Unit 14: Building ARIMA Models

Jeffrey Woo

Department of Statistics, University of Virginia

Spring 2020

Readings for Unit 14

Textbook chapter 3.6, 3.7.

Last Unit

1. Method of Moments Estimation
2. Maximum Likelihood Estimation

This Unit

1. Integrated models for nonstationary data.
2. Building ARIMA models.

Motivation

Thus far, we have assumed stationary ARMA models. We next consider ARIMA models, which basically consider differences of the observations to coerce stationarity. We then see how to go about building an ARIMA model using techniques we've looked at previously (and some new ones). These techniques are namely: exploratory data analysis (plots, transformations, identifying potential models), model estimation, model diagnostics, and model selection.

1 Integrated Models for Nonstationary Data

2 Building ARIMA Models

3 Worked Example

Integrated Models for Nonstationary Data

Recall when we discussed exploratory data analysis and smoothing that we sometimes began with a model such as

$$x_t = \mu_t + y_t$$

where μ_t was a non-stationary component (trend) and y_t was a stationary zero-mean process.

Integrated Models for Nonstationary Data

If we assume μ_t was of the form $\beta_0 + \beta_1 t$, then differencing yields

$$\nabla x_t = \beta_1 + \nabla y_t$$

which is a stationary time series. In general if μ_t is a polynomial in t, such as $\beta_0 + \beta_1 t + ... + \beta_p t^d$, then

$$\nabla^d x_t = d! \beta_d + \nabla^d y_t.$$

Integrated Models for Nonstationary Data

A process is ARIMA(p, d, q) if

$$\nabla^d x_t = (1 - B)^d x_t$$

is ARMA(p, q). Note: ARIMA models are sometimes called Box-Jenkins models.

Integrated Models for Nonstationary Data

If the mean of $\nabla^d x_t$ is zero then we can write the ARIMA model as

$$\phi(B)(1-B)^d x_t = \theta(B) w_t \tag{1}$$

If a mean does exist then we can write

$$\phi(B)(1-B)^d x_t = \alpha + \theta(B) w_t \tag{2}$$

where $\alpha = \mu(1 - \phi_1 - ... - \phi_p)$.

Integrated Models for Nonstationary Data

We usually use a first difference to account for a **linear** trend in the data. A second difference may be used to account for a **quadratic** trend in the data.

Integrated Models for Nonstationary Data

Recall that differencing can sometimes introduce more **dependence** in the data. Let's assume that our data has the following model

$$x_t = \beta_0 + \beta_1 t + y_t$$

where

$$y_t = \phi_1 y_{t-1} + \phi_2 y_{t-2} + w_t.$$

Consider $\nabla x_t = \beta_1 + \nabla y_t$.

Integrated Models for Nonstationary Data

Define

$$z_t = \nabla y_t = y_t - y_{t-1}$$

So,

$$\nabla y_t = y_t - y_{t-1} = \phi_1(y_{t-1} - y_{t-2}) + \phi_2(y_{t-2} - y_{t-3}) + w_t - w_{t-1}$$

which we could write as

$$z_t = \phi_1 z_{t-1} + \phi_2 z_{t-2} + w_t - w_{t-1}$$

which is ARMA(2,1). Notice that the original noise term y_t is an AR(2).

1. Integrated Models for Nonstationary Data

2. Building ARIMA Models

3. Worked Example

ARIMA Models

We'll explore some techniques for identifying and estimating non-seasonal ARIMA models, as well as how to analyze the residuals after a model is estimated. Recall that ARIMA models are specified as ARIMA(p,d,q).

Elements of ARIMA Models

- The elements in the ARIMA model are typically specified, in the following order: AR order, differencing, MA order.
- If $d = 0$, we may call the model an ARMA(p,q) model.
- If $q = 0$, we have an AR(p) model.
- If $p = 0$, we have an MA(q) model.

Building ARIMA Models

The main steps in building an ARIMA model are

- Exploratory data analysis (plots, transformations, identifying potential models).
- Model estimation.
- Model diagnostics.
- Model selection.

Exploratory Data Analysis

We typically look at

- the time series plot,
- the ACF,
- and the PACF

of the data. This step guides us to our choice for the elements of the ARIMA model, p, d, q.

Exploratory Data Analysis: Time Series Plot

We usually check for stationarity in a time series plot. Recall for stationarity, the plot should suggest the mean and the variance are constant. What we look out for in a time series plot:

- Trend (increasing, decreasing, quadratic).
- Increasing variability.
- Seasonality (talk about next unit).
- Outliers.

Exploratory Data Analysis: Time Series Plot

We usually try to stabilize the variance first. If the variance appears to be increasing, we can try transforming the data, usually a log transform.

