UNIVERSITY of LIMERICK

OLLSCOIL LUIMNIGH

Time Series Analysis

MS 4218

joseph.lynch@ul.ie

Outline

Main features of time series:

- Trend
- Seasonality
- Correlation between successive observations

Descriptive techniques

Usually in statistics, when given some data to analyse, we first calculate the summary statistics.

Entering

```
summary(dataname)
```

in R outputs min., max., quartiles, ± mean and standard deviation.

With time series, because of auto-correlation, these outputs can be seriously misleading.

Understanding time series

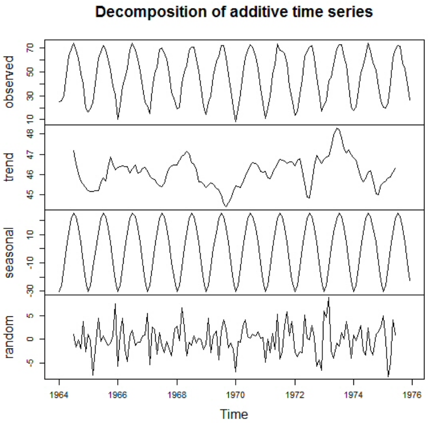
The variation in observed time series is decomposed into trend, seasonal and cyclic variation.

After these have been accounted for, any residual variation is classed as irregular variation.

```
data(tempdub)
Decomptempdub<-decompose(tempdub)
plot(Decomptempdub)
names(Decomptempdub)
Decomptempdub$trend
```

If there is no seasonal component, e.g., in data(larain), decompose() will not work.

Decomposed time series data



Trend estimation

Trend estimates represent the underlying directions of the time series: the long term change in the mean level.

It is important to take into account length of observation time. How long is long term?

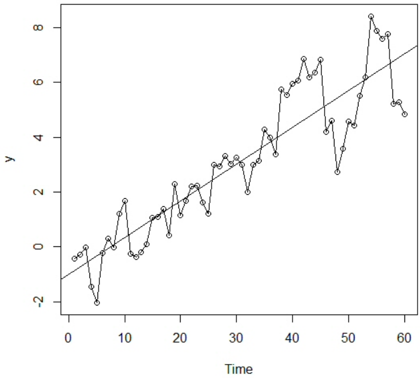
Useful if irregular component is large.

Can be removed so as data can be examined for other sources of variation.

Deterministic v stochastic trends

The apparent trend in the random walk is not deterministic; it is a function of the auto-correlation of the data.

```
data(rwalk);plot(rwalk,type=’o’,ylab=’y’)
modelRW<-lm(rwalk~time(rwalk))
abline(modelRW)
```



Analysing series that contain a trend

Is the trend constant?, i.e., time-independent: $\mu_t = \mu$

Is it a linear function of time?: $\mu_t = \beta_0 + \beta_1 t$

Lines go on forever, so is the trend global or local?

Is it seasonal(regular) or is it cyclical(not regular)?

Estimation of a constant mean

Let $Y_t = \mu + X_t$ be stationary, where $E(X_t) = 0 \; \forall \; t$.

The sample mean $\bar{Y} = \frac{1}{n} \sum_{t=1}^{n} Y_t$ is a sample statistic.

From the Central Limit Theorem, $\bar{Y} \sim N\left(\mu, \frac{\sigma^2}{n} = \frac{\gamma_0}{n}\right)$.

If the data are correlated, its variance is a function of the correlation within the data.

$$Var(\bar{Y}) \;=\; \frac{\gamma_0}{n} \left\{ 1 + 2 \sum_{k=1}^{n-1} \left(1 - \frac{|k|}{n} \right) \rho_k \right\}.$$

Example of variance of constant mean estimator

Let Y_t be a stationary process such that:

$$
\begin{aligned}
Y_t &= e_t - \frac{1}{2} e_{t-1}; \qquad e' \text{s are white noise.} \\
Var(Y_t) &= \gamma_0 \\
&= Var\left(e_t - \frac{1}{2} e_{t-1}\right) \\
&= \sigma_e^2 + \left(-\frac{1}{2}\right)^2 \sigma_e^2 \\
&= \frac{5}{4}\sigma_e^2.
\end{aligned}
$$

