# Applied Time Series Analysis

Young K. Truong
Department of Statistics
and Applied Probability
The National University of Singapore

yktruong@math.nus.edu.sg
http://www.bios.unc.edu/~truong
truong@bios.unc.edu

## Time series

1. The construction data set.

2. The lead production data set: totals of US lead production (in tons) for the period January 1986 to September 1992.

3. The Dow Jones Index data set: stock prices of 30 US firms from January 1984 to February 1994.

4. The steel data set: steel (thousands of tons) shipped from US for the period January 1984 to December 1991.

## Creating and plotting time series

- SAS
- SPLUS
- MINITAB

**Modeling Trend**

The trend model is

$$y_t = \mu_t + e_t,$$

where

$$\begin{aligned} y_t &= \text{the value of the time series at time } t, \\ \mu_t &= \text{the trend at time } t, \\ e_t &= \text{the error term at time } t. \end{aligned}$$

## Mathematical models

- No trend:

$$\mu_t = \beta_0.$$

- Linear trend:

$$\mu_t = \beta_0 + \beta_1 t.$$

- Quadratic trend:

$$\mu_t = \beta_0 + \beta_1 t + \beta_2 t^2.$$

- Polynomial trend:

$$\mu_t = \beta_0 + \beta_1 t + \beta_2 t^2 + \cdots + \beta_k t^k.$$

## Detecting autocorrelation

A simple model for autocorrelation is

$$e_t = \phi e_{t-1} + \epsilon_t,$$

where $e_t$ is the residual and $\epsilon_t$ is the random error.

- $\phi > 0$ positive autocorrelation
- $\phi < 0$ negative autocorrelation
- $\phi = 0$ no autocorrelation

**Durbin-Watson statistic**

$$d = \frac{\sum_{t=2}^{n}(e_t - e_{t-1})^2}{\sum_{t=1}^{n} e_t^2}.$$

The statistic $d$ is compared with critical values $d_{L,\alpha}$ and $d_{U,\alpha}$, where $\alpha$ is the level of the test. See Tables A5 and A6.

**Durbin-Watson test for positive correlation**

Consider the hypothesis

$H_0$ : The error terms are not autocorrelated

vs

$H_1$ : The error terms are positively autocorrelated.

1. Reject $H_0$ when $d < d_{L,\alpha}$.

2. Do not reject $H_0$ when $d > d_{U,\alpha}$.

3. The test is inconclusive when

$$d_{L,\alpha} \leq d \leq d_{U,\alpha}.$$

**Example.** For the calculator sales time series, $d = 1.682$, $d_{L,.05} = 1.27$ and $d_{L,.05} = 1.45$. We do not reject $H_0$.

**D-W test for negative correlation**

Consider the hypothesis

$H_0$ : The error terms are not autocorrelated

vs

$H_1$ : The error terms are negatively autocorrelated.

1. Reject $H_0$ when $(4 - d) < d_{L,\alpha}$.

2. Do not reject $H_0$ when $(4 - d) > d_{U,\alpha}$.

3. The test is inconclusive when

$$d_{L,\alpha} \leq (4 - d) \leq d_{U,\alpha}.$$

**Example.** For the calculator sales time series, $4-d = 2.318 > d_{L,.05} = 1.45$. We do not reject $H_0$.

**D-W test for autocorrelation**

Consider the hypothesis

$H_0$ : The error terms are not autocorrelated

vs

$H_1$ : The error terms are autocorrelated.

1. Reject $H_0$ when $d < d_{L,\alpha/2}$ or $(4 - d) < d_{L,\alpha/2}$.

2. Do not reject $H_0$ when $d > d_{L,\alpha/2}$ and $(4 - d) > d_{U,\alpha/2}$.

3. The test is inconclusive when

   $d_{L,\alpha/2} \leq d \leq d_{U,\alpha/2}$ and $d_{L,\alpha/2} \leq (4-d) \leq d_{U,\alpha/2}$.

