Lecture 2
Autoregressive Processes

Dennis Sun
Stats 253

June 25, 2014

Agenda

Outline of Lecture

1. Last Class
2. Bootstrap Standard Errors
3. Maximum Likelihood Estimation
4. Spatial Autoregression
    - Case Study
    - Simultaneous vs. Conditional Autoregression
    - Non-Gaussian Data
5. Wrapping Up

Where are we?

Motivation for AR processes

- The linear regression model

$$\boldsymbol{y} = X\boldsymbol{\beta} + \boldsymbol{\epsilon},\ \boldsymbol{\epsilon} \sim N(0, \sigma^2 I)$$

  assumes observations y_t are independent.
- We can introduce dependence by adding a lag term:

$$y_t = \boldsymbol{x}_t^T \boldsymbol{\beta} + \phi y_{t-1} + \epsilon_t$$

Least Squares Estimation

- We can still estimate $\boldsymbol{\beta}$ and ϕ by least squares:

$$\begin{bmatrix} y_2 \\ \vdots \\ y_n \end{bmatrix} = \begin{bmatrix} & & y_1 \\ \text{---}X_{2:n}\text{---} & & \vdots \\ & & y_{n-1} \end{bmatrix} \begin{bmatrix} | \\ \boldsymbol{\beta} \\ | \\ \phi \end{bmatrix} + \boldsymbol{\epsilon}$$

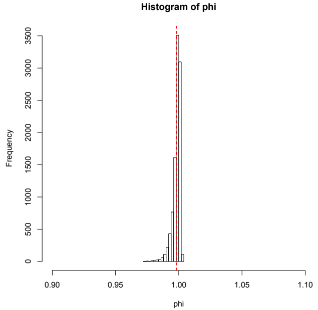
- Advantages: consistent estimate of $\boldsymbol{\beta}$ and ϕ
- Disadvantages: discard 1 observation, standard errors are incorrect

Simulation Study

- Simulated many instances of a length 1000 random walk

$$y_t = \phi y_{t-1} + \epsilon_t, \ \phi = 1$$

- Estimate ϕ by autoregression.



$$\text{Var}(\hat{\phi}) = .003$$

Simulation Study

- Simulated many instances of a length 1000 random walk

$$y_t = \phi y_{t-1} + \epsilon_t, \ \phi = 1$$

- Estimate ϕ by autoregression.

```
Call:
lm(formula = y ~ x - 1)

Residuals:
    Min      1Q  Median      3Q     Max
-3.2497 -0.6678  0.0396  0.6699  4.3311

Coefficients:
  Estimate Std. Error t value Pr(>|t|)
x 0.998757   0.001835   544.3   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.9882 on 998 degrees of freedom
Multiple R-squared: 0.9966, Adjusted R-squared: 0.9966
F-statistic: 2.962e+05 on 1 and 998 DF,  p-value: < 2.2e-16
```

Good (0.998757) — Bad (0.001835)

Million dollar question

How do we obtain correct standard errors?

The (Parametric) Bootstrap

- In the simulation, we knew ϕ and so were able to simulate many instances of

$$y_t = \phi y_{t-1} + \epsilon_t$$

to estimate $\mathrm{Var}(\hat{\phi})$.

- In practice, we do not know ϕ—that's why we're estimating it!
- **Idea:** We have a (pretty good) estimate of ϕ. Why not simulate many instances of

$$y_t = \hat{\phi} y_{t-1} + \epsilon_t$$

to estimate $\mathrm{Var}(\hat{\phi})$?

- This is the **(parametric) bootstrap**.

Where are we?

Review of the MLE

- Another general approach for estimating parameters is **maximum likelihood estimation**.
- The likelihood is the probability distribution, viewed as a function of ϕ:

$$L(\phi) \stackrel{def}{=} p(y_1, ..., y_n|\phi)$$

- The MLE estimates ϕ by choosing the ϕ maximizes L for the observed data:

$$\hat{\phi}_{mle} = \underset{\phi}{\operatorname{argmax}} \; \log L(\phi)$$

MLE of an AR process

We need to calculate $p(y_1, ..., y_n|\phi)$.

$$p(y_1, ..., y_n) = p(y_1) \cdot p(y_2|y_1) \cdot p(y_3|y_1, y_2) \cdot ... \cdot p(y_n|y_1, ..., y_{n-1}).$$

Recall that for an AR process, we have $y_t = \phi y_{t-1} + \epsilon_t$.

$$p(y_t|y_1, ..., y_{t-1}) = p(y_t|y_{t-1}) \text{ for t=2, ..., n}$$

is the density of a $N(\phi y_{t-1}, \sigma^2)$.