Constant mean estimator properties cont.

$$
\begin{aligned}
\gamma_1 &= Cov(Y_t, Y_{t-1}) \\
&= E(Y_t Y_{t-1}) - E(Y_t)E(Y_{t-1}) \\
&= E\left\{\left(e_t - \frac{1}{2}e_{t-1}\right)\left(e_{t-1} - \frac{1}{2}e_{t-2}\right)\right\} - (0)(0) \\
&= E\left(-\frac{1}{2}e_{t-1}^2\right) \\
&= -\frac{1}{2}\sigma_e^2.
\end{aligned}
$$

Constant mean estimator properties cont.

$$\begin{aligned}
\rho_1 &= \frac{\gamma_1}{\gamma_0} \\
&= -\frac{1}{2}\sigma_e^2 \div \frac{5}{4}\sigma_e^2 \\
&= -0.4.
\end{aligned}$$

Constant mean estimator properties cont.

$$
\begin{aligned}
Var(\bar{Y}) &= \frac{\gamma_0}{n}\left\{1 + 2\sum_{k=1}^{n-1}\left(1 - \frac{|k|}{n}\right)\rho_k\right\} \\
&= \frac{\gamma_0}{n}\left\{1 - 0.8\left(\frac{n-1}{n}\right)\right\} \quad (k = 1) \\
&\approx 0.2\frac{\gamma_0}{n} \text{ when } n \text{ is large.}
\end{aligned}
$$

Estimation of a constant mean cont.

In white noise $\rho_k = 0$ and thus $Var(\bar{Y}) = \frac{\gamma_0}{n}$.

If $\rho_k < 0$, $Var(\bar{Y})$ decreases as we have just seen.

If $\rho_k > 0$, $Var(\bar{Y})$ increases.

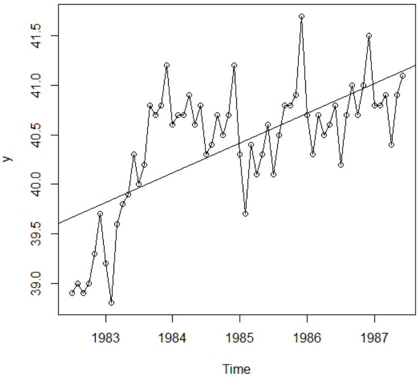
If process is non-stationary, $Var(\bar{Y})$ can be even bigger.

Regression methods

If there is a deterministic mean trend e.g., $\mu_t = \beta_0 + \beta_1 t$, we need to estimate $\hat{\beta}_0$ and $\hat{\beta}_1$.

```
data(hours)
modelhours<-lm(hours~time(hours))
plot(hours,type="o",ylab="y",)
abline(modelhours)
```

Hours data with linear time trend



Linear coefficients

```
summary(modelhours)

#              Estimate Std. Error t value Pr(>|t|)
#(Intercept) -556.31190   86.31998  -6.445 2.49e-08 ***
#time(hours)    0.30062    0.04349   6.913 4.11e-09 ***

cor(hours[-1],hours[-length(hours)])
#[1] 0.7896543
```

Interpretation of R output

We see that the data are positively correlated.

Both $\hat{\beta}_0$ and $\hat{\beta}_1$ parameters are significant.

For any correlation, positive or negative, the standard errors of these parameter estimates:

$se(\hat{\beta}_0) = 86.319$ and
$se(\hat{\beta}_1) = 0.043$

in the regression output are likely to be underestimated.

Regression methods cont.

Instead of ordinary least squares regression, we use generalised least squares to estimate the parameters.

```
library(nlme)
modelhours2=gls(hours~time(hours),cor=corAR1(0.7896543))
summary(modelhours2)
#                Value Std.Error   t-value p-value
#(Intercept) -613.3437 196.67083 -3.118631  0.0028
#time(hours)    0.3293   0.09908  3.323909  0.0015
```

The respective $(\hat{\beta}_0, \hat{\beta}_1)$ parameter values have gone from (-556.312 and 0.301) to (-613.344 and 0.329).

The respective standard errors (86.319 and 0.043) have increased to (196.671 and 0.099) resulting in a drop in the $|t|$ values.

