Times Series and Forecasting (I)

Chapter 1. Introduction and Examples

Jianhua Hu

School of Statistics and Management
Shanghai University of Finance and Economics

Spring 2013

Example 1.1. Earth temperature data

- "Global warming" refers to the increase in the average temperature of the Earth's near-surface air and oceans since the mid-20th century and its projected continuation.
- The question of interest for global warming proponents and opponents is whether the overall trend is natural or whether it is caused by some human-induced interface (human being greedy).
- The data of interest are annual temperature deviations (1856-1997) in degrees C. The data are a combination of land-air average temperature "anomalies", measured from a 1961-1990 baseline average.

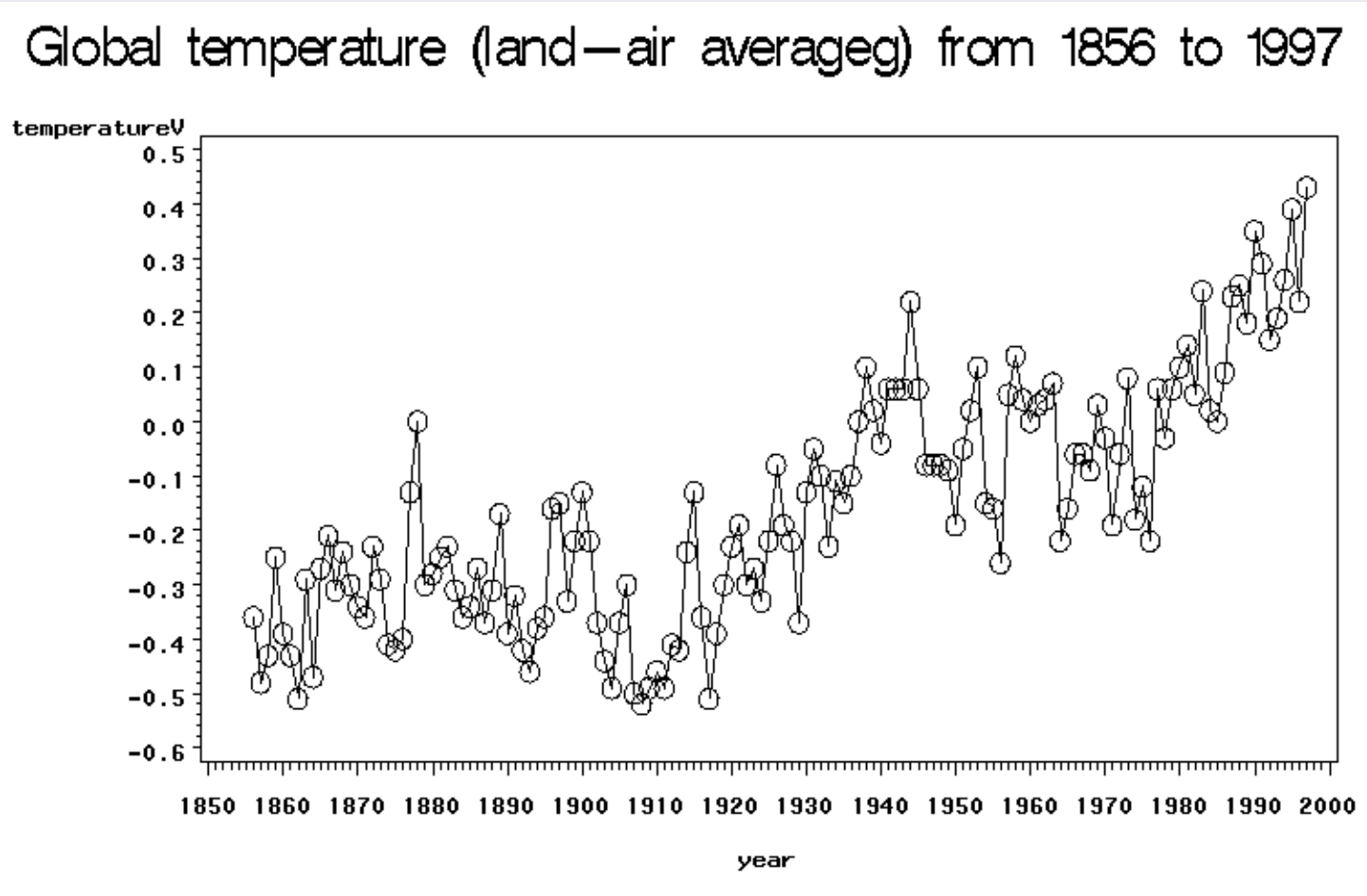
Data of example 1.1 and SAS program

```
Data chapter1.example11;
input temperatureV @@;
year = intnx( 'year', '1jan1856'd, _N_ - 1 );
format year year4.; datalines;
-.36 -.48 -.43 -.25 -.39 -.43 -.51 -.29 -.47 -.27 -.21 -.31 -.24 -.30 -.34 -.36
-.23 -.29 -.41 -.42 -.40 -.13 .00 -.30 -.28 -.25 -.23 -.31 -.36 -.34 -.27 -.37
-.31 -.17 -.39 -.32 -.42 -.46 -.38 -.36 -.16 -.15 -.33 -.22 -.13 -.22 -.37 -.44
-.49 -.37 -.30 -.50 -.52 -.49 -.46 -.49 -.41 -.42 -.24 -.13 -.36 -.51 -.39 -.30
-.23 -.19 -.30 -.27 -.33 -.22 -.08 -.19 -.22 -.37 -.13 -.05 -.10 -.23 -.11 -.15
-.10 -.00 .10 .02 -.04 .06 .06 .06 .22 .06 -.08 -.08 -.08 -.09 -.19 -.05 .02
.10 -.15 -.16 -.26 .05 .12 .04 .00 .03 .04 .07 -.22 -.16 -.06 -.06 -.09 .03
-.03 -.19 -.06 .08 -.18 -.12 -.22 .06 -.03 .06 .10 .14 .05 .24 .02 .00 .09 .23
.25 .18 .35 .29 .15 .19 .26 .39 .22 .43
;
run;
```

