Unit 6: Exploratory Data Analysis II

Taylor R. Brown PhD

Department of Statistics, University of Virginia

Spring 2020

Readings for Unit 6

Textbook chapter 2.2 (page 64).

Last Unit

1. Detrending
2. Differencing for Stationarity
3. Backshift Operator

This Unit

1. Periodic functions
2. Exploratory data tools to access frequency

Motivation

We've already seen how we can use differencing to obtain stationary processes. We are assuming that our observations can be written in the form

$$x_t = \mu_t + y_t \tag{1}$$

where μ_t is some function of time and y_t is a stationary process. What if μ_t were not a trend, but a periodic function?

1. Periodic Functions
2. Exploratory Data Tools to Access Frequency
3. Worked Example

Setup

A basic type of **periodic** function would be

$$\mu_t = Acos(2\pi\omega t + \phi),$$

where

- A: amplitude,
- ω: frequency,
- $1/\omega$: period,
- ϕ: phase.

Note that a cosine function is equal to a sine function for some phases, e.g. $cos(2\pi\omega t) = sin(2\pi\omega t + \frac{\pi}{2})$.

Setup

For now we assume that y_t in model (1) is white noise. Model (1) is now written as

$$x_t = A cos(2\pi\omega t + \phi) + w_t. \tag{2}$$

Setup

We could try to use non-linear least squares to fit A, ω and ϕ.
Recall the identity $cos(\alpha + \beta) = cos(\alpha)cos(\beta) - sin(\alpha)sin(\beta)$.
Thus, we can rewrite model (2) as

$$\begin{aligned} x_t &= \underline{\hspace{6cm}} \\ &= \underline{\hspace{6cm}} \\ &= \underline{\hspace{6cm}} \end{aligned} \tag{3}$$

Frequencies

In many settings, certain frequencies are natural. For example, in monthly data a frequency, ω, of $1/12$ (corresponding to a period of 12) is quite natural. We may want to remove a periodic signal by fitting

$$x_t = \beta_1 cos(2\pi/12 \ t) + \beta_2 sin(2\pi/12 \ t) + w_t$$

and then analyzing the residuals to understand w_t. This is regular regression by treating $cos(2\pi/12 \ t)$ and $sin(2\pi/12 \ t)$ as the **predictor variables** and we may use OLS to estimate β_1, β_2.

OLS Estimation

There are solutions for the estimates of these parameters.

$$\hat{\beta}_1 = \frac{2}{n}\sum_{t=1}^{n} x_t cos(2\pi\frac{1}{12}t).$$

$$\hat{\beta}_2 = \frac{2}{n}\sum_{t=1}^{n} x_t sin(2\pi\frac{1}{12}t).$$

Choosing Frequency

If we do not have an intuition regarding the frequency, we could try various regressions with different frequencies, ω, of the form $\frac{j}{n}$ for $j = 0, ..., \frac{n-1}{2}$. This guarantees evenly spaced frequencies from zero to 0.5. The parameters can be estimated by

$$\hat{\beta}_1(j/n) = \frac{2}{n}\sum_{t=1}^{n} x_t cos(2\pi tj/n).$$

$$\hat{\beta}_2(j/n) = \frac{2}{n}\sum_{t=1}^{n} x_t sin(2\pi tj/n).$$

Choosing Frequency

We then obtain the value of $\hat{\beta}_1^2(j/n) + \hat{\beta}_2^2(j/n)$ for all these frequencies, which can be interpreted as the amount of variation at a certain frequency. A measure of the presence of a frequency of oscillation of j cycles would be

$$P(j/n) = \hat{\beta}_1^2(j/n) + \hat{\beta}_2^2(j/n).$$

The quantity $P(j/n)$ is called the **periodogram**. Note that:

Exploratory Data Tools

- Periodogram (as described in the previous slide).
- ACF plot. Recall that the ACF is a correlation of lags; this makes sense in a stationary time series.
- Lag plot. Useful with **periodic** data, this is a scatter plot that has x_t on the y-axis and x_{t-h} on the x-axis.