Seasonal variation

Commonly found in sales figures, T° records, unemployment statistics etc.

These have a regular annual period.

The seasonal component can be estimated if it is of direct interest.

It can be removed (seasonally-adjusted) if it is not of direct interest. This enables in-year comparisons.

If irregular component is large, these seasonally-adjusted estimates may be misleading.

Seasonal variation cont.

- Additive: $Y_t = T(t) + S(t) + \epsilon(t)$

  With additive seasonality, seasonal fluctuations exhibit constant amplitude with respect to the trend.

  This is the default.

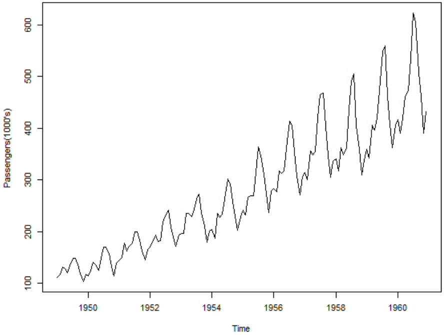
- Multiplicative: $Y_t = T(t) \times S(t) \times \epsilon(t)$.

  With multiplicative seasonality, seasonal fluctuations is a function of the trend, e.g., the amplitude of the seasonal fluctuations increases as the trend increases.

Multiplicative seasonal variation

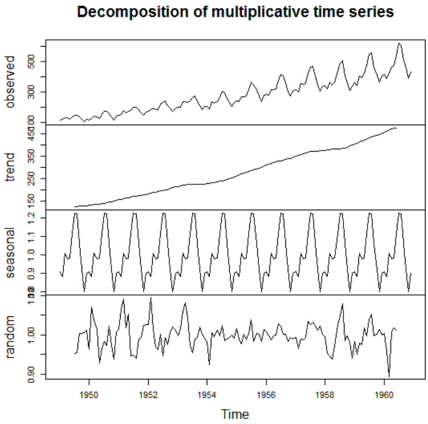
```
data(AirPassengers)
plot(AirPassengers, ylab="Passengers(1000’s)")
```

Pan-Am air passenger time plot



Multiplicative decomposition

```
plot(decompose(AirPassengers,
type="multiplicative"))
```



Linear models with seasonal variables

Let $Y_t = \mu_t + X_t$ with again, $E(X_t) = 0 \; \forall \; t$.

In the Dubuque T°, μ_t has 12 $\hat{\beta}$'s requiring estimation, i.e., the average for each separate month throughout the data.

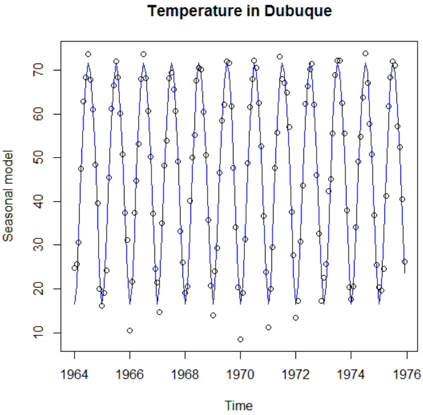
No intercept is used via (-1) in the R code:

```
data(tempdub)
month.=season(tempdub)
modelSM<-lm(tempdub~month.-1)
```

If an intercept term is included, the first month becomes a reference month for the other 11 months. The β's are thus changed accordingly.

```
modelSM2=lm(tempdub~month.)
summary(modelSM2)
```

```
plot(ts(fitted(modelSM),freq=12,start=c(1964,1)),
main="Temperature in Dubuque",ylab="Seasonal model",
col=4,ylim=range(c(fitted(modelSM),tempdub)))
points(tempdub)
```



Cosine trends

In a Cosine model, only three parameters need estimating.

Let $f = 1/12$ be the frequency (period$= 1/f =$12), Φ the phase and β the amplitude.

$$\mu_t = \beta cos(2\pi ft + \Phi)$$

$$cos(A + B) = cosAcosB - sinAsinB \Rightarrow$$

$$\mu_t = \beta_0 + \beta_1 cos(2\pi ft) + \beta_2 sin(2\pi ft),$$

where $\beta_1 = \beta cos(\Phi)$, $\beta_2 = -\beta sin(\Phi)$, $\Phi = tan^{-1}(-\beta_2/\beta_1)$ and β_0 is a cosine with zero frequency.