SAS programming for Figure 1.1

```
Proc gplot data= chapter1.example11;
symbol i=spline v=circle h=1.5;
plot temperatureV * year;
title 'Global temperature (land-air average) from 1856 to 1997';
run;
```

Figure 1.1


Global temperature (land−air averageg) from 1856 to 1997
temperatureV
0.5
0.4
0.3
0.2
0.1
0.0
-0.1
-0.2
-0.3
-0.4
-0.5
-0.6
1850 1860 1870 1880 1890 1900 1910 1920 1930 1940 1950 1960 1970 1980 1990 2000
year

Comments of Example 1.1

- There are $n = 142$ observations. Measurements are taken **each year**.
- We note an apparent upward trend in the series during the latter part of the twentieth century that has been used as an argument for the global warming hypothesis.
- Note also the leveling off at about 1945 and then another rather sharp upward trend at about 1970.
- What are the noticeable patterns? Predictions?

How to model the data in Example 1.1

Consider the following model to fit the temperature data

- Let Y be the random variable of global temperature deviation, then

$$Y = \mathsf{E}(Y) + e.$$

- If we consider the times t and use a linear regression model to fit the data, then we observe the following model

$$Y_i = \beta_0 + \beta_1 t_i + e_i, i = 1, \cdots, 142,$$

equivalently, the sample (matrix) version

$$\mathbf{Y}_{n\times 1} = X\boldsymbol{\beta} + \mathcal{E}, \quad \mathcal{E} = (e_1, \cdots, e_n)'$$

Is the regression model reasonable for Example 1.1?

- Assumption 1: $\text{Cov}(\mathcal{E}) = \sigma_e^2 I_n$.
  By Gauss-Markov Theorem, $\hat{\boldsymbol{\beta}} = (X'X)^{-1}X'Y$ is BLUE of the regression coefficient $\boldsymbol{\beta}$.
  Assumption 1 is reasonable?

- Assumption 2: $\text{Cov}(\mathcal{E}) = \sigma_e^2 V$ for known $V > \mathbf{0}$.
  By Gauss-Markov Theorem, $\hat{\boldsymbol{\beta}} = (X'V^{-1}X)^{-1}X'V^{-1}Y$ is BLUE of the regression coefficient $\boldsymbol{\beta}$.
  How do you get the known V? For example,

$$V = \frac{1}{1-\rho^2}\begin{pmatrix} 1 & \rho & \rho^2 & \cdots & \rho^{n-1} \\ \rho & 1 & \rho & \cdots & \rho^{n-2} \\ \cdots & \cdots & \cdots & \cdots & \cdots \\ \rho^{n-1} & \rho^{n-2} & \rho^{n-3} & \cdots & 1 \end{pmatrix}$$

  How to get the parameter ρ? Or the covariance structure $\sigma_e^2 V$ contains two unknown parameters σ_e^2 and ρ.

Is the regression model feasible for Example 1.1?

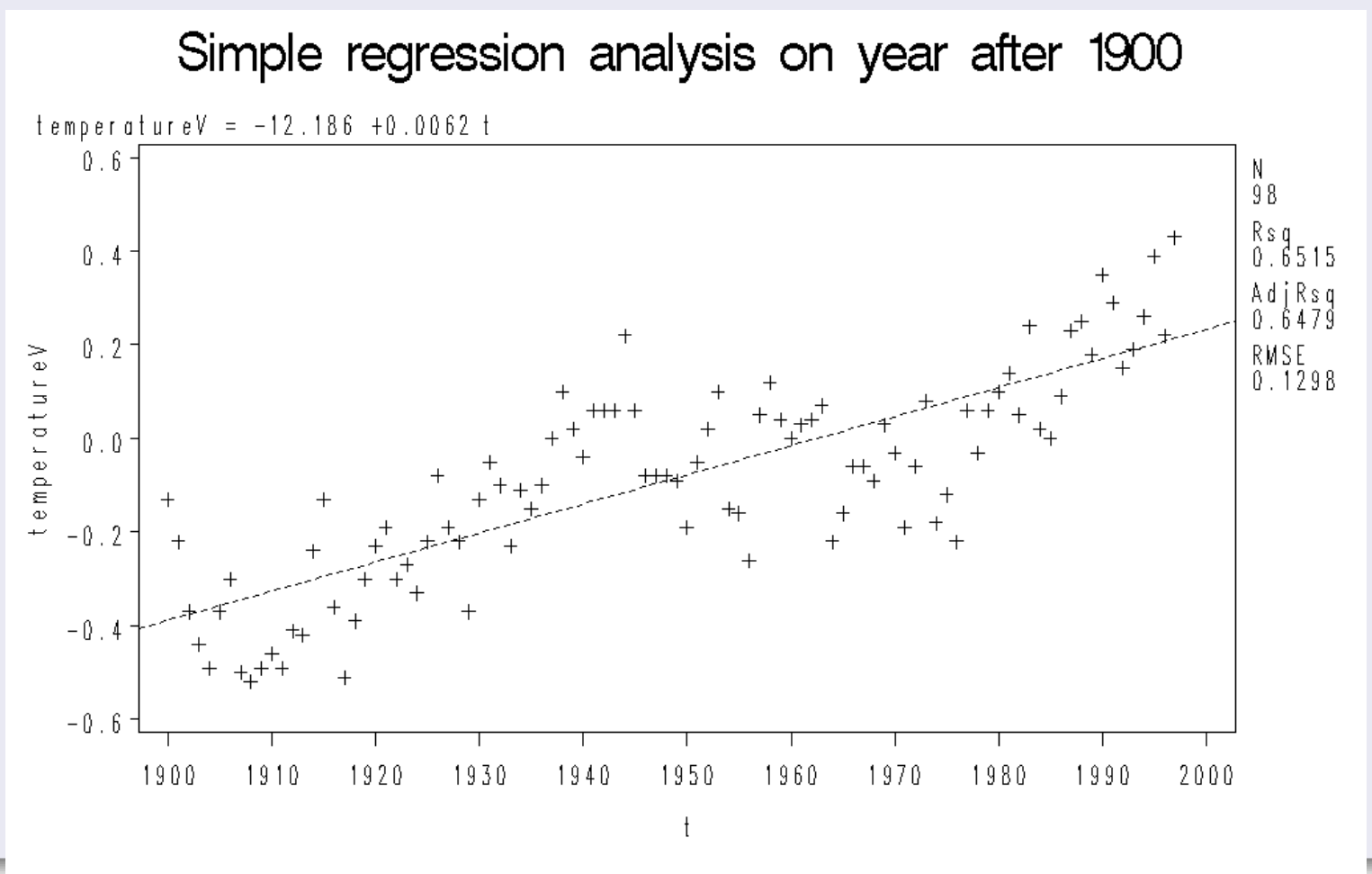
- Assumption 3: $\mathcal{E} \sim N(\mathbf{0}, \sigma_e^2 I_n)$.
  $\hat{\boldsymbol{\beta}} = (X'X)^{-1}X'Y$ is MVUE of $\boldsymbol{\beta}$.