**Homework.** Carry out the test of autocorrelation for the calculator sales time series.

**Seasonal variation**

Two types of seasonal variation:

- constant seasonal variation
- increasing seasonal variation (Fig 6.14)

We will focus on analyzing the constant one, since the increasing one can be handled via transformation such as log, or power transformation.

Example. Monthly hotel room averages for 1977–1990 (p.312).

**Modeling seasonal variation**

A time series is often modeled as

$$\begin{aligned} y_t &= \{\text{trend}\}_t + \{\text{seas}\}_t + \{\text{error}\}_t \\ &= \mu_t + S_t + e_t. \end{aligned}$$

Trend is handled via regression method.

The seasonal variation is further modeled by

- dummy variables — monthly average,
- trigonometric functions — regression

## Dummy variables

Suppose that there are $L$ seasons (months, quarters ...) per year.

$$\begin{aligned} S_t &= \beta_{s1}x_{s1,t} + \cdots + \beta_{s(L-1)}x_{s(L-1),t} \\ &= \beta_1 x_{1,t} + \cdots + \beta_{L-1}x_{L-1,t} \end{aligned}$$

where, for $j = 1, 2, \ldots, L-1$,

$$x_{sj,t} = \begin{cases} 1 & \text{if time period is season } j, \\ 0 & \text{otherwise.} \end{cases}$$

For example, if $L = 12$ and $t = 2$, then $S_2 = \beta_2$.

For monthly time series, this is basically the method of taking the corresponding monthly averages. Why?

SAS Example. Monthly hotel room averages for 1977–1990.

**Trigonometric model for constant seasonal variation**

$$S_t = \beta_1 \sin\left(\frac{2\pi t}{L}\right) + \beta_2 \cos\left(\frac{2\pi t}{L}\right).$$

Or,

$$\begin{aligned} S_t &= \beta_1 \sin\left(\frac{2\pi t}{L}\right) + \beta_2 \cos\left(\frac{2\pi t}{L}\right) \\ &\quad + \beta_3 \sin\left(\frac{4\pi t}{L}\right) + \beta_4 \cos\left(\frac{4\pi t}{L}\right). \end{aligned}$$

**Trigonometric model for increasing seasonal variation**

$$
\begin{aligned}
S_t &= \beta_1 \sin\left(\frac{2\pi t}{L}\right) + \beta_2 \cos\left(\frac{2\pi t}{L}\right) \\
&\qquad + \beta_3 t \sin\left(\frac{2\pi t}{L}\right) + \beta_4 t \cos\left(\frac{2\pi t}{L}\right).
\end{aligned}
$$

Or,

$$
\begin{aligned}
S_t &= \beta_1 \sin\left(\frac{2\pi t}{L}\right) + \beta_2 \cos\left(\frac{2\pi t}{L}\right) \\
&\qquad + \beta_3 t \sin\left(\frac{2\pi t}{L}\right) + \beta_4 t \cos\left(\frac{2\pi t}{L}\right) \\
&\qquad + \beta_5 \sin\left(\frac{4\pi t}{L}\right) + \beta_6 \cos\left(\frac{4\pi t}{L}\right) \\
&\qquad + \beta_7 t \sin\left(\frac{4\pi t}{L}\right) + \beta_8 t \cos\left(\frac{4\pi t}{L}\right).
\end{aligned}
$$

SAS Example. Monthly hotel room averages for 1977–1990.

## Stationary time series

A time series is *stationary* if its stochastic behavior do not change with time. This implies that it has the following characteristics:

- constant mean,
- constant variance,
- constant lagged covariance or correlation.

Any violation of the above will be called *non-stationary*.

Sometimes nonstationarity can be handled by differencing the series:

$$z_t = x_t - x_{t-1}.$$

## Sample autocovariance function

Given a time series $x_1, x_2, \ldots, x_n$. We define the sample autocovariance function as

$$\begin{aligned} \widehat{c}(u) \equiv \widehat{c}^n(u) &= \frac{1}{n}\sum_{t=1}^{n-u}(x_{t+u} - \bar{x})(x_t - \bar{x}) \\ \widehat{c}(-u) &= \widehat{c}(u), \end{aligned}$$

where

$$\bar{x} = \frac{1}{n}\sum_{t=1}^{n} x_t.$$

Similarly, the sample autocorrelation function is defined by

$$\widehat{\rho}(u) = \frac{\widehat{c}(u)}{\widehat{c}(0)}.$$

In practice, the sample autocorrelation function is widely used than the autocovariance function because of its asymptotic properties.