Cosine trends cont.

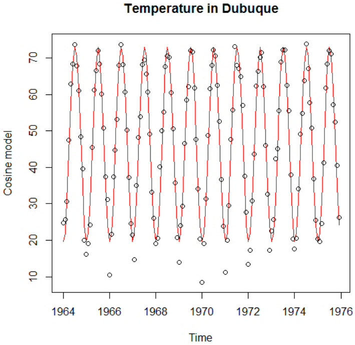
$cos(2\pi ft)$ and $sin(2\pi ft)$ are explanatory variables and their parameter estimates, $\hat{\beta}_1$ and $\hat{\beta}_2$ are seen to be highly significant when we run the following code:

```
modelCos<-lm(tempdub~harmonic(tempdub,m=1))
summary(modelCos)
```

We can plot it via:

```
plot(ts(fitted(modelCos),freq=12,start=c(1964,1)),
main="Temperature in Dubuque",ylab="Cosine model",
col=2,ylim=range(c(fitted(modelCos),tempdub)))
points(tempdub)
```

Cosine trend for temperature series



Seasonal means v cosine model

Cosine model has far fewer parameters to be estimated.

As the seasonal $\hat{\beta}'$s are monthly averages, their variances are a function of ρ.

$$Var(\hat{\beta}_j) = \frac{\gamma_0}{N}\left\{1 + 2\sum_{k=1}^{N-1}\left(1 - \frac{k}{N}\right)\rho_k\right\} \quad j = 1, \ldots, 12.$$

It turns out that, even if the data are correlated, the standard errors associated with the $\hat{\beta}'$s are much smaller for the parsimonious cosine trend model than for the seasonal means model.

Cyclical variation

Variation can be non-seasonal but fixed, e.g., daily T°.

Periodic variation can also be not fixed, e.g., business cycles might be somewhere between 3 and 10 years.

See also hare data in lecture 1.

Despite not having regular periods, they may be predictable to a certain extent.

Other possible effects

- Holiday effects; may be moving
- Trading day effects; Monday or Friday may be different from other days
- Outliers, e.g., “one-off" events that have large effects
- Level shifts, also called discontinuities, where the measurement (level) suddenly takes a step change and does not revert to its previous value
- Seasonal breaks, where the seasonal pattern changes permanently

Irregular variation: Residual analysis

After the trend and seasonal or cyclical variation have been removed, what remains is the residual variation.

This might be random, or more likely in a time series, the variation may be explained in terms of probability models such as moving average models or autoregressive models, of which, more later.

$$\hat{X}_t \quad = \quad Y_t - \hat{\mu}_t.$$

The predicted unobserved residual is the difference between what is observed and the estimated trend (constant, linear, quadratic(see lab3), seasonal, cosine).

Residuals

The usual assumption is that they follow $N(0, \sigma_e^2)$.

When standardised by subtraction of the residual mean and dividing the result by the residual standard deviation, they follow a standard Normal distribution i.e., $N(\mu = 0, \sigma^2 = 1)$.

This being the case, 95% of the standardised residuals should lie within ± 2.

Studentised residuals are similar but make allowance for outliers.

Residual analysis cont.

If the standardised or the studentised residuals are:

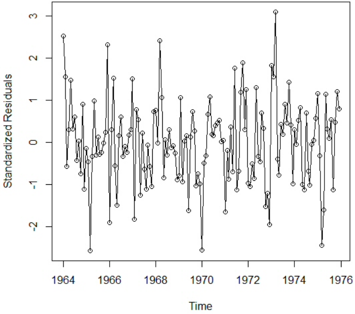
- not normally distributed, i.e., do not display random variation
- do not have zero mean
- do not have constant variance

then the chosen models may have to be altered.