- SAS program for the regression model is as follows:

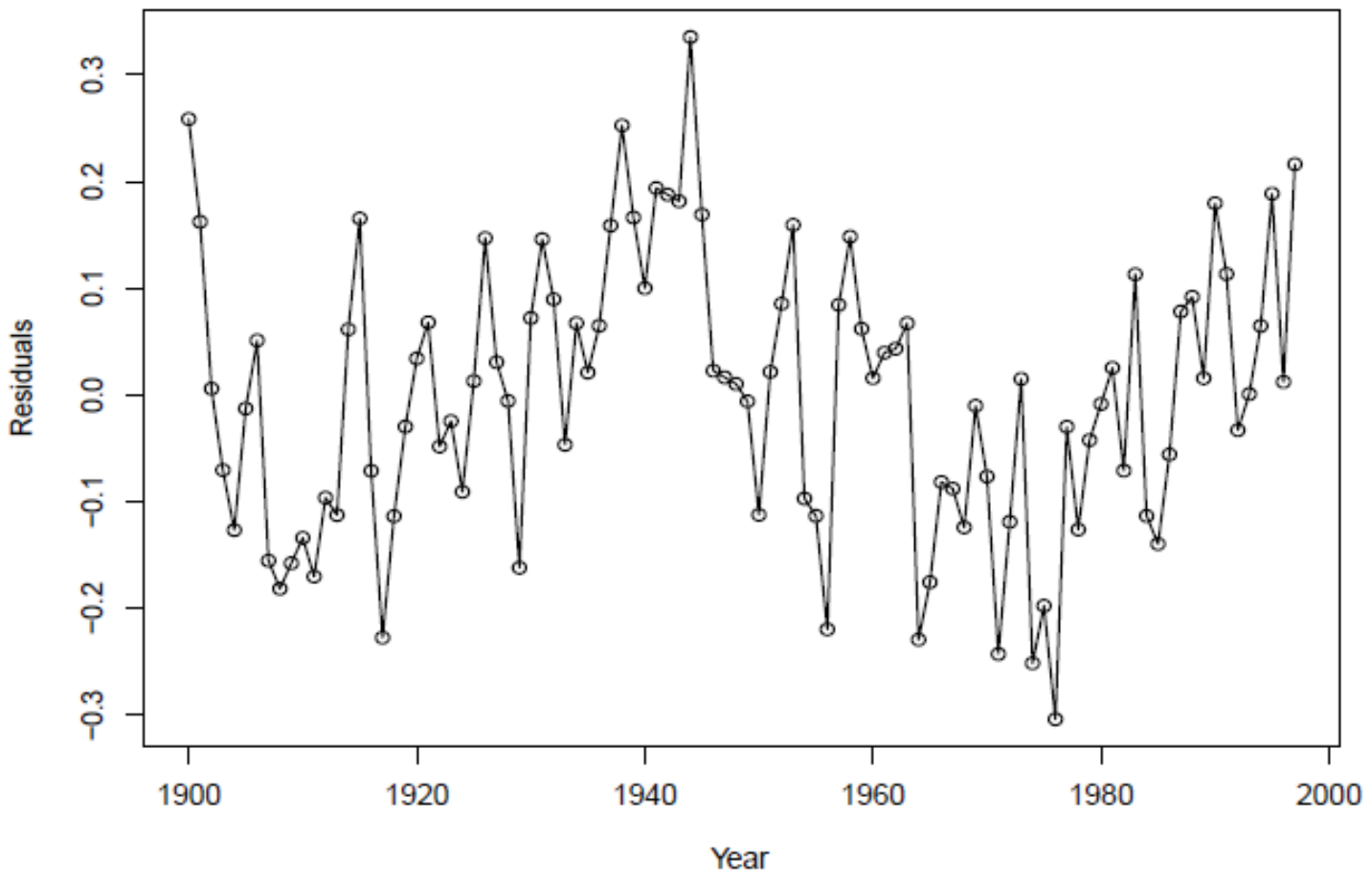
```
data chapter1.example11_1;
set chapter1.example11;
t=year(year); if t < 1900 then delete; run;
proc reg data=chapter1.example11_1;
model temperatureV=t;
title 'Simple regression analysis on year after 1900';
plot temperatureV*t; run;
```

- The fitted model is $\hat{Y} = -12.19 + 0.0062t$.