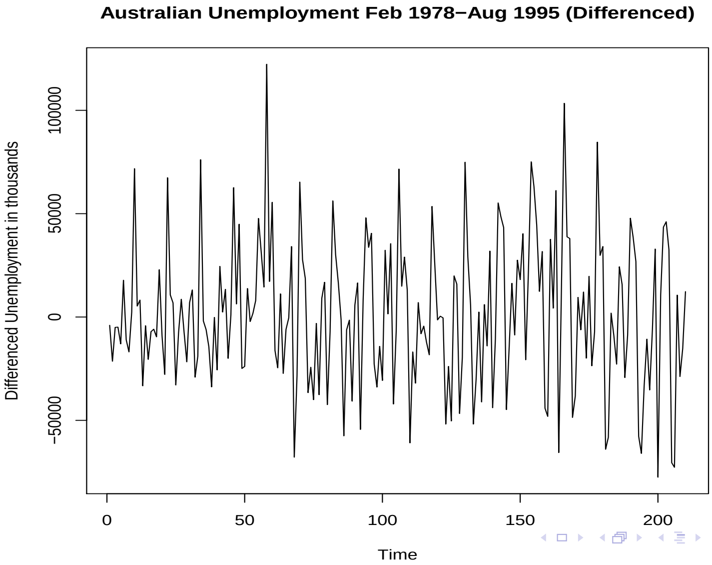
Example: Australian Unemployment

Let's look at these techniques with an example. The data consist of Australian unemployment numbers recorded monthly from Feb 1978 to Aug 1995 (in thousands).

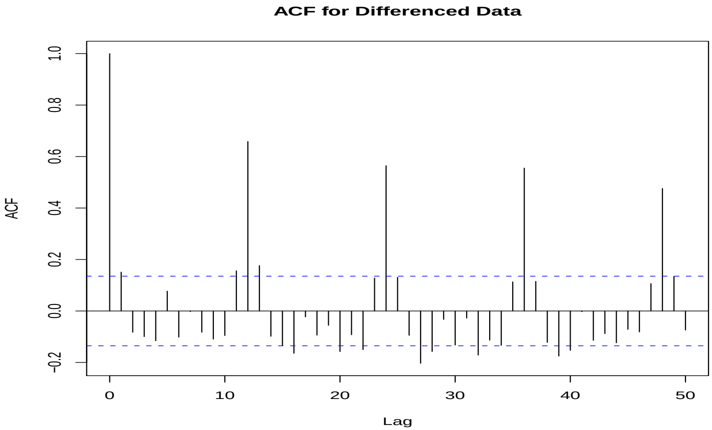
Example: Australian Unemployment

Question: Do the data look stationary?

Example: Australian Unemployment

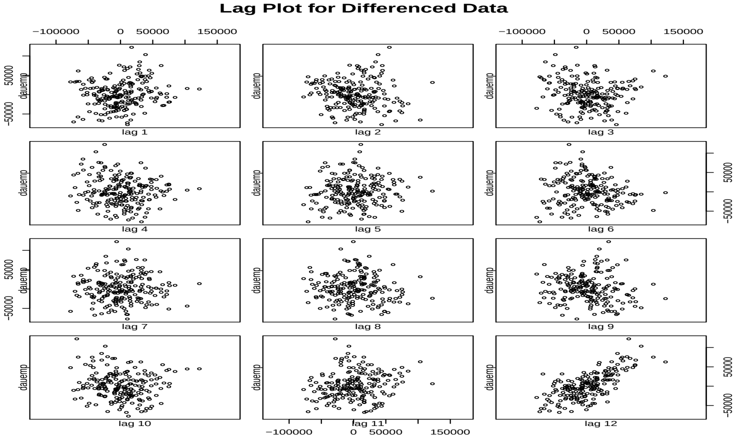


Example: Australian Unemployment



Question: What does the ACF indicate?