Temperature seasonal means model

```
plot(y=rstudent(modelSM),x=as.vector(time(tempdub)),
xlab='Time',ylab='Standardized Residuals',type='o')
```

Residuals v time

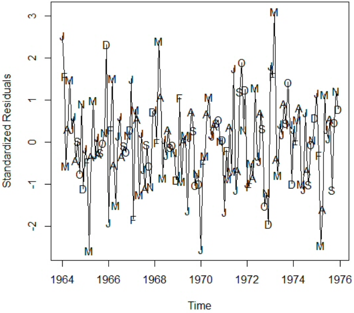


The residuals are randomly distributed.

Residuals v time with seasonal plotting symbols

Supplement the last graph with the following code:

```
points(y=rstudent(modelSM),x=as.vector(time(tempdub)),
pch=as.vector(season(tempdub)))
mean(rstudent(modelSM))
```

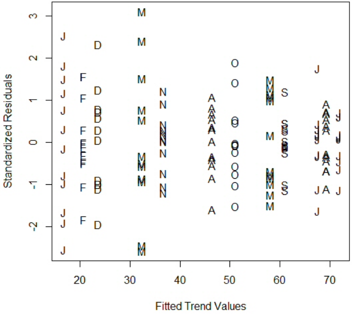


The monthly patterns are random with mean close to zero.

Standardised residuals v fitted $\hat{\mu}_t$ values

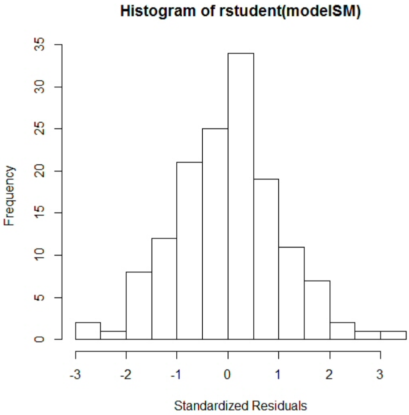
```
plot(y=rstudent(modelSM),x=as.vector(fitted(modelSM)),
xlab='Fitted Trend Values',ylab='Standardized Residuals',type='n')
points(y=rstudent(modelSM),x=as.vector(fitted(modelSM)),
pch=as.vector(season(tempdub)))
```

More variation in January and March, but overall variation relatively constant.



Histogram of standardised residuals from seasonal means model

```
hist(rstudent(modelSM),xlab='Standardized Residuals')
```

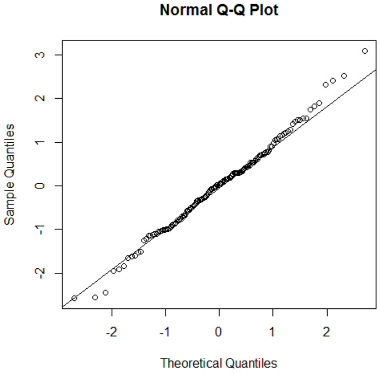


The residuals appear normally distributed.

QQ plot: Data quantiles v theoretical normal quantiles

```
qqnorm(rstudent(modelSM)); qqline(rstudent(modelSM))
```

Standardised residuals of temperature seasonal means model



Residuals are normally distributed.

Shapiro-Wilk test

The Null Hypothesis is the residuals are normally distributed.

Reject H_0 if $p < 0.05$.

```
shapiro.test(rstudent(modelSM))
```

The p value $= 0.6954$ is high and hence we cannot reject H_0.

Runs test

A test for checking any non-random patterns in data.

H_0 is the pattern is random.

Reject if $p < 0.05$.

```
runs(rstudent(modelSM))
```

The p value $= 0.216 > 0.05$ and so we cannot reject H_0.

Sample auto-correlation(acf) function (covered in detail later)

A stationary process has constant mean and variance.

