The output also provides the estimate of the spatial correlation coefficient `lambda`, which is 0.971. The LR test value is 712.59 with a p value of virtually zero. There is very strong evidence that the spatial correlation coefficient `lambda` is not zero.

The log likelihood value for the fitted spatial CAR model is -8105.608 and the AIC value is 16,229, which is smaller than the AIC values of 16,940, 16,532, 16,346 for the standard linear regression, the spatial lag model, and the spatial SAR model, respectively. It is larger than the AIC for the spatial Durbin error model (AIC = 16,142). The maximum likelihood estimate of the error variance `sigma`^{2} is 9.8422, with the error standard deviation `sigma` estimated to be its square root 3.1372.

We apply the R function `moran.mc()` to test for spatial dependence in the error term of the spatial CAR model:

```
moran.mc(resid(m3_car), listw_povW, zero.policy = TRUE, nsim=999)

##
##  Monte-Carlo simulation of Moran I
##
## data:  resid(m3_car)
## weights: listw_povW
## number of simulations + 1: 1000
##
## statistic = -0.20909, observed rank = 1, p-value = 0.999
## alternative hypothesis: greater
```

The observed Moran's I test statistic is -0.20909 with an observed rank of 1. Thus, the p value is 0.001×2, which is 0.002. There is strong evidence of additional spatial dependence unaccounted for by the spatial CAR model. The spatial CAR model has a smaller AIC value than the spatial SAR model. The AIC value of the spatial SAR model is close to that for the spatial CAR model and the fitted regression coefficients and their standard errors are similar between the two models with qualitatively the same interpretation of the relationship between poverty rates and the social-economic explanatory variables. There is evidence of spatial dependence in the error terms of both the spatial SAR and CAR models.

8 Summary

In this manuscript we illustrate the general steps of spatial regression modeling of poverty in R. The spatial methods are limited to spatial dependence, one aspect of spatial regression modeling. At least two other spatial aspects could be considered in poverty research—spatial heterogeneity and spatial panel.

Spatial heterogeneity could refer to the fact that individual variables or the regression coefficients between a response variable and explanatory variables vary

systematically across space (Dutilleul 2011; LeSage 1999). Frequently it is found that the associations of the response variable with the explanatory variables vary across the studied area. There are at least three approaches to deal with spatial heterogeneity. The first approach, widely used by sociologists, is to use aspatial methods, such as using dummy variables indicating the category of regions, combining dummy variables with explanatory variables, partitioning the study area into several regions that exhibit different spatial patterns, and then separately fitting standard linear regression for each region (Baller and Richardson 2002). The disadvantage of this approach is that it makes it difficult to practically control spatial dependence when any of the partitioned regions are not contiguous or when they change over time. The second approach is using the geographically weighted regression (GWR) method (Fotheringham, Brunsdon, and Charlton 1998), which enables modeling of the spatially-varying coefficients. However, GWR does not consider the spatial lag and error dependence in the spatial regression context; this makes it difficult to consider the spatial lag and error dependence simultaneously. The third approach is to apply a spatial regime model to estimate coefficients separately for each regime (Anselin 1990; Patton and McErlean 2003). This approach allows diagnosing each variable's coefficient stability as well as the overall structural stability.

Spatial panel data are geographically referenced and have observations at each areal unit over multiple time points. Spatial panel data might exhibit both spatial dependence among observations of areal units at each time point and temporal dependence among observations of each areal unit over time. Such spatial panel data present researchers with various modeling possibilities. Spatio-temporal regression models refer to regression models that consider both spatial and temporal dependence exhibited in the data, i.e., a combination of the capacity of spatial regression modeling and time-series analysis. Generally, there are two approaches for spatio-temporal regression modeling.

The first approach for spatio-temporal regression modeling is to fit spatial regression models separately for each time point (or period) and then compare the results, especially model parameters (including regression coefficients, variance components, and spatial parameters), across the multiple time points (or periods). However, the temporal dimension is considered only by comparing the temporal difference of model parameters rather than through temporal dependence. Therefore, this approach does not consider both spatial dependence and temporal dependence simultaneously. One advantage, though, is that it allows us to conduct spatial panel data analysis without knowledge beyond the basic spatial regression models—spatial lag models and spatial error models—while at the same time providing insights into the spatial dependence in the data and the temporal variation of the model parameters.

