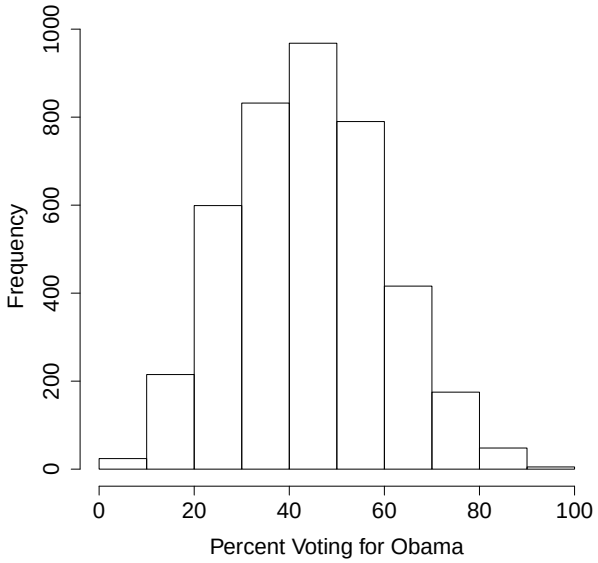
Stats 253 Summer 2015
Homework 2: Autoregressive Processes
Due: Friday, July 24

Policies

- Please submit all homework files through the online homework system on the course website. You can find more instructions there.
- You may work on this homework in pairs. If you choose to do so, one person should submit all of the files (writeup, code, predictions, etc.). However, both partners should submit a statement indicating their contributions.

Description

In this assignment, we will explore the possibility of spatial autocorrelation in the U.S. 2012 presidential election results. In particular, we will model the percentage of people who voted for Barack Obama in each county, which plausibly follows a normal distribution:



Note that each county in the U.S. is assigned a unique ID called a FIPS code. This will allow you to merge different data sources.

Tasks

1. Load `2012USElection.csv` into `R`. We will analyze the lower 48 states only, so please remove Alaska and Hawaii from this data set. Also, note that this file contains both county data and state data. We will only analyze county-level data and not the aggregated state-level data.

2. Make a map, where the color of each county represents the percentage voting for Barack Obama. There are two ways to do this: (1) using the `map()` function within the `maps` package in `R`, specifying the colors in the `col` argument in the appropriate order, or (2) read the shapefiles provided in `UScounties.zip` into `R` using `readShapeSpatial` in the `maptools` package. If you choose the second option, you should read Chapter 3 in Bivand *et al.* on "Visualizing Spatial Data." (You can access a PDF on this book for free online through Stanford libraries. The link is on the course webpage.)

3. We will use `spautolm` in the `spdep` package in `R` to fit either the SAR or CAR models. This function takes in a formula, much like the `lm` function, as well as a `listw` object specifying the weight matrix W. Form a `listw` object for the counties. You may find it easier to form a matrix or an `nb` object first and then use the `mat2listw` or `nb2listw` objects. Make sure the ordering of rows in the `listw` object matches the ordering of the observations in your data frame.

   To form the `listw` object, you need to know which counties are adjacent to which other counties. You can find this information on the U.S. Census Bureau's website:

   `https://www.census.gov/geo/reference/county-adjacency.html`

   Of course, the best way to map this to your main data (the election results) is using the FIPS code.

4. Now call `spautolm` using the election data and the `listw` object you just formed to fit a SAR or CAR model (your choice). We will start by fitting an intercept-only model, i.e., `y ~ 1`. Report the fitted model.

5. Now let's introduce some predictors into the model. To do this, you will have to find additional information about each county, such as the population, average income, etc. The U.S. Census Bureau also provides this information:

   `http://quickfacts.census.gov/qfd/download_data.html`

   Again, the best way to align this data with the election data is using the FIPS code. You may want to look at the `merge()` function in `R` for merging two data frames along a column.

6. Now make a linear model for the trend using the predictors you just obtained and call `spautolm` with this model. Report the fitted model. Are you able to make the spatial autocorrelation go away with a sufficiently good trend model?