Example: Australian Unemployment



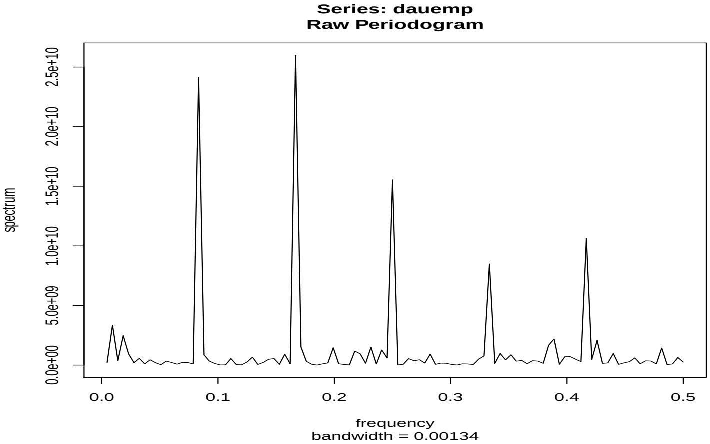
Question: What do the lag plots indicate?

Example: Australian Unemployment

```
auemp<-ts(scan("unemploy.dat", skip=1))
dauemp<-ts(diff(auemp))

plot(auemp)
plot(dauemp)
acf(dauemp,50)
lag.plot(dauemp, lags=12, diag=F)
```

Example: Australian Unemployment



Question: What does the periodogram indicate?

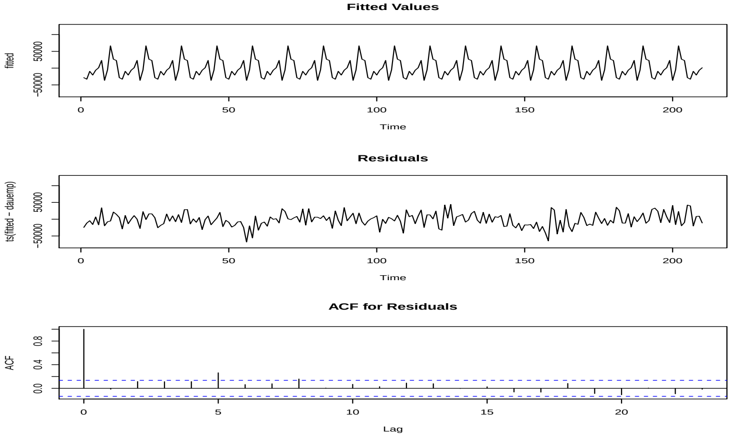
Example: Australian Unemployment

```
temp<-spec.pgram(dauemp, taper=0, log="no")

freq<-temp$freq[temp$spec>5e9]
freq
[1] 0.08333333 0.16666667 0.25000000 0.33333333 0.41666667
```

Example: Australian Unemployment

Let's perform regression at these 5 peaks.



Non-Stationary Residuals

Question: If the residuals exhibit non-stationarity, what does that suggest?

Example: Australian Unemployment

```
t=1:length(dauemp)
c1<-cos(2*pi*t*freq[1])
s1<-sin(2*pi*t*freq[1])
c2<-cos(2*pi*t*freq[2])
s2<-sin(2*pi*t*freq[2])
c3<-cos(2*pi*t*freq[3])
s3<-sin(2*pi*t*freq[3])
c4<-cos(2*pi*t*freq[4])
s4<-sin(2*pi*t*freq[4])
c5<-cos(2*pi*t*freq[5])
s5<-sin(2*pi*t*freq[5])
fit<-lm(dauemp~c1+s1+c2+s2+c3+s3+c4+s4+c5+s5)
```

Example: Australian Unemployment

```
fitted<-ts(fit$coef[2]*c1+fit$coef[3]*s1+fit$coef[4]
*c2+fit$coef[5]*s2+fit$coef[6]*c3+fit$coef[7]
*s3+fit$coef[8]*c4+fit$coef[9]*s4+fit$coef[10]
*c5+fit$coef[11]*s5)

par(mfrow=c(3,1))
plot(fitted, ylim=c(miny,maxy), main="Fitted Values")
plot(ts(fitted-dauemp), ylim=c(miny,maxy), main="Residuals'
acf(ts(fitted-dauemp), main="ACF for Residuals")
```

Recap

To recap, we consider model (1) which take the form

$$x_t = \mu_t + y_t$$

where μ_t is a either a polynomial or periodic function, and y_t is a zero mean stationary process.

Recap

If μ_t is polynomial, we can

- Difference to coerce stationarity.
- Use least squares regression, if estimating y_t is the goal.

If μ_t is periodic, we can use the periodogram to identify prevalant frequencies.