The fitted linear model



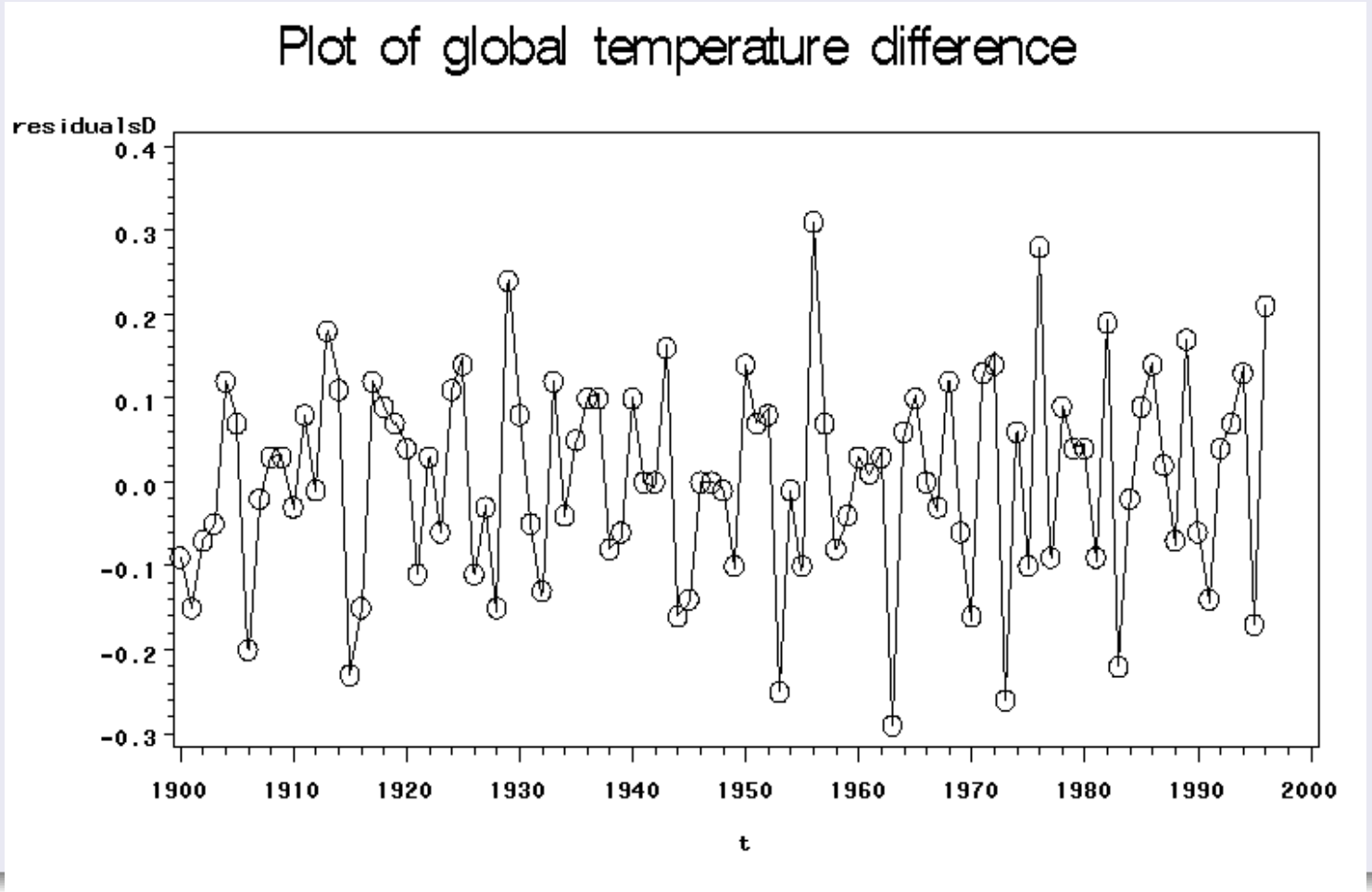
Residuals plot, detrending series plot



Remarks about the fitting of a regression model

- Global temperature data (1900-1997) has a straight line trend fit.
- The sample is not iid. They may be correlated. This implies that the Gauss-Markov model does not fit these "global warming" temperature data. Why?
- From above analysis, due to dependence between global temperatures of two consecutive years, investigating other models is necessary to fit the global warming temperature data.

Considering global temperature difference $\delta_i = y_i - y_{i-1}$



- The difference series $\{\delta_i = y_i - y_{i-1}\}_{i=1901}^{1997}$ "seems to be" uncorrelated.

- This strongly hints that it is reasonable for us to assume that the first difference process $\Delta_i = Y_i - Y_{i-1}$ are identically uncorrected variables, namely,

$$Y_i - Y_{i-1} = e_i \quad \text{or} \quad Y_i = Y_{i-1} + e_i$$

  where $e_{1901}, \cdots, e_{1997}$ are uncorrelated with zero mean.

- Generalization: further assume that the global temperature data can be fitted by

$$Y_t - \mu = \rho(Y_{t-1} - \mu) + e_t, \text{ for } t = 0, 1, \cdots, \tag{1}$$

  where $e_t \sim$ uncorrelated $(0, \sigma_e^2)$ with unknown σ_e^2, ρ is unknown parameter and t denotes time points by year.