The standard error of $\widehat{\rho}(u)$ is

$$\mathrm{se}(\widehat{\rho}(u)) = \sqrt{\frac{1 + 2\sum_{j=1}^{u-1} \widehat{\rho}^2(u)}{n}}.$$

The test statistic is

$$t_u = \frac{\widehat{\rho}(u)}{\mathrm{se}(\widehat{\rho}(u))}.$$

Examples using SAS.

**Moving average models**

$$x(t) = e(t) - \theta e(t-1)$$

where $x(t)$ is the time series at time $t$, and $e(t)$ is the unobserve random error.

More generally,

$$x(t) = \delta + e(t) - \theta e(t-1).$$

An estimate of $\theta$ can be obtained by solving

$$\rho(1) = -\frac{\theta}{1+\theta^2}.$$

See my lecture note.

The ACF plot cuts off after lag 1. This is an important property for identification. (See also Table 9.4).

Examples 9.5 and 9.6 — Absorbent Paper Towel sales in Table 9.1.

**Autoregressive models**

$$x(t) = \phi x(t-1) + e(t).$$

From my note (p.24),

$$\phi \approx \rho(0).$$

Thus

$$\widehat{\phi} \approx \widehat{\rho}(0).$$

The PACF plot cuts off after lag 1 — an important property for identification. (See also Table 9.4).

Example 9.7 with data given in Table 9.3.

## SAS examples

```
DATA mydata;
INPUT Y;
DATALINES;
...

PROC ARIMA DATA=mydata;
I VAR = Y;  * I = IDENTIFY
I VAR = Y(1);  * First difference
I VAR = Y(1,1); * Second difference
RUN;
```

```
/*    1 January 1960    */
 data _null_;
   date = '28sep98'd;
   put date=; run;

/*  Log window  */
. . .

DATE=14150

. . .
t=intnx( 'month', '31dec1985'd, _n_ );
format t monyy.;
. . .
```

```
goptions cback=white colors=(black) border reset=(axis symbol);

axis1 offset=(1 cm)
      label=('Year') minor=none
      order=('01jan1986'd to '01jan1993'd by year);
axis2 label=(angle=90 'Lead Production')
      vminor=1
      order=(15000 to 50000 by 5000);
```

Box-Jenkins Seasonal Modeling (Chapter 10)

$$x_t - \sum_{j=1}^{p} \phi_j x_{t-j} = w_t + \sum_{k=1}^{q} \theta_k w_{t-k}, \qquad w_k \text{iid}(0, \sigma^2).$$

**Causality**

This comes from the AR process. The AR characteristic polynomial $\phi(z)$ has no roots inside the unit circle of the complex plane. That is, $\phi(\cdot)$ satisfies

$$\phi(z) \neq 0, \quad z \in C, \quad |z| \leq 1.$$

**Invertibility**

This comes from the MA component, for which the characteristic polynomial satisfies

$$\theta(z) := 1 + \theta_1 z + \cdots + \theta_q z^q \neq 0, \quad |z| \leq 1.$$

The ARMA process can be abbreviated as

$$\phi(B)x_t = \theta(B)w_t, \qquad \text{or} \qquad x_t = \phi^{-1}(B)\theta(B)w_t.$$

- $t$ and $P$ values

- Multicolinearity and SE (p.495)

- $Q^*$-statistics (Ljung-Box) p 496

- Read Section 10.4: A case study.

Box-Jenkins Seasonal Modeling (Chapter 11)

- Transformation: logarithm, power ...
  - Behavior of SAC at the nonseasonal and seasonal levels (p 527)

- Lags 1 to (L-3) corresponds to the nonseasonal lags

- Lags 1 to 3 corresponds to the low nonseasonal lags

- Lags 4 to (L-3) corresponds to the high nonseasonal lags

- Read page 528

Differencing ...

$$Bx_t = x_{t-1}, \quad B^p x_t = x_{t-p}.$$

We use $(1 - B)x_t$ to remove the (linear) trend, and $(1 - B^{12})x_t$ to remove the seasonal effect.