Exploratory Data Analysis: Time Series Plot

If there's a linear trend, we consider a first difference. A quadratic trend suggests a second difference. We rarely go beyond $d = 2$, unless there is a contextual or scientific reason to do so.

As mentioned earlier, over-differencing can introduce unnecessary dependency in the model. **Smoothing techniques** may be used, where we analyze the smoothed data instead.

Exploratory Data Analysis: ACF and PACF

The ACF and PACF should be used together. Recall that

- AR(p) models have theoretical PACF with non-zero values for $h \leq p$, and zero values for $h > p$. The ACF should decay exponentially to zero.
- MA(q) models have theoretical ACF with non-zero values for $h \leq q$, and zero values for $h > q$. The PACF should decay exponentially to zero.
- ARMA(p,q) models have ACF and PACF that both decay exponentially to zero. The order will not be obvious. In such a case, we may just start off with $p, q = 1$ or 2, and see what happens during model estimation and diagnostics.

Exploratory Data Analysis: ACF and PACF

- If the ACF and PACF decay slowly (do not decay exponentially), then the time series is likely to be not stationary (was differencing performed earlier?).
- If all the autocorrelations are insignificant for $h \geq 1$, than the series is random (e.g. shifted white noise).
- If all the autocorrelations are insignificant for $h \geq 1$ for a first difference, then we may have a **random walk**.

Model Estimation

After exploratory data analysis, you should have an idea (or ideas) about the values of p, d, q. We use computer software (e.g. R) to estimate the parameters. Maximum likelihood estimation is usually used.

Model Estimation: Significance of Estimates

For the estimated coefficients of the parameters, use the t statistic

$$t = \frac{\text{estimated coefficient}}{\text{s.e. of coefficient}}.$$

If $|t| > t_{1-\alpha/2,df}$, the estimated coefficient is significantly different from 0 at 0.05 significance level.

Model Diagnostics

For model diagnostics, we usually check the following:

- ACF of residuals.
- Ljung-Box-Pierce statistic.
- Plot of residuals against fitted values or time series of residuals.

If our model diagnostics are acceptable, we expect our residuals to behave like **white noise**. If something appears unreasonable, you might have to revise your thought at what the model might be.

Model Diagnostics: ACF of Residuals

If you have a good model, all estimated ACFs of residuals should be **insignificant**. If this isn't the case, you probably need to explore a different model.

Model Diagnostics: Ljung-Box-Pierce Statistic

Recall that for white noise, the sample autocorrelations are approximately independent and normally distributed with mean 0 and variance $\frac{1}{n}$. The Ljung-Box-Pierce Q-statistic takes into account the magnitudes of the sample autocorrelations as a whole.

Model Diagnostics: Ljung-Box-Pierce Statistic

The Ljung-Box-Pierce statistic is a function of accumulated sample autocorrelations, $\hat{\rho}(h)$, up to a specified time lag H. The Ljung-Box-Pierce Q-statistic is given by

$$Q(H) = n(n+2)\sum_{h=1}^{H}\frac{\hat{\rho}(h)^2}{n-h}. \tag{3}$$

The choice of H is somewhat arbitrary. In R, H is 20.

Model Diagnostics: Ljung-Box-Pierce Statistic

Under the null hypothesis that the model fits the data adequately, $Q \sim \chi^2_{H-p-q}$ as $n \to \infty$. A large Q-statistic leads to the rejection of the null hypothesis, i.e. **model is not an adequate fit for the data**.

Model Diagnostics: Residual Plot

Using either a plot of residuals against fitted values, or a time series plot of the residuals, we check if the variance is constant. If the variance is not constant, you may need to transform the data.

Model Selection

Sometimes you may have more than one set of values for p, d, q from exploratory data analysis. To be thorough, you may want to investigate the model estimation and diagnostics for more than one model. If model diagnostics suggest more than one model works, here are some issues to keep in mind when comparing models:

- Simpler model.
- Standard errors of forecasts.
- Smaller MSE, AIC, etc.

Model Selection

Two different ARIMA models can be nearly equivalent when converted to an infinite order MA model using the causal representation.