Structure of covariance for the model (1)

From model (1), without loss of generality, let $\mu = 0$ and we have

$$\gamma_0 = \mathsf{var}(Y_t) = \mathsf{var}(\rho Y_{t-1} + e_t) = \rho^2 \gamma_0 + \sigma_e^2.$$

So

$$\gamma_0 \equiv \mathsf{var}(Y_t) = \frac{\sigma_e^2}{1-\rho^2}.$$

And

$$\gamma_k \equiv \mathsf{Cov}(Y_t, Y_{t-k}) = \mathsf{E}(Y_t Y_{t-k}) = \mathsf{E}(\rho Y_{t-1} Y_{t-k} + e_t Y_{t-k}) = \rho \gamma_{k-1}$$

for $k = 1, 2, \cdots$. Thus

$$\mathsf{Cov} \begin{pmatrix} Y_1 \\ \vdots \\ Y_n \end{pmatrix} = \frac{\sigma_e^2}{1-\rho^2} \begin{pmatrix} 1 & \rho & \rho^2 & \cdots & \rho^{n-1} \\ \rho & 1 & \rho & \cdots & \rho^{n-2} \\ \cdots & \cdots & \cdots & \cdots & \cdots \\ \rho^{n-1} & \rho^{n-2} & \rho^{n-3} & \cdots & 1 \end{pmatrix}$$

Comparison between model (1) and regression model

- For the simple regression model,

$$Y_i = \mu + \varepsilon_i = \mathbf{x}_i'\boldsymbol{\beta} + \varepsilon_i,$$

where $\varepsilon_1, \cdots, \varepsilon_n$ are correlated with zero mean and $\text{cov}(\varepsilon_i, \varepsilon_j) = \sigma_e^2 \frac{\rho^{|i-j|}}{1-\rho^2}$. We regress Y on explanatory $\mathbf{x}$.

- For model (1),

$$Y_t - \mu = \rho(Y_{t-1} - \mu) + e_t,$$

where $e_1, \cdots, e_n$ are uncorrelated with $e_t \sim (0, \sigma_e^2)$. Y_t is regressed on Y_{t-1}, meaning "regression on itself".

Example 1.2. USC Fall enrollment data

The data in the below Figure 1-2 are the annual fall enrollment counts for USC (Columbia campus only, 1956-2008). The data were obtained from the USC website

http://www.ipr.sc.edu/enrollment/,

which contains the enrollment counts for all campuses of USC.

- There are $n = 53$ observations.
- Measurements are taken **each year**.
- What are the noticeable patterns?
- Predictions?

Figure1-2

Figure 1.2: *University of South Carolina fall enrollment data. Number of students registered for classes on the Columbia campus during 1956-2008.*

Example 1.3. Exchange rate data (US vs British Pound)

The pound sterling, often simply called "the pound", is the currency of the United Kingdom and many of its territories. The data in Figure 1.3 are weekly exchange rates of the US dollar and the British pound between the years 1980 and 1988.

- There are $n = 470$ observations.
- Measurements are taken **each week**.
- What are the noticeable patterns?
- Predictions?

Figure1-3

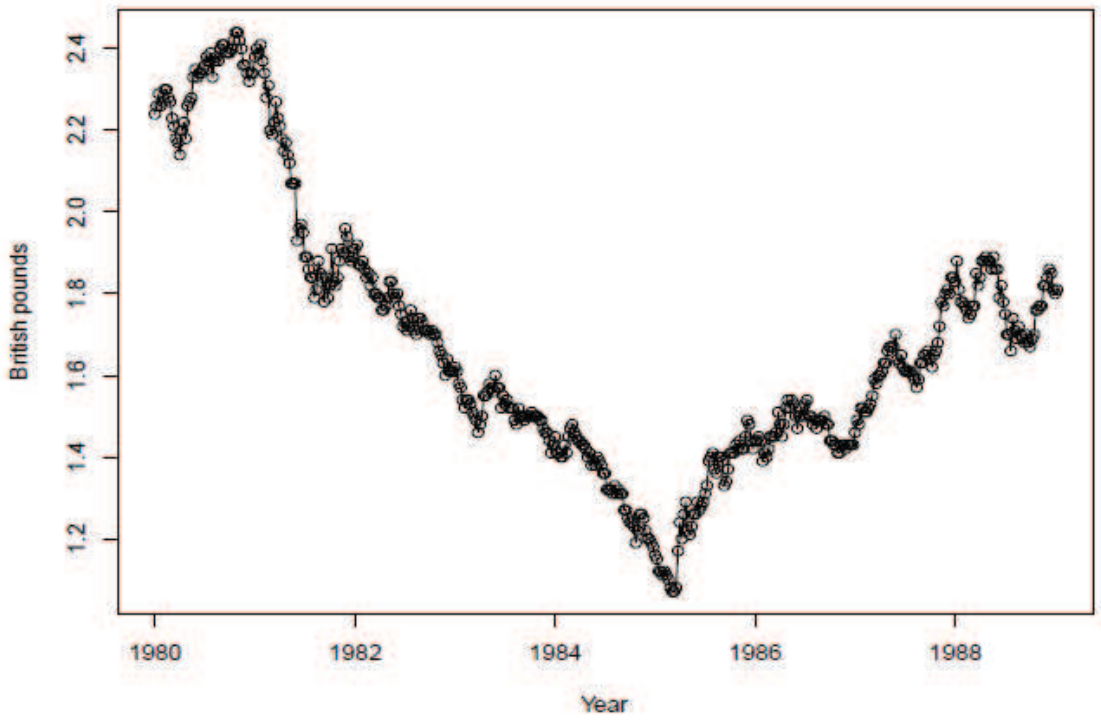


Figure 1.3: *Exchange rate data. Weekly exchange rate of US dollar compared to the British pound, from 1980-1988.*

Example 1.4. CREF bond fund data

TIAA-CREF is the leading provider of retirement accounts and products to employees in academic, research, medical, and cultural institutions. The data in Figure 1.4 are daily values of one unit of the CREF (College Retirement Equity Fund) Bond fund from 8/26/04 to 8/15/06. This time series (and more current values of it) is of particular interest to your instructor!