Two examples

- Monthly hotel room averages

- Airline passengers

`p=4` means $(1-\phi_1 B-\phi_2 B^2-\phi_3 B^3-\phi_4 B^4)$

`p=(4)` means $(1-\phi_4 B^4)$

`p=( 1 12 13 )` means $(1-\phi_1 B-\phi_{12} B^{12}-\phi_{13} B^{13})$

`p=( 1 ) ( 12 )` means $(1-\phi_1 B)(1-\phi_{12} B^{12})$

ARIMA $(p,d,q) \times (P,D,Q)_s$

The following model is usually used to fit seasonal data:

$$\Phi(B^s)\phi(B)\nabla_s^D \nabla^d X_t = \Theta(B^s)\theta(B)e_t,$$

where

$$\nabla^d = (1-B)^d, \qquad \nabla_s^D = (1-B^s)^D,$$

$$\phi(z) = 1 - \phi_1 z - \cdots - \phi_p z^p,$$

$$\theta(z) = 1 + \theta_1 z + \cdots + \theta_q z^q$$

$$\Phi(z) = 1 - \Phi_1 z - \cdots - \Phi_P z^P,$$

$$\Theta(z) = 1 + \Theta_1 z + \cdots + \Theta_Q z^Q.$$

Steps for tentatively identifying a model.

1. Identify a nonseasonal model using SAC and SPAC

2. Identify a seasonal model using SAC and SPAC

3. Combine the above models to yield an overal tentative model.

Heuristics in modeling the airline data:

$$Y_t = m_t + S_t + e_t$$

then

$$Z_t = Y_t - Y_{t-1} \doteq [S_t - S_{t-1}] + [e_t - e_{t-1}],$$

$$\begin{aligned} Z_t - Z_{t-12} \;\; &\doteq \;\; [S_t - S_{t-12}] - [S_{t-1} - S_{t-13}] \\ &\qquad + [e_t - e_{t-12}] - [e_{t-1} - e_{t-13}]. \end{aligned}$$

We approximate $[e_t - e_{t-12}] - [e_{t-1} - e_{t-13}]$ by

$$\begin{aligned} &(1 - \theta_1 B)(1 - \theta_{12} B^{12}) e_t \\ &\qquad = e_t - \theta_1 e_{t-1} - \theta_{12} e_{t-12} - \theta_1 \theta_{12} e_{t-13}. \end{aligned}$$

**Time Series Regression Model**

Continue with the hotel.dat. We use dummy variables to model the seasonal component as follow (via the transformed series $X$).

$$\begin{aligned} X_t &= a_0 + a_1 t + \beta_{s1} x_{s1,t} + \cdots + \beta_{s11} x_{s11,t} + e_t \\ &= a_0 + a_1 t + \beta_1 x_{1,t} + \cdots + \beta_{11} x_{11,t} + e_t \end{aligned}$$

where

$$x_{s1,t} = \begin{cases} 1 & \text{if time period is January,} \\ 0 & \text{otherwise;} \end{cases}$$

$$\vdots$$

$$x_{s11,t} = \begin{cases} 1 & \text{if time period is November,} \\ 0 & \text{otherwise.} \end{cases}$$

What about December?

Read Chapter 13, Examples 13.1–2.

Forecast future values at time $t+1$.