$$p(y_t|y_{t-1}) = \frac{1}{\sigma\sqrt{2\pi}} \exp\left\{-\frac{1}{2\sigma^2}(y_t - \phi y_{t-1})^2\right\}$$

Putting it all together, we have:

$$p(y_1, ..., y_n) = p(y_1) \cdot \left(\frac{1}{\sigma\sqrt{2\pi}}\right)^{n-1} \exp\left\{-\frac{1}{2\sigma^2}\sum_{t=2}^{n}(y_t - \phi y_{t-1})^2\right\}$$

MLE of an AR process

$$p(y_1, ..., y_n) = p(y_1) \cdot \left(\frac{1}{\sigma\sqrt{2\pi}}\right)^{n-1} \exp\left\{-\frac{1}{2\sigma^2}\sum_{t=2}^{n}(y_t - \phi y_{t-1})^2\right\}$$

- The log-likelihood is:

$$\log p(y_1) - (n-1)\log(\sigma\sqrt{2\pi}) - \frac{1}{2\sigma^2}\sum_{t=2}^{n}(y_t - \phi y_{t-1})^2$$

  and we **maximize** this over ϕ.
- How does this compare with regression (least squares)?
- In least squares, we **minimize**

$$\sum_{t=2}^{n}(y_t - \phi y_{t-1})^2.$$

- Maximum likelihood and least squares are identical for AR time series!

Summary

- Maximum likelihood is another "recipe" for coming up with a good estimator.
- The MLE for an AR process turns out to be the same as the least squares estimator.

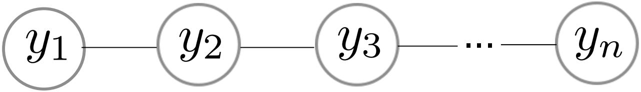
$$\hat{\phi} = \hat{\phi}_{mle}$$

- The parametric bootstrap is a general way to get an estimate of $\mathrm{Var}(\hat{\phi})$.

Where are we?

Graphical Representation of AR(1) process

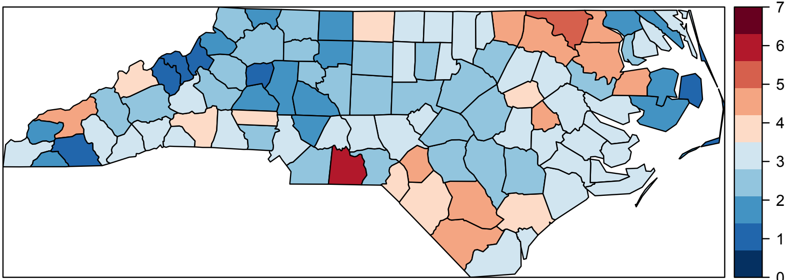
AR(1) process: $y_t = \phi y_{t-1} + \epsilon_t$



An edge between y_i and y_j indicates that y_i and y_j are dependent, conditional on the rest.

North Carolina SIDS Data

- Sudden infant death syndrome (SIDS): unexplained infant deaths.
- Is it genetic? environmental? random?
- Number of SIDS cases $S_i, i = 1, ..., 100$ collected for 100 North Carolina counties.

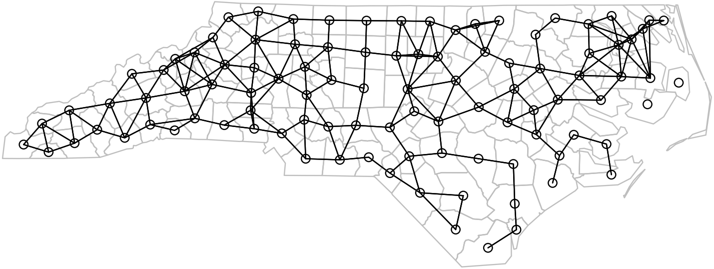


Freeman-Tukey transformed data:

$$y_i = (1000 S_i/n_i)^{1/2} + (1000(S_i + 1)/n_i)^{1/2}$$

An Autoregressive Model

Let's try to model this as a spatial process.



Let $N(i)$ denote the neighbors of county i. Consider the model:

$$y_i - \mu_i = \phi \frac{1}{|N(i)|} \sum_{j \in N(i)} (y_j - \mu_j) + \epsilon_i,$$

where e.g., $\mu_i = \boldsymbol{x}_i^T \boldsymbol{\beta}$. What happens if $\phi = 0$?