- There are $n = 500$ observations.
- Measurements are taken **each trading day**.
- What are the noticeable patterns?
- Predictions? (Prof. Ai's retirement depends on these).

Figure1-4

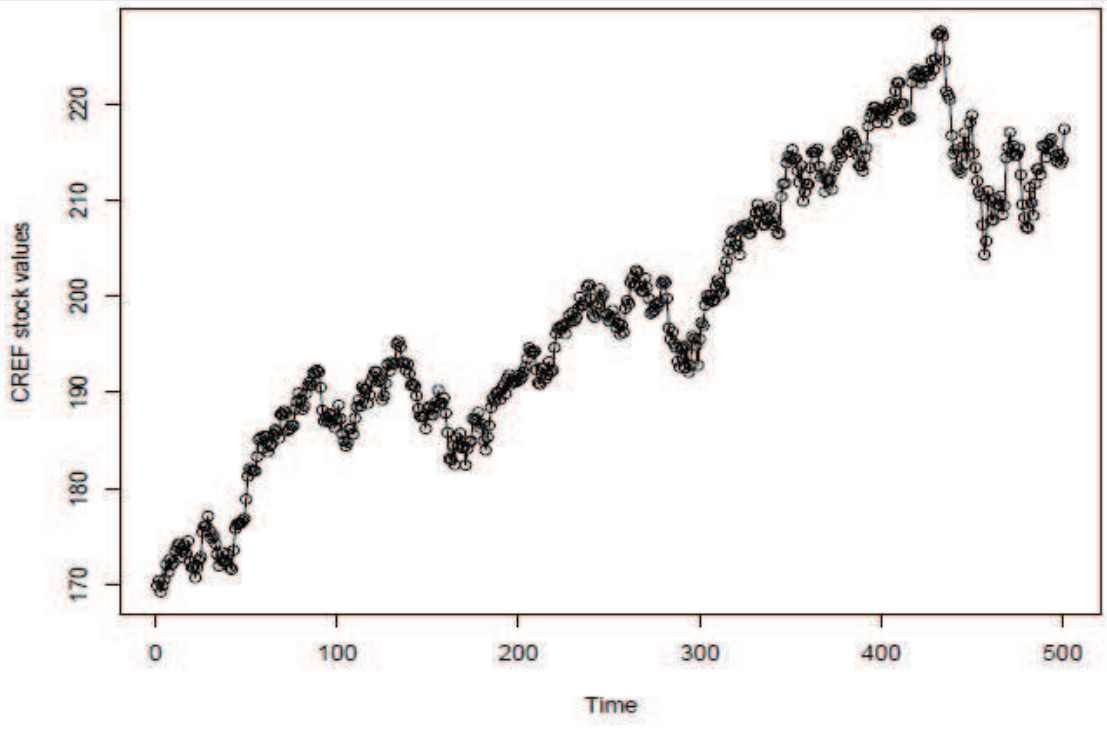


Figure 1.4: *CREF bond fund data. Daily values of one unit of CREF stock values: August 26, 2004 to August 15, 2006.*

Example 1.5. S & P500 Index data

The S&P500 is a capitalization-weighted index (published since 1957) of the prices of 500 large-cap common stocks actively traded in the United States. The stocks included in the S&P 500 are those of large publicly held companies that trade on either of the two largest American stock market companies: the NYSE and the NASDAQ. The data in Figure 1.5 are the daily S&P500 Index prices measured during June 6, 1999 to June 5, 2000.

- There are $n = 254$ observations.
- Measurements are taken **each trading day**.
- What are the noticeable patterns?
- Predictions?

Figure1-5

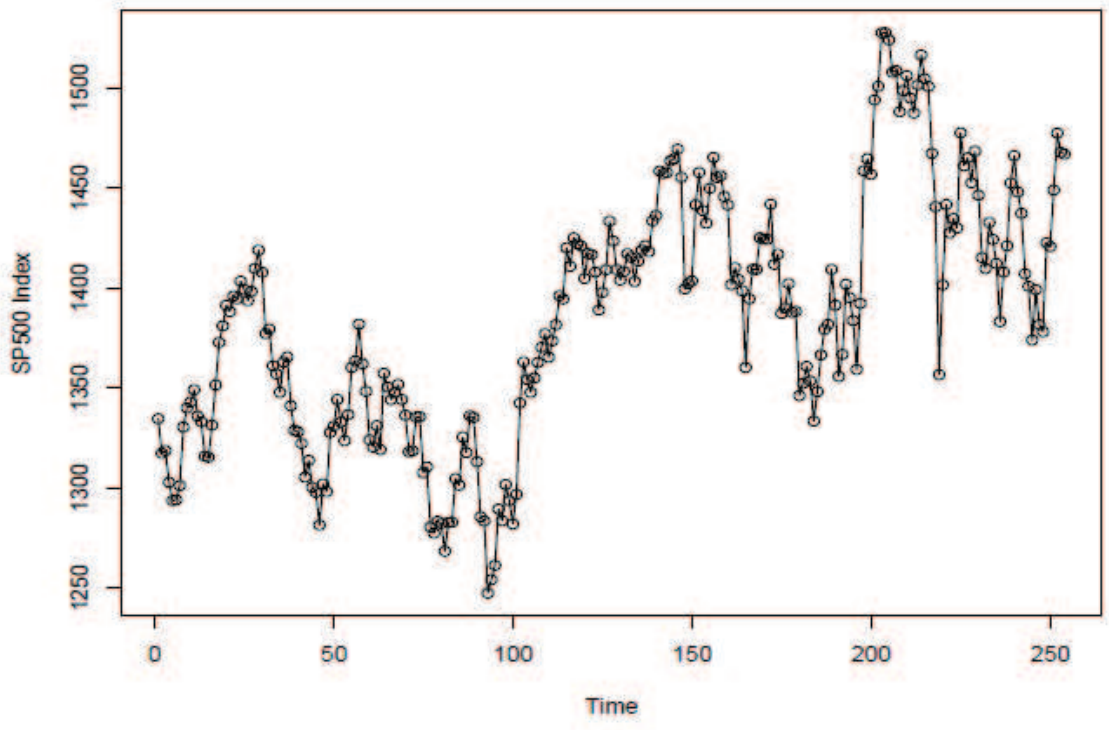


Figure 1.5: *S&P Index price data. Daily values of the index from June 6, 1999 to June 5, 2000.*