Suppose $t+1$ is January, then the predicted value for $Y_{t+1}$ is

$$\widehat{Y}_{t+1} = \widehat{a}_0 + \widehat{a}_1(t+1) + \widehat{\beta}_1 + \widehat{e}_{t+1},$$

where

$$\begin{aligned}
\widehat{\phi}(B)\widehat{\Phi}(B)\widehat{e}_{t+1} &= \eta_{t+1}, \\
\widehat{\phi}(B) &= 1 - \widehat{\phi}_1 B - \widehat{\phi}_3 B^3 - \widehat{\phi}_5 B^5, \\
\widehat{\Phi}(B) &= 1 - \widehat{\Phi}_1 B^{12}.
\end{aligned}$$

That is,

$$\begin{aligned}
\widehat{e}_{t+1} &= \widehat{\phi}_1 \widehat{e}_t + \widehat{\phi}_3 \widehat{e}_{t-2} + \widehat{\phi}_5 \widehat{e}_{t-4} + \widehat{\Phi}_1 \widehat{e}_{t-12} \\
&\quad + \widehat{\phi}_1 \widehat{\Phi}_1 \widehat{e}_{t-13} + \widehat{\phi}_3 \widehat{\Phi}_1 \widehat{e}_{t-15} + \widehat{\phi}_5 \widehat{\Phi}_1 \widehat{e}_{t-17},
\end{aligned}$$

with

$$\begin{aligned}
\widehat{e}_t &= Y_t - \widehat{a}_0 - \widehat{a}_1 t, \\
\widehat{e}_{t-2} &= Y_{t-2} - \widehat{a}_0 - \widehat{a}_1(t-2) - \widehat{\beta}_2, \\
&\cdots
\end{aligned}$$

**Seasonal intervention models**

Suppose that a season has 11 months.

$$\begin{aligned} Y_t - Y_{t-12} &= \beta_0 + \beta_{s1} x_{s1,t} + \cdots + \beta_{s11} x_{s11,t} + e_t \\ &= \beta_0 + \beta_1 x_{1,t} + \cdots + \beta_{11} x_{11,t} + e_t. \end{aligned}$$

If $t$ is January, then $Y_t = Y_{t-12} + \beta_0 + \beta_1$.
If $t$ is February, then $Y_t = Y_{t-12} + \beta_0 + \beta_2$.

$$\vdots$$

If $t$ is November, then $Y_t = Y_{t-12} + \beta_0 + \beta_{11}$.
If $t$ is December, then $Y_t = Y_{t-12} + \beta_0$.

Note that $Y_t - Y_{t-12}$ depends on the season. This allows us to model increasing seasonal variation.

We model the error term using ARIMA models.
Read Example 13.3. Do Tutorial 7.

Examples.

- hotel.sas: Box-Jenkins methodology (random trend and seasonal component)
- reg1.sas: deterministic trend and seasonal component. Model
  - trend by regression
  - seasonal variation by dummy variables
  - error term by ARIMA
- reg2.sas: model trend and seasonal component as above. Select the 'best' ARIMA model for the error term by `proc autoreg` via backward elimination.
- reg3.sas: model the trend by regression, the seasonal component by trigonometric function.

Attach

- hotel.tex,
- reg1.tex,
- reg2.tex and
- reg3.tex

**Transfer function models** (Ch 14)

- These are used to forecast a time series ($Y_t$) using another (explanatory) series ($X_t$).

- Model specification:

  $$Y_t = \mu + \beta_0 X_t + \beta_1 X_{t-1} + \cdots + \beta_k X_{t-k} + Z_t.$$

  - $Y =$ output, response, dependent series
  - $X =$ input, predictor, independent series
  - $X$ can have more than one variable
  - $Y$ leads $X$ (or $X$ lags $Y$)
  - No feedback yet (see `proc statespace`)

Examples:

- Advertising-sales
  ($Y$ = sales, $X$ = advertising)

- Blood pressure ($Y$ = SBP, $X$ = DBP)

- EEG data ($Y$ = brain, $X$ = wrist)

- Environmental data
  ($Y$ = death, $X_1$ = temp, $X_2$ = humid,... )

- Construction data ($Y$ = hstarts,
  $X_1$ = constrcts, $X_2$ = intrate)

- Gas-furnace data ($Y$ = Co2, $X$ = gas rate)

- Leading economic indicator and sales
  ($Y$ = sales, $X$ = econ-indicator)

