Example 1: Yearly average global temperatures deviations, 1900–1997

Figure 1: Different variants of a time series plot
Example 2: US population every ten years, 1790–1990

Figure 2: The most significant feature is an increasing trend

Figure 3: The most significant feature is a quasi-periodic behavior
Example 4: Quarterly earning per share of the company Johnson & Johnson, 1960–1980

Figure 4: The most significant features are an increasing trend and a quasi-periodic behavior. In addition, note that the variability of the data tend to increase with the mean.
Some common goals of time series analysis

- Trend estimation (Examples 1 and 2)
- Seasonally adjusted trend estimation (Example 4)
- Signal detection. For instance, estimate a “signal” from observations of it corrupted by noise. Suppose we have observations \( \{x_t; t = 1, \ldots, n\} \) assumed to have been generated as \( x_t = \sin\left(\frac{2\pi t}{P}\right) + \epsilon_t \), where \( \epsilon_t \) is “noise” and \( P \) is an unknown period. The goal is to estimate the signal \( \sin\left(\frac{2\pi t}{P}\right) \).
- Prediction/Forecasting (Examples 2 and 4)
- Quantifying the type and strength of a relation between two time series. For instance, determine the relation between number of cardiovascular deaths in a city and the levels of certain pollutant in that city
- Many others ...

Time series

The starting point for the statistical analysis of time series data is to propose a probability model that serves to describe the variability in the data. A time series is a collection of random variables indexed by time, \( \{X_t : t \in T\} \), where \( T \subset \mathbb{R} \) is the index set. In this course we will assume that \( T = \mathbb{Z} = 0, \pm 1, \pm 2, \ldots \) The observed data consist of a realization of the above random variables for times \( t \in T_{obs} \subset T \) (a subset of \( T \)), that is, data = \( \{x_t : t \in T_{obs}\} \).

For most time series data, the original observation times correspond to days, months, quarters, years, etc. But for many data analyzes it is more convenient to re-scale (measure) time so the observation times become \( T_{obs} = \{1, 2, 3, \ldots, n\} \), where \( n \) is the number of observations. Consider the data in Example 3, the monthly number of accidental deaths in the US for the period 1973–1978. We will have:
Analysis of time series using linear regression

Example 5: Analysis of US population data using linear regression

[vdo@field Brock-Dav.dat]$ more USPOP.TSM
3929214
5308483
7239881
9638453
12860702
17063353
23191876
.
.
.

[vdo@field ST5383]$ Splus
S-PLUS : Copyright (c) 1988, 2002 Insightful Corp.
S : Copyright Lucent Technologies, Inc.
Version 6.1.2 Release 2 for Linux 2.2.12 : 2002
Working data will be in .Data

## Read the time series data into a vector
> uspop.tsm <- scan("/home/vdo/Work/ST5383/Datasets/Brock-Dav.dat/USPOP.TSM")

## Transform the vector into a time series object
> uspop.ts <- rts(uspop.tsm,start=1790,deltat=10)

## Plot the time series
> tsplot(uspop.ts,type="p",xlab="year",ylab="population in the US") # Fig. 2

## Fitting a linear regression model
## For the purpose of fitting a linear regression, it is better to code time
## as t=1,2,...,21 instead of the original t=1790,1800,...,1990.
## The two are related by t_org = 1790 + 10*(t_new - 1)
> t <- 1:21
> uspop.lsfit <- lm(uspop.ts ~ 1 + t + t^2)
> coefficients(uspop.lsfit)
  (Intercept)      t  I(t^2)
6957920   -2159870 650633.9

## The prediction equation using this model is
## \( \hat{X} = 6957920 - 2159870t + 650633.9t^2 \), so in particular the predicted
## US population in 2010 (t=23) using this model is 301463470 people.
## Data and fitted model

```r
> tsplot(uspop.ts,type="p",xlab="year",ylab="population in the US",
+ main="US population and fitted model")
> lines(1790 + 10*(t-1),fitted(uspop.lsf)) # Fig. 5
```

## The model provides an apparent good fit since R^2 = 0.9989, but ...

## Looking at residuals

```r
> plot(1790 + 10*(t-1),residuals(uspop.lsf),xlab="year",ylab="residuals",
+ main="Residuals of fitting linear model to US population") # Fig. 6
> abline(0,0)
```

## The residuals do not behave as "white noise", and the model can be improved
## by using correlated errors. The model with correlated errors might not improve
## much the fit to the observed data, but will likely improve the predictions
Example 6: Analysis of J & J quarterly earning data using linear regression