Example 1.6 Milk production data

In the United States, cow's milk is produced on an industrial scale, and is by far the most commonly consumed form of milk. Commercial dairy farming using automated milking equipment produces the vast majority of milk in the U.S. The data in Figure 1.6 are the monthly U.S. milk production (measured in millions of pounds) from January, 1994 to December, 2005.

- There are $n = 144$ observations.
- Measurements are taken **each month**.
- What are the noticeable patterns?
- Predictions?

Figure1-6. U.S. milk production data

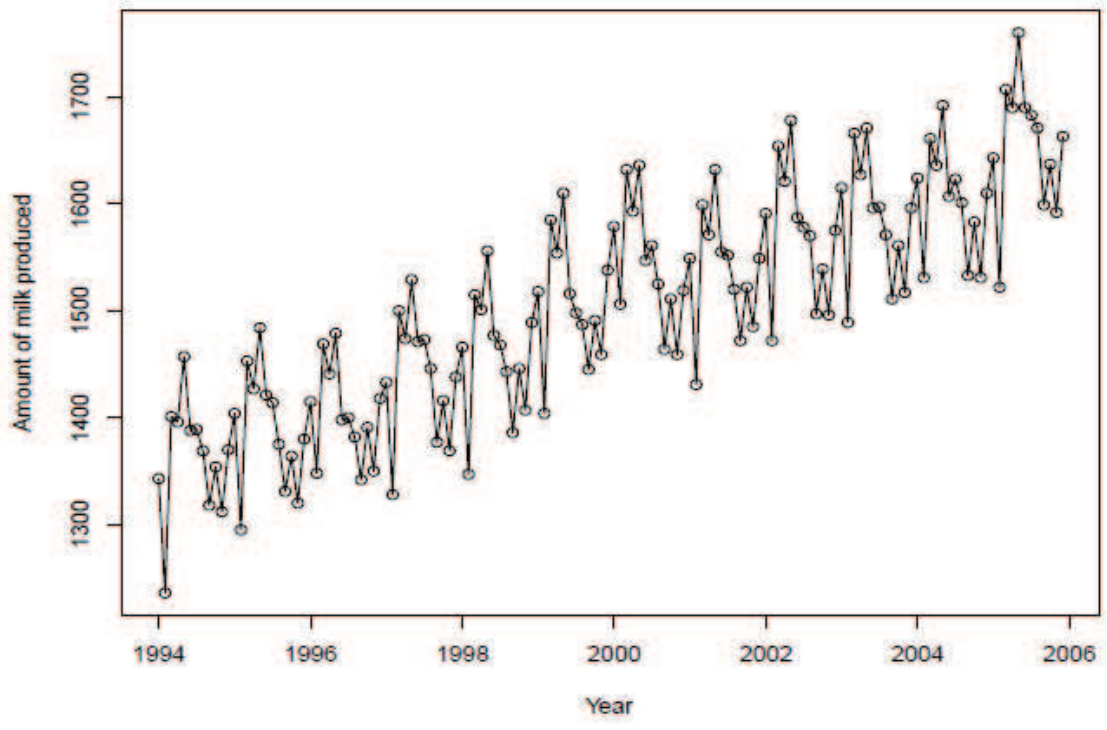


Figure 1.6: *United States milk production data. Monthly production figures, measured in billions of pounds, from January, 1994 to December, 2005.*

Example 1.7. Star brightness data

Two factors determine the brightness of a star: its luminosity (how much energy it puts out in a given time) and its distance from the Earth. The data in Figure 1.7 are nightly brightness measurements (in magnitude) of a single star over a period of 600 nights.

- There are $n = 600$ observations.
- Measurements are taken **each night**.
- What are the noticeable patterns?
- Predictions?