Procedures for finding the model

1. Prewhitening of $X$ and $Y$ (Ch 9–13)

2. Estimate the crosscorrelation

3. Tentative model

4. Model the error term

5. Final model

## Prewhitening

The idea here is to transform $X$ to a white noise (hence the name whiten). This is carried out by fitting an ARIMA model to it. For example,

$$\phi(B)\nabla^d X_t = \theta(B)u_t,$$

$$u_t = \frac{\phi(B)}{\theta(B)}\nabla^d X_t.$$

Now transform $Y$ using the following

$$z_t = \frac{\phi(B)}{\theta(B)}\nabla^m Y_t.$$

```
proc arima data=sales;
 i var=x(1);
 e p=1 q=1 noconstant plot;
 i var=y(1) crosscor=(x(1)) nlag=10;
```

## Estimate the crosscorrelation

This is utilized for examining

$$\mathrm{corr}(u_t, z_{t+v}), \quad v = 0, 1, 2, \ldots$$

```
proc arima data=sales;
 ... ...
 i var=y(1) crosscor=(x(1)) nlag=10;
```

If there is a significant correlation at negative lags, see `proc statespace`.

## Tentative modeling

$$\nabla^m y_t = \mu + C\frac{w(B)}{\delta(B)}B^b\nabla^d x_t + \eta_t,$$

where

- $\nabla^m y_t$ and $\nabla^d x_t$ are the differenced series;
- $\mu$ is an unknown constant;
- $C$ is an unknown scale parameter;
- $b$ is the *delay* — the first significant spike occurs in the crosscorrelation plot;
- $w(B) = 1 - w_1 B - w_2 B^2 - \cdots - w_s B^s$ (p 666);
- $\delta(B) = 1 - \delta_1 B - \delta_2 B^2 - \cdots - \delta_r B^r$
  - set $r = 1$ for damped exponential,
  - set $r = 2$ for damped sine-wave fashion;
- $\eta_t$ is the error term.

## Model specification

```
Example:
e input = (2 $ (1,2) / (1,2) x)
noconstant printall
altparm maxit=30 backlim=-3  plot;  /* Output 4 */

INPUT = (<form-1>var-1 ...<,<form-n>var-n>)
where form-i describes the structure of the TF
for the corresponding variable. The
specification of the form is given by:
b $ (L11,L12, ...)(L21,...) ... / (Lj1,...) ...
where
b    is the delay
Lij  are the lag number for the parameters of TF.
     Terms before the slash (/) are numerator
     factors analog to MA parameters. Terms
     after / are denominator factors, similar to
     AR parameters. See the text (p666) for method
     of selecting b and the order of Lij.
```

Note. `method=ml` is a useful alternative.

## Handling the error term and final model

This is carried out using the standard tools (such as ACF, PACF, RSAC, RPSAC) for fitting an ARIMA model to the residuals:

$$\eta_t \sim \text{ARMA}(p, q).$$

The residual plots were produced in the previous estimation step thru the option `plot`.

The following example shows how to fit a model to the error term and then examine the residuals thru ACF and PACF.

```
/*
  The RSAC, RPSAC suggests an AR(2) model.
  Fit the final model
*/
e p=(1,2) input = (2 $ (1,2) / (1,2) x)
      noconstant printall
      altparm maxit=30 backlim=-3
      plot;  /* Output 5 */
```

**Putting things together ...**

The final model has the form:

$$\nabla^m y_t = \mu + C\frac{w(B)}{\delta(B)}B^b\nabla^d x_t + \frac{\theta(B)}{\phi(B)}e_t,$$

or,

$$\begin{aligned}\phi(B)\delta(B)\nabla^m y_t &= \phi(B)\delta(B)\mu \\ &\quad + C\phi(B)w(B)B^b\nabla^d x_t \\ &\quad + \delta(B)\theta(B)e_t,\end{aligned}$$

where

$$\begin{aligned}w(B) &= 1 - w_1 B - \cdots - w_s B^s, \\ \delta(B) &= 1 - \delta_1 B - \cdots - \delta_r B^r, \\ \theta(B) &= 1 - \theta_1 B - \cdots - \theta_q B^q, \\ \phi(B) &= 1 - \phi_1 B - \cdots - \phi_r B^r.\end{aligned}$$

This forms the basis for computing the forecast values.