Estimating Parameters

$$y_i - \mu_i = \phi \frac{1}{|N(i)|} \sum_{j \in N(i)} (y_j - \mu_j) + \epsilon_i$$

- Should we estimate parameters by least squares? **No! It's inconsistent.** (Whittle 1954)
- Let's try maximum likelihood.
  - First, write in vector notation as

$$\boldsymbol{y} - \boldsymbol{\mu} = \phi W (\boldsymbol{y} - \boldsymbol{\mu}) + \boldsymbol{\epsilon}$$

$$(I - \phi W)(\boldsymbol{y} - \boldsymbol{\mu}) = \boldsymbol{\epsilon}$$

so $\boldsymbol{y} = \boldsymbol{\mu} + (I - \phi W)^{-1} \boldsymbol{\epsilon} \sim N(\boldsymbol{\mu}, (I - \phi W)^{-1} \sigma^2 I (I - \phi W^T)^{-1})$.
  - Now we can write down the likelihood and maximize it.

Data Analysis

R Code:

```
model <- spautolm(ft.SID74 ~ 1, data=nc,
                   listw=nb2listw(neighbors, zero.policy=T))
summary(model)
```

R Output:

```
Coefficients:
            Estimate Std. Error z value  Pr(>|z|)
(Intercept)   2.8597     0.1445  19.791 < 2.2e-16

Lambda: 0.38891 LR test value: 11.286 p-value: 0.00078095
Numerical Hessian standard error of lambda: 0.10761

Log likelihood: -133.3255
ML residual variance (sigma squared): 0.80589, (sigma: 0.89771)
Number of observations: 100
Number of parameters estimated: 3
AIC: 272.65
```

Data Analysis

R Code:

```
model <- spautolm(ft.SID74 ~ ft.NWBIR74, data=nc,
                  listw=nb2listw(neighbors, zero.policy=T))
summary(model)
```

R Output:

```
Coefficients:
             Estimate Std. Error z value  Pr(>|z|)
(Intercept) 1.5444201  0.2161106  7.1464 8.906e-13
ft.NWBIR74  0.0416524  0.0060981  6.8303 8.471e-12

Lambda: 0.083728 LR test value: 0.38241 p-value: 0.53632
Numerical Hessian standard error of lambda: 0.13428

Log likelihood: -117.7629
ML residual variance (sigma squared): 0.616, (sigma: 0.78486)
Number of observations: 100
Number of parameters estimated: 4
AIC: 243.53
```

Different Specifications?

- Previously, we considered the **simultaneous** specification:

$$y_i - \mu_i = \phi \frac{1}{|N(i)|} \sum_{j \in N(i)} (y_j - \mu_j) + \epsilon_i$$

- We might also consider the **conditional** specification:

$$y_i \Big| (y_j, j \in N(i)) \sim N\left(\mu_i + \phi \frac{1}{|N(i)|} \sum_{j \in N(i)} (y_j - \mu_j), \quad \sigma^2\right)$$

- Issues:
  - Are the two specifications equivalent?
  - Is the conditional specification even well defined?

Difficulties with the Conditional Specification

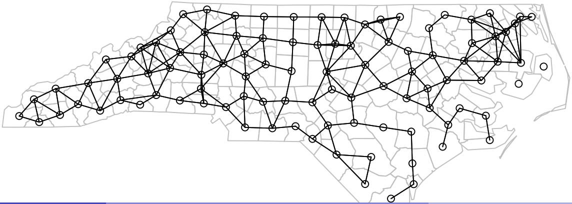
- Recall that with temporal data, we had the conditional specification

$$y_t \Big| (y_1, ..., y_{t-1}) \sim N(\mu_t + \phi y_{t-1}, \sigma^2)$$

- We were able to write the joint distribution in terms of these conditionals using:

$$p(y_1, ..., y_n) = p(y_1) \cdot p(y_2|y_1) \cdot ... \cdot p(y_n|y_1, ..., y_{n-1})$$

- This formula doesn't help us here.



Difficulties with the Conditional Specification

- In general, given a set of conditionals $p(y_i|y_j, j \neq i)$, there does not necessarily exist a joint distribution $p(y_1, ..., y_n)$ with those conditionals.
- However, in this case, we can show that

$$\boldsymbol{y} \sim N(\boldsymbol{\mu}, (I - \phi W)^{-1}\sigma^2 I)$$

Data Analysis

R Code:

```
model <- spautolm(ft.SID74 ~ ft.NWBIR74, data=nc,
                  listw=nb2listw(neighbors, zero.policy=T), family="CAR")
summary(model)
```

R Output:

```
Coefficients:
             Estimate Std. Error z value  Pr(>|z|)
(Intercept) 1.5446517  0.2156409  7.1631 7.889e-13
ft.NWBIR74  0.0416498  0.0060856  6.8440 7.704e-12

Lambda: 0.078486 LR test value: 0.3631 p-value: 0.54679
Numerical Hessian standard error of lambda: 0.12741

Log likelihood: -117.7726
ML residual variance (sigma squared): 0.6151, (sigma: 0.78428)
Number of observations: 100
Number of parameters estimated: 4
AIC: 243.55
```

What to do about non-Gaussian data?