For any sample of data items in a stationary process, the correlation between two observations, *k* lags apart is given by:

$$r_k = \frac{\sum_{t=1}^{n-k}(Y_t - \bar{Y})(Y_{t-k} - \bar{Y})}{\sum_{t=1}^{n}(Y_t - \bar{Y})^2}, \text{ for } k = 1, 2\ldots$$

Acf cont.

$$r_k \sim N\left(-\frac{1}{n} \approx 0, \frac{1}{\sqrt{n}}\right).$$

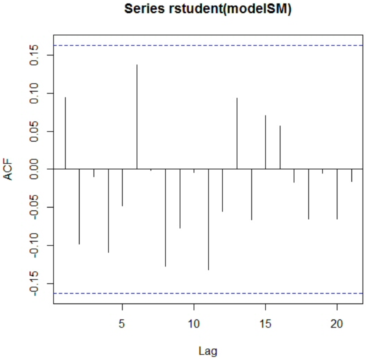
95% of r_k values should lie between the dotted lines (shown overleaf) of ± 2 standard errors $= \pm\frac{2}{\sqrt{n}}$.

In the case of Dubuque temperature data,
$\frac{2}{\sqrt{n}} = \pm\frac{2}{\sqrt{144}} = \pm 0.16$,

where $n =$ number of data points in tempdub.

Sample acf of Residuals seasonal means model without intercept

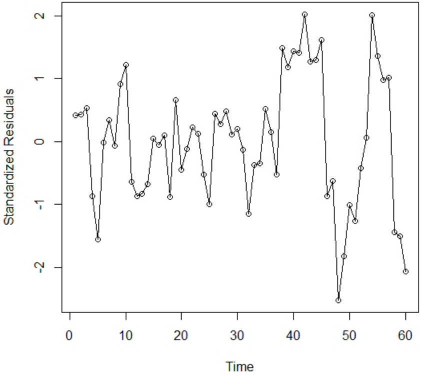
```
acf(rstudent(modelSM))
```



$H_0 : \rho_k = 0$ holds each time.

Residuals from straight line fit of random walk

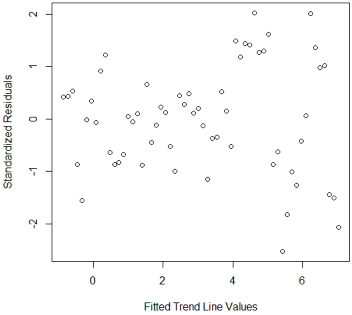
```
plot(y=rstudent(modelRW),x=as.vector(time(rwalk)),
ylab='Standardized Residuals',xlab='Time',type='o')
```



Residuals hang together here too much.

Residuals v fitted values from straight line fit of random walk

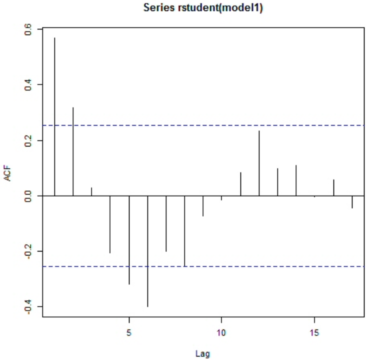
```
plot(y=rstudent(modelRW),x=fitted(modelRW),
ylab='Standardized Residuals',
xlab='Fitted Trend Line Values',type='p')
```



Funnel effect of increasing variance for larger $\hat{\mu}_t$ values.

Sample acf of residuals from straight line model of random walk

```
acf(rstudent(modelRW))
```

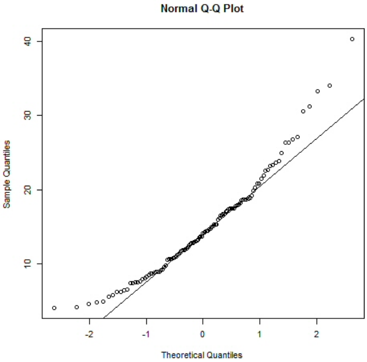


Lags 1, 2, 5 and 6 are beyond the $\pm\frac{2}{\sqrt{n}} = \pm\frac{2}{\sqrt{60}} = 0.26$.

QQ plot of Los Angeles rainfall series in lecture 1

```
data(larain)
qqnorm(larain); qqline(larain)
```

Data appeared to be random, but was the distribution normal?



Not normally distributed in either tail.

Normality test for LA rainfall series

```
shapiro.test(larain)
```

yields a p value of $0.0001614 < 0.05$

Reject H_0.

Data are not normally distributed.

Next two lectures

Models for stationary time series

- General linear processess
- Moving Average (MA) processes
- Auto-Regressive (AR) processes
- Mixed ARMA processes