[vdo@field ST5383]$ Splus

## Read the data
> jj.dat <- scan("/home/vdo/Work/ST5383/Datasets/Shum-Stoff.dat/jj.dat")

## Transform the vector into a time series object
> jj.da <- rts(jj.dat,start=c(1960,1),end=c(1980,4),frequency=4)

## Plot the time series
> tsplot(jj.da,type="p",xlab="year",ylab="J & J quarterly earnings") # Fig. 4

## Fitting a linear regression model
## The series presents both trend and seasonal behavior. A way to model the
## latter is to use as explanatory variables quarter indicators
## Note also that the variability of the series increases with the level of
## the series, so a log transformation reduces this problem

> par(mfrow=c(1,2))
> tsplot(jj.da,type="p",xlab="year",ylab="J & J quarterly earnings") # Fig. 7
> tsplot(log(jj.da),type="p",xlab="year",ylab="log of J & J quarterly earnings")

## Create matrix with quarter indicators as dummy explanatory variables:
## Q_1(t) = 1 if time t is a first quarter and 0 otherwise,
## Q_2(t) = 1 if time t is a second quarter and 0 otherwise, etc
> z <- t(matrix( contr.sum(4,contrast=F), 4, 84))

> z

```r
[1,]  1  0  0  0
[2,]  0  1  0  0
[3,]  0  0  1  0
[4,]  0  0  0  1
[5,]  1  0  0  0
[6,]  0  1  0  0
[7,]  0  0  1  0
[8,]  0  0  0  1
[9,]  1  0  0  0
[10,]  0  1  0  0
[11,]  0  0  1  0
[12,]  0  0  0  1
```

...
## Fitting a linear regression model

For computational convenience code time as \( t=1,2,\ldots,84 \)
\[
\begin{align*}
  \text{tt} & \leftarrow 1: \text{length}(\text{jj} \cdot \text{da}) \\
  \text{logjj.lsfit} & \leftarrow \text{lm}(\log(\text{jj} \cdot \text{da}) \sim -1 + \text{tt} + z) \\
  \text{coefficients}(\text{logjj.lsfit})
  \begin{array}{ccccc}
  \text{tt} & z1 & z2 & z3 & z4 \\
  0.04179304 & -0.6607215 & -0.6325988 & -0.5624905 & -0.8312482
  \end{array}
\end{align*}
\]

The prediction equation using this model is
\[
\hat{X\cdot t} = 0.04179304 \cdot t - 0.6607215 \cdot Q_1(t) - 0.6325988 \cdot Q_2(t)
- 0.5624905 \cdot Q_3(t) - 0.8312482 \cdot Q_4(t).
\]

Hence, the predicted log-earnings for the third quarter of 1983
\((t=95)\) using this model is
\[
0.04179304 \cdot 95 - 0.5624905 = 3.40784, \quad \text{and}
\]
the predicted quarterly earning is \(\exp(3.407848) = 30.20018\).

Data and fitted model
\[
\begin{align*}
  \text{tsplot}(\log(\text{jj} \cdot \text{da}), \text{type}="p", \text{xlab}="\text{year}", \text{ylab}="\text{log of J \\& J quarterly earnings}", \\
  \quad + \text{main}="\text{J \\& J log-quarterly earnings and fitted linear model}\") \\
  \text{lines}(1960 + 0.25 \cdot (\text{tt}-1), \text{fitted}(\text{logjj.lsfit})) \quad \# \text{Fig. 8}
\end{align*}
\]

The model provides an apparent good fit since \(R^2 = 0.9935\), but ...

Looking at residuals
\[
\begin{align*}
  \text{plot}(1960 + 0.25 \cdot (\text{tt}-1), \text{residuals}(\text{logjj.lsfit}), \text{xlab}="\text{year}", \text{ylab}="\text{residuals}", \\
  \quad + \text{main}="\text{Residuals of fitting linear model to J \\& J log-quarterly earning}\") \quad \# \text{Fig. 9}
  \text{abline}(0,0)
\end{align*}
\]

The residuals do not behave as "white noise", since they have a periodic
pattern, and the model can be improved by using correlated errors.

The model with correlated errors might not improve much the fit to the
observed data, but will likely improve the predictions
US population and fitted model

Figure 5:

Residuals of fitting linear model to US population

Figure 6:
Figure 7:

J & J log-quarterly earnings and fitted linear model

Figure 8:
Residuals of fitting linear model to J & J log-quarterly earning

Figure 9: