Time Series Analysis

Lecture 3

Autoregressive Models and Moving Average Models

datascience@berkeley

Autoregressive Models

Example 2: Estimation, Model Selection, Model Diagnostics, and Assumption Testing

Estimation and the True Data-Generating Process

Let's apply an AR model to the series we just simulated.

Recall that the "true" model (or true underlying data-generating process (DGP)) has the following functional form:

$$y_t = 1.5y_{t-1} - 0.9y_{t-2} + \epsilon_t$$

As we have already examine the series, we will proceed directly to estimation.

Pretend that we do not have knowledge about the underlying DGP, but based on the histogram, t-plot, ACF, and PACF, we will try using a autoregressive model.

Estimation: R Commands

We will first apply the `AR()` function, which selects the model with the lowest AIC.
Similar to example 1, the following commands are used to (1) estimate a series of AR model, (2) list the objects within the estimated AR objects, (3) determine the order of the AR model with the lowest AIC, (4) determine the estimated AR parameters, and (5) determine the difference between AICs and the best AIC.

```
x3.arfit <- ar(x3, method = "mle")
summary(x3.arfit)

x3.arfit$order # order of the AR model with lowest AIC
x3.arfit$ar # parameter estimate
x3.arfit$aic # AICs of the fit models
sqrt(x3.arfit$asy.var) # asy. standard error
x3.arfit$mean
```

Estimation Results

The results:

```
> x3.arfit$order # order of the AR model with lowest AIC
[1] 5
> x3.arfit$ar # parameter estimate
[1]  1.4787350 -0.9085119 -0.0658531  0.1367330 -0.1021317
> x3.arfit$aic # AICs of the fit models
          0           1           2           3           4           5 
2516.786599 1579.744987    4.590775    6.550943    8.410185    0.000000 
          6           7           8           9          10          11 
   1.778044    2.165644    2.169896    4.129782    5.383153    7.338368 
         12 
   9.025748 
> sqrt(x3.arfit$asy.var) # asy. standard error
            [,1]       [,2]       [,3]       [,4]        [,5]
[1,] 0.031273080        NaN 0.02990534        NaN 0.004045091
[2,]         NaN 0.05587106        NaN 0.03040702         NaN
[3,] 0.029905338        NaN 0.06275694        NaN 0.029905338
[4,]         NaN 0.03040702        NaN 0.05587106         NaN
[5,] 0.004045091        NaN 0.02990534        NaN 0.031273080
Warning message:
In sqrt(x3.arfit$asy.var) : NaNs produced
```

Estimation Results Explained

Several points worth noticing:

1. A series of 12 AR models were estimated. This is the default set in the `AR()` function. We can change the maximum order using the option `order.max`. See the R documentation for more details.

2. The "best" model (in terms of AIC) is an AR(5) while the underlying DGP is an AR(2) process, although the first two parameters (1.48 and -0.91 with standard errors being 0.0312 and 0.0559) are very similar to those of the true DGP.

3. The AICs of the AR(0) and AR(1) models are huge, but those of AR(2), AR(5)–AR(10) are within the same ballpark.

4. Some of the off-diagonal elements in the (asymptotic) variance-covariance matrix cannot be computed.

Model Diagnostics

Examine the residuals:

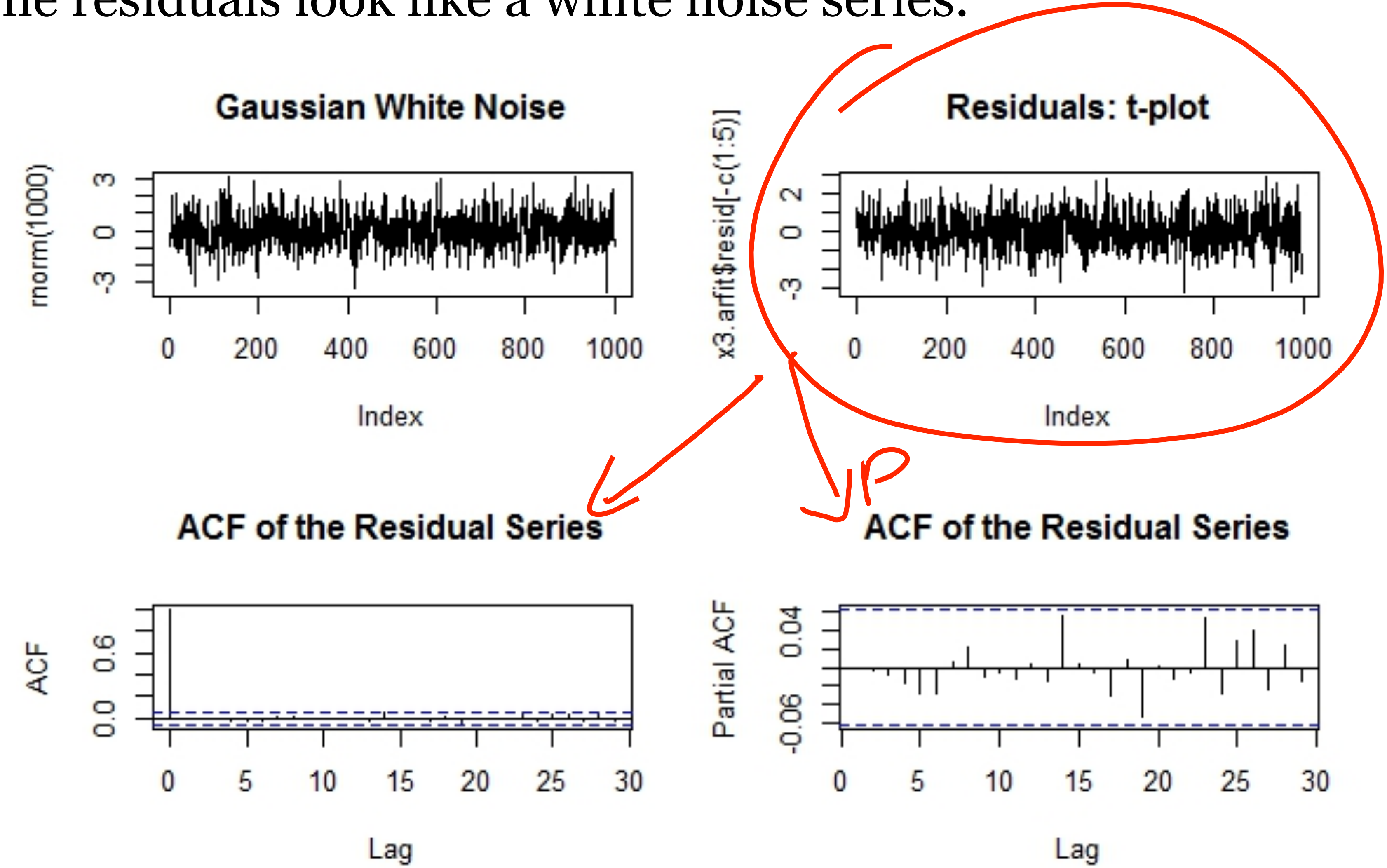
Note that the first five observations in the estimated residual series are missing. It is because the first five observations from the sample path are excluded in the estimation of an AR(5) model.

For this reason, to plot ACF and PACF, these missing observations from the series need to be excluded.

```
> head(x3.arfit$resid, 15)
 [1]          NA          NA          NA          NA          NA  1.19943972
 [7] -0.74451763  0.04727915  0.79347539  0.94224939  0.55571479 -0.81977882
[13]  0.34283689  0.26066242  0.86765075
> head(x3.arfit$resid[-c(1:5)], 15)
 [1]  1.19943972 -0.74451763  0.04727915  0.79347539  0.94224939  0.55571479
 [7] -0.81977882  0.34283689  0.26066242  0.86765075  0.02029650 -0.43598030
[13] -0.19129438  1.37263278  0.06466725
```

Model Diagnostics

The residuals look like a white noise series.



Model Selection

These results call into the question of which model (AR(5) or AR(2)) should we use.

- The AR(5) model gives a lower AIC.
- Or, the AR(2) is the “true” model.
  - But, of course, in reality we do not know what the true model really is.

Before we make other considerations, remember:

1. AIC is but one model performance measure.
2. In this example, we have only used it for in-sample measure (i.e., goodness-of-fit).

Model Selection (2)

Other considerations:

1. Are the differences among the AICs from the several AR models really that big?

2. Even if the focus is “in-sample” fit, we should consider other model performance measure such as Bayesian information criteria, which penalize more on the number parameters used in the model: `-2ln(likelihood) + ln(N)k`
   - It is possible that AIC and BIC give different results. We will discuss AIC and BIC more in the next lecture.

3. What about out-of-sample test?
   - We have not talked about out-of-sample tests.

4. Among all of these considerations, the most important question we need to ask is what is the objective of building this model? What question are we trying to answer?

Model Selection (3)

Suppose the objective is for forecasting. Then, the performance measure should be focusing on using forecast errors to compare among models.

- We will have to ask how long the forecast horizon is.
- Is the forecast short-term or long-term?
- How long is the sample we want to use to estimate the model? Do we always want to use the entire sample, as we did in the previous two simulated examples?
- Should we leave a subset of the observed samples for out-of-sample test?
  - ➢ In other courses, you may have come across this concept as dividing the sample into a **training set** and a **testing (or validation) set**. In time series analysis, the sample is often very limited. How much sample to leave out requires a thorough discussion between data scientists/modelers and the business stakeholders.

Berkeley SCHOOL OF INFORMATION