The second approach for spatio-temporal regression modeling is to formally consider spatial and temporal dependence simultaneously in linear regression models. There are a number of spatio-temporal regression models, and each has different strengths and limitations. Readers are suggested to refer to these methods for a comprehensive review of spatio-temporal regression models (e.g., Anselin 1988; Anselin and Bera 1998; Baltagi and Li 2004; Cressie 1993; Elhorst 2001, 2010; Huang et al. 2010; Lee and Yu 2010; LeSage and Pace 2009).

Acknowledgements This research was supported in part by the National Science Foundation (Awards # CMMI-1541136, # OPP-1745369, # SES-1823633, and # DGE-1806874), the National Aeronautics and Space Administration (Award # NNX15AP81G), the Eunice Kennedy Shriver National Institute of Child Health and Human Development (Award # P2C HD041025), the National Institute on Alcohol Abuse and Alcoholism (Award # U24 AA027684-01), and the Social Science Research Institute, Population Research Institute, and the Institutes for Energy and the Environment of the Pennsylvania State University.

References

Anselin, L. (1988). *Spatial econometrics: Methods and models*. Dordrecht: Kluwer Academic Publishers.

Anselin, L. (1990). Spatial dependence and spatial structural instability in applied regression analysis. *Journal of Regional Science, 30,* 185–207.

Anselin, L., & Bera, A. (1998). Spatial dependence in linear regression models with an introduction to spatial econometrics. In A. Ullah & D. Giles (Eds.), *Handbook of applied economic statistics* (pp. 237–289). New York, NY: Marcel Dekker.

Baller, R. D., & Richardson, K. K. (2002). Social integration, imitation, and the geographic patterning of suicide. *American Sociological Review, 67,* 873–888.

Baltagi, B., & Li, D. (2004). Prediction in the panel data model with spatial autocorrelation. In L. Anselin, R. J. G. M. Florax, & S. Rey (Eds.), *Advances in spatial econometrics: Methodology, tools, and applications* (pp. 283–295). New York, NY: Springer.

Bennett, K. J., Probst, J. C., & Pumkam, C. (2011). Obesity among working age adults: The role of county-level persistent poverty in rural disparities. *Health and Place, 17,* 1174–1181.

Bivand, R., Pebesma, E., & Gomez-Rubio, V. (2013). *Applied spatial data analysis with R*. New York, NY: Wiley.

Cressie, N. (1993). *Statistics for spatial data*. New York, NY: Wiley.

Curtis, K. J., Lee, J., O'Connell, H. A., & Zhu, J. (2018). The spatial distribution of poverty and the long reach of the industrial makeup of places: New evidence on spatial and temporal regimes. *Rural Sociology*. https://doi.org/10.1111/ruso.12216.

Curtis, K. J., Voss, P. R., & Long, D. D. (2012). Spatial variation in poverty-generating processes: Child poverty in the United States. *Social Science Research, 41*(1), 146–159.

Duncan, C. M. (1999). *Worlds apart: Why poverty persists in rural America*. New Haven, CT: Yale University Press.

Dutilleul, P. R. L. (2011). *Spatio-temporal heterogeneity: Concepts and analyses*. New York, NY: Cambridge University Press.

Elhorst, J. P. (2001). Dynamic models in space and time. *Geographical Analysis, 33,* 119–140.

Elhorst, J. P. (2010). Applied spatial econometrics: Raising the bar. *Spatial Economic Analysis, 5*(1), 9–28.

Fotheringham, A. S., Brunsdon, M., & Charlton, M. (1998). Geographically weighted regression: A natural evolution of the expansion method for spatial data analysis. *Environment and Planning A: Economy and Space, 30,* 1905–1927.

Fox, J., Weisberg, S. (2011). *An R companion to applied regression*, 2nd edn. Thousand Oaks CA: Sage. http://socserv.socsci.mcmaster.ca/jfox/Books/Companion.

Goetz, S. J., & Swaminathan, H. (2006). Wal-Mart and county-wide poverty. *Social Science Quarterly, 87,* 211–226.

Golgher, A. B., & Voss, P. R. (2016). How to interpret the coefficients of spatial models: Spillovers, direct and indirect effects. *Spatial Demography, 4,* 175–2015.

Greenlee, R. T., & Howe, H. L. (2009). County-level poverty and distant stage cancer in the United States. *Cancer Causes and Control, 20,* 989–1000.

Huang, B., Wu, B., & Barry, M. (2010). Geographically and temporally weighted regression for modeling spatio-temporal variation in house prices. *International Journal of Geographical Information Science, 24*(3), 383–401.