- What if instead of

$$y_i \big| (y_j, j \in N(i)) \sim N\left(\mu_i + \phi \frac{1}{|N(i)|} \sum_{j \in N(i)} (y_j - \mu_j), \quad \sigma^2\right)$$

  we had

$$y_i \big| (y_j, j \in N(i)) \sim \text{Pois}\left(\mu_i + \phi \frac{1}{|N(i)|} \sum_{j \in N(i)} (y_j - \mu_j)\right)?$$

- Issues:
  - Impossible to write down joint distribution.
  - Challenging to simulate.

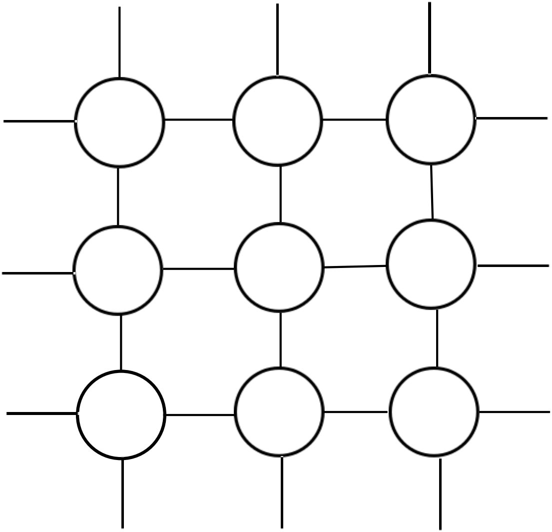
Some Preliminary Solutions

- **Simulation:** Gibbs sampler
  Start with an initial $(y_1, ..., y_n)$, simulate sequentially:
  - $y_1 | y_j, j \neq 1$
  - $y_2 | y_j, j \neq 2$

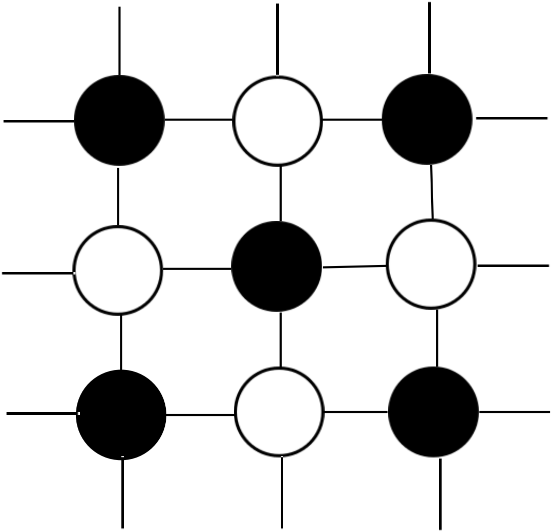
    ...
  - $y_n | y_j, j \neq n$

  and repeat.
  In the long run, the samples $\boldsymbol{y} = (y_1, ..., y_n)$ will be samples from the joint distribution.
- **Estimation:** coding and pseudo-likelihood

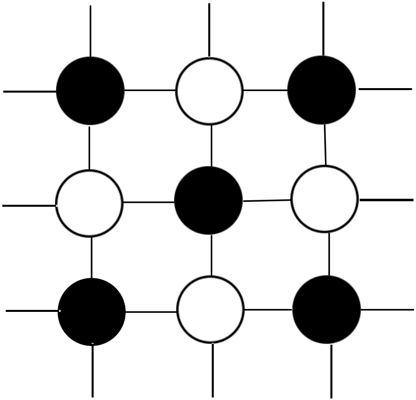
Coding



Coding



Coding



- Consider maximizing the *pseudo*-likelihood $\tilde{L}(\phi) = p(\boldsymbol{y}_{black}|\boldsymbol{y}_{white})$.
- This is easy because the y_i's at the black nodes are **independent**, given the y_i's at the white nodes.

Where are we?

What We've Learned

- The (parametric) bootstrap can be used to get valid standard errors.
- The MLE is a general way of coming up with an estimator: equivalent to least squares in the temporal case, but better in the spatial case.
- There are two similar, but different formulations of spatial autoregression: simultaneous and conditional.
- Things are easiest in the Gaussian setting, but Gibbs sampling and coding can be used with non-Gaussian data.

Administrivia

- Piazza
- Enrollment cap?
- Homework 1: autoregression and bootstrap
  - Will be posted by tomorrow night.
  - Remember that you can work in pairs! (Hand in only one problem set per pair.)
  - Will be graded check, resubmit, or zero.
- Edgar will be lecturing next Monday on R for spatial data.
- Jingshu and Edgar will be holding workshops starting next week.