1 Integrated Models for Nonstationary Data

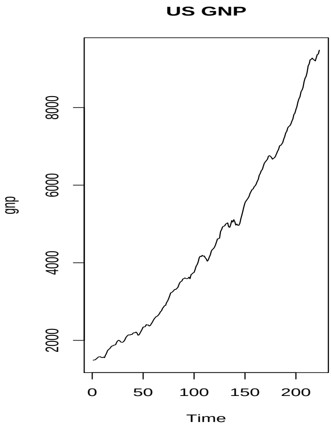
2 Building ARIMA Models

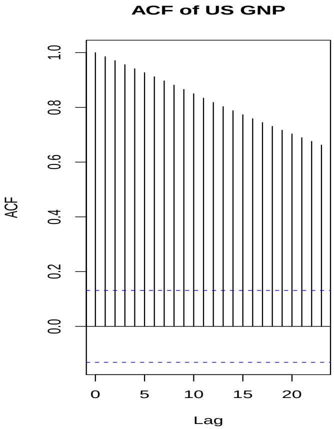
3 Worked Example

Worked Example: US GNP

To tie everything in, we'll look at an example on how to build an ARIMA model. (Don't worry you'll get chances in the homework $\ddot\smile$) We will use data on US gross national product in billions of chained 1996 dollars. The observations are quarterly data from 1947 to 2002.

Worked Example: Exploratory Data Analysis

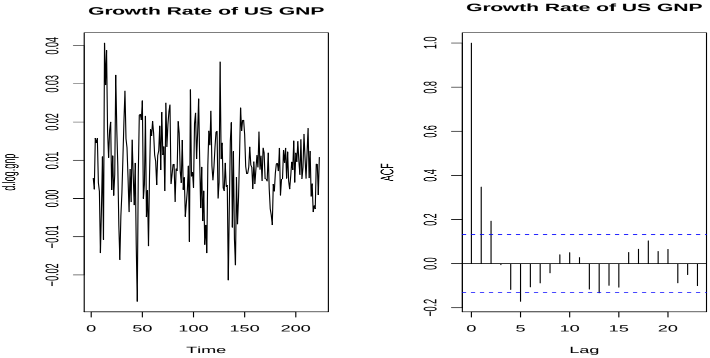

US GNP


ACF of US GNP

Question: Any comments on stationarity?

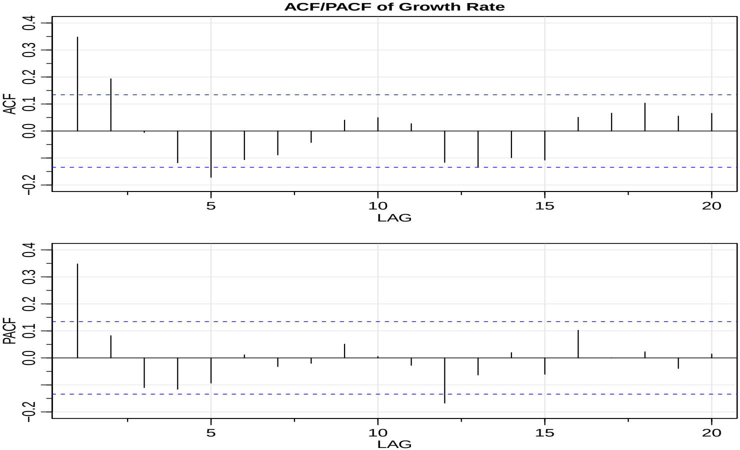
Worked Example: Exploratory Data Analysis

Let's look at growth rate, usually defined as $\nabla \log(x_t)$.



Question: Any comments on stationarity?

Worked Example: Exploratory Data Analysis



Question: What do you think for p, q?

Worked Example: Model Estimation

Fit using `sarima()` function in R.

```
      Estimate     SE t.value p.value
ar1     0.3467 0.0627  5.5255       0
xmean   0.0083 0.0010  8.5398       0
```

Question: How to write this model?

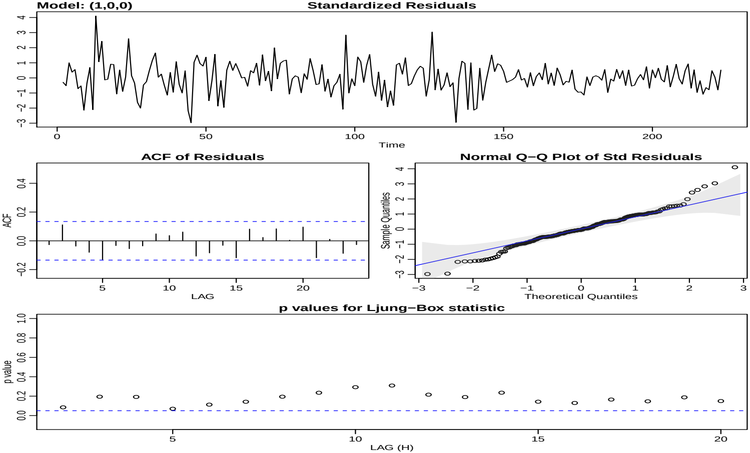
Worked Example: Model Estimation

Fit using `sarima()` function in R.

```
      Estimate     SE t.value p.value
ma1     0.3028 0.0654  4.6272  0.0000
ma2     0.2035 0.0644  3.1594  0.0018
xmean   0.0083 0.0010  8.7178  0.0000
```