Iceland, J. (2013). *Poverty in America: A handbook* (3rd ed.). Berkeley, CA: University of California.

Jennings, J. (1999). Persistent poverty in the United States: Review of theories and explanations. In Louis Kushnick & James Jennings (Eds.), *A new introduction to poverty: The role of race, power and politics* (pp. 13–38). New York, NY: New York University Press.

Lee, L., & Yu, J. (2010). Some recent developments in spatial panel data models. *Regional Science and Urban Economics, 40*(5), 255–271.

LeSage, J. P. (1999). A spatial econometric examination of China's economic growth. *Geographic Information Sciences, 5,* 143–153.

LeSage, J. P., & Pace, R. K. (2009). *Introduction to spatial econometrics*. Boca Raton, FL: CRC Press.

Levernier, W., Partridge, M. D., & Rickman, D. S. (2000). The causes of regional variations in U.S. poverty: A cross-country analysis. *Journal of Regional Science, 40,* 473–497.

Lichter, D. T., & Johnson, K. M. (2007). The changing spatial concentration of america's rural poor population. *Rural Sociology, 72,* 331–358.

Lobao, L. M., Hooks, G., & Tickamyer, A. R. (2008). Poverty and inequality across space: Sociological reflections on the missing-middle subnational scale. *Cambridge Journal of Regions, Economy and Society, 1,* 89–113.

Lovelace, R., Nowosad, J., & Muenchow, J. (2019). *Geocomputation with R*. Boca Raton: CRC Press.

Nord, M., Luloff, A. E., & Jensen, L. (1995). Migration and the spatial concentration of poverty. *Rural Sociology, 60,* 399–415.

Patton, M., & McErlean, S. (2003). Spatial effects within the agricultural land market in Northern Ireland. *Journal of Agricultural Economics, 54,* 35–54.

Pebesma, E. (2018). Simple features for R: Standardized support for spatial vector data. *The R Journal*, *10*(1), 439–446.

Pebesma, E., Bivand, R., Rowlingson, B., Gomez-Rubio, V., Hijmans, R., Sumner, M., et al. (2018). *Package 'sp'*. R package version 1.3-1. https://cran.r-project.org/web/packages/sp/sp.pdf.

Sandoval, D. A., Mark, R., & Thomas, H. (2009). The increasing risk of poverty across the american life course. *Demography, 46*(4), 717–737.

Sparks, C. (2013a). Spatial analysis in R: Part 1. *Spatial Demography, 1,* 131–139.

Sparks, C. (2013b). Spatial analysis in R: Part 2. *Spatial Demography, 1,* 219–226.

Thiede, B., Kim, H., & Valasik, M. (2018). The spatial concentration of America's rural poor population: A postrecession update. *Rural Sociology, 83,* 109–144.

Tickamyer, A. R., & Duncan, C. M. (1990). Poverty and opportunity. *Annual Review of Sociology, 16,* 67–86.

Vaughan, A. S., Rosenberg, E., Shouse, R. L., & Sullivan, P. S. (2014). Connecting race and place: A county-level analysis of White, Black, and Hispanic HIV prevalence, poverty, and level of urbanization. *American Journal of Public Health, 104,* 77–84.

Voss, P. R., Long, D. D., Hammer, R. B., & Friedman, S. (2006). County child poverty rates in the US: A spatial regression approach. *Population Research and Policy Review, 25,* 369–391.

Walker, K. (2018). Tigris: Load census TIGER/Line Shapefiles. R package version 0.7. https://CRAN.R-project.org/package=tigris.

Weber, B., Jensen, L., Miller, K., Mosley, J., & Fisher, M. (2005). A critical review of rural poverty literature: Is there truly a rural effect? *International Regional Science Review, 28,* 381–414.

Wickham, H. (2016). *ggplot2: Elegant graphics for data analysis*. New York: Springer.

Wickham, W., Hester, J., François, R. (2018a). readr: Read Rectangular Text Data. R package version 1.3.1. https://CRAN.R-project.org/package=readr.

Wickham, H., François, R., Henry, L., Müller, K. (2018b). dplyr: A Grammar of Data Manipulation. R package version 0.7.8. https://CRAN.R-project.org/package=dpylr.

Wimberley, R. C., & Morris, L. (2002). The regionalization of poverty: Assistance for the black belt south? *Southern Rural Sociology, 18,* 294–306.