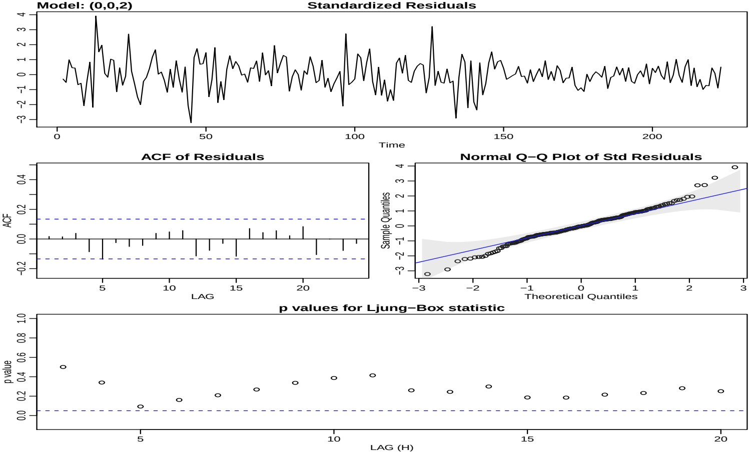
Question: How to write this model?

Worked Example: Model Diagnostics



Diagnostic plots for AR(1) model.

Worked Example: Model Diagnostics



Diagnostic plots for MA(2) model.

Worked Example: Model Selection

AR(1) model.

```
sigma^2 estimated as 9.03e-05:

$AIC
[1] -6.44694

$AICc
[1] -6.446693

$BIC
[1] -6.400958
```

Worked Example: Model Selection

MA(2) model.

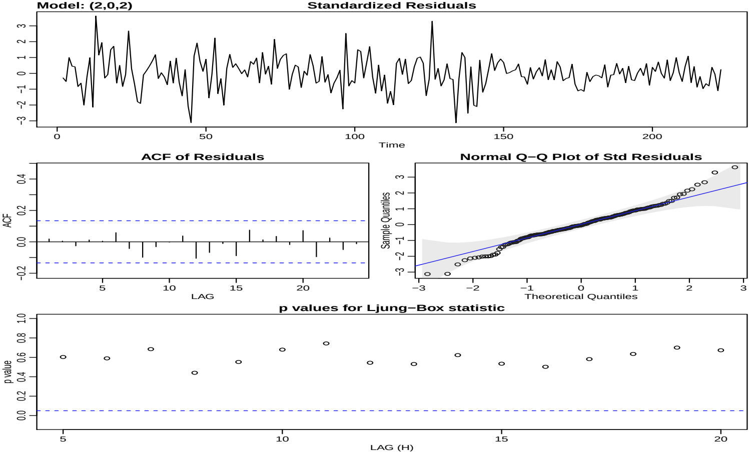
```
sigma^2 estimated as 8.919e-05

$AIC
[1] -6.450133

$AICc
[1] -6.449637

$BIC
[1] -6.388823
```

Worked Example: ARMA Model

How about an ARMA(2,2) model?

```
      Estimate     SE t.value p.value
ar1     1.3459 0.1374  9.7983  0.0000
ar2    -0.7378 0.1540 -4.7923  0.0000
ma1    -1.0634 0.1873 -5.6787  0.0000
ma2     0.5621 0.1971  2.8518  0.0048
xmean   0.0083 0.0008 10.4121  0.0000
```

Worked Example: Model Diagnostics



Diagnostic plots for ARMA(2,2) model.

Worked Example: Model Selection

ARMA(2,2) model.

```
sigma^2 estimated as 8.649e-05:

$AIC
[1] -6.462032

$AICc
[1] -6.46078

$BIC
[1] -6.370067
```

Worked Example: Recap

	AR(1)	**MA(2)**	**ARMA(2,2)**
Est Coeffs	All significant	All significant	All significant
Diagnostics	Ok	Ok	Ok
AIC	-6.44694	-6.450133	**-6.462032**
AICc	-6.446693	-6.449637	**-6.46078**
BIC	**-6.400958**	-6.388823	-6.370067

Question: Which model will you choose? What else could be done to help us decide which model to choose?

A Few More Comments about Model Building

- Real data cannot be **exactly** modeled using a finite number of parameters.
- What we want typically is a simple and accurate model.