Figure1-7. Star brightness data

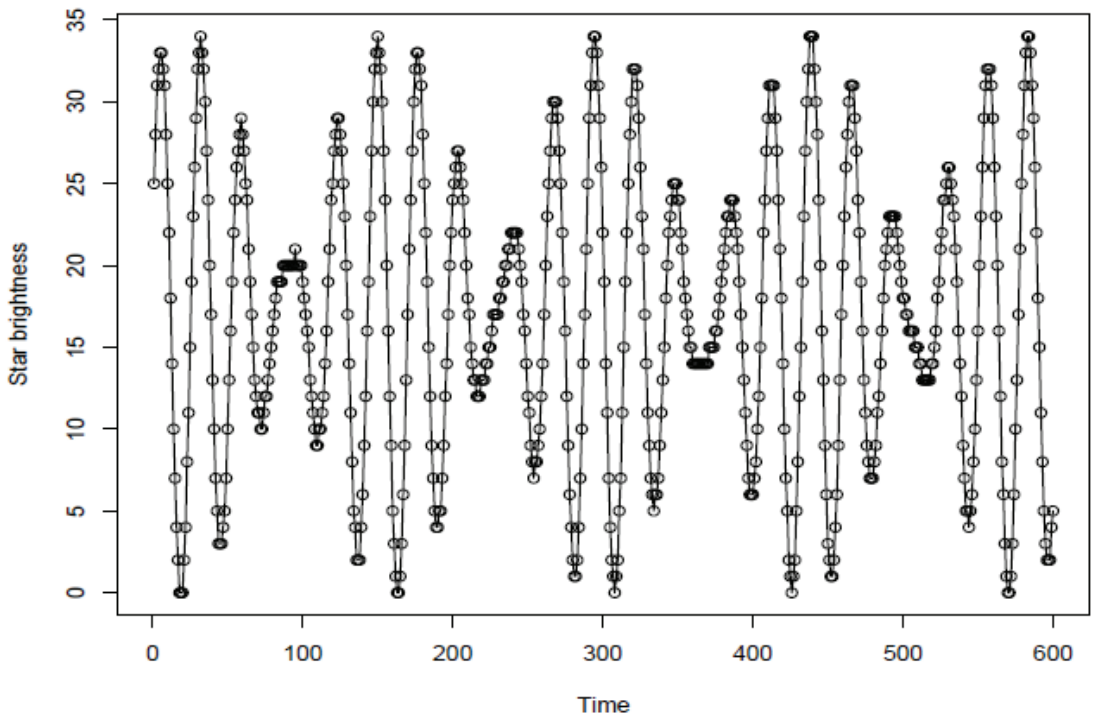


Figure 1.7: *Star brightness data. Measurements for a single star taken over 600 consecutive nights.*

Example 1.8. Airline mile data

The Bureau of Transportation Statistics publishes monthly passenger traffic data reflecting 100 percent of scheduled operations for U.S. airlines. The data in Figure 1.11 are monthly U.S. airline passenger miles traveled from 1/1996 to 5/2005.

- There are $n = 113$ observations.
- Measurements are taken **each month**.
- What are the noticeable patterns? Which months correspond to the most-traveled? See pp 14 (notes).
- Predictions?

Figure1-8. Airline mile data

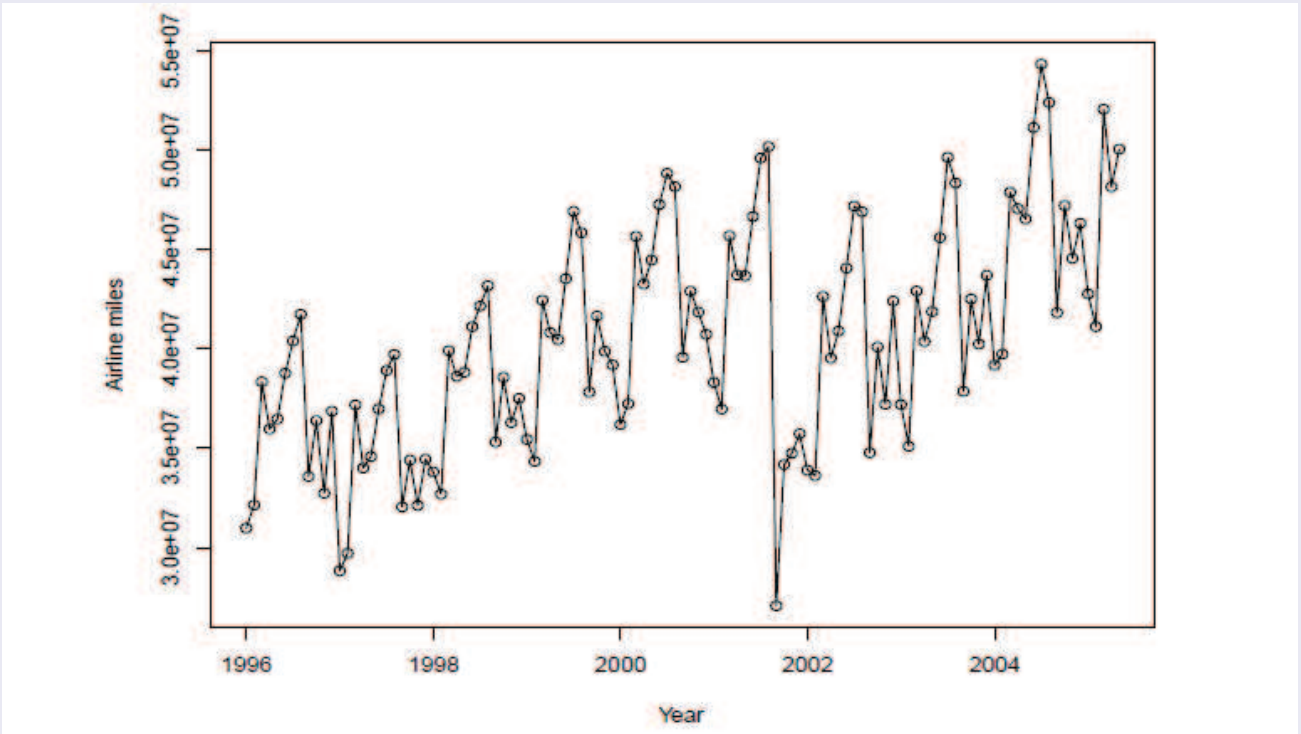


Figure 1.8: *Airline passenger mile data. The number of miles, in thousands, traveled by passengers in the United States from January, 1996 to May, 2005.*

Example 1.9. Crude price data

Crude oil prices behave much as any other commodity with wide price swings in times of shortage or oversupply. The crude oil price cycle may extend over several years responding to changes in demand as well as OPEC and non-OPEC supply. The data in Figure 1.9 are monthly spot prices for crude oil (measured in U.S. dollars per barrel) from Cushing, OK.

- There are $n = 321$ observations.
- Measurements are taken **each month**.
- What are the noticeable patterns?
- Predictions?

Figure1-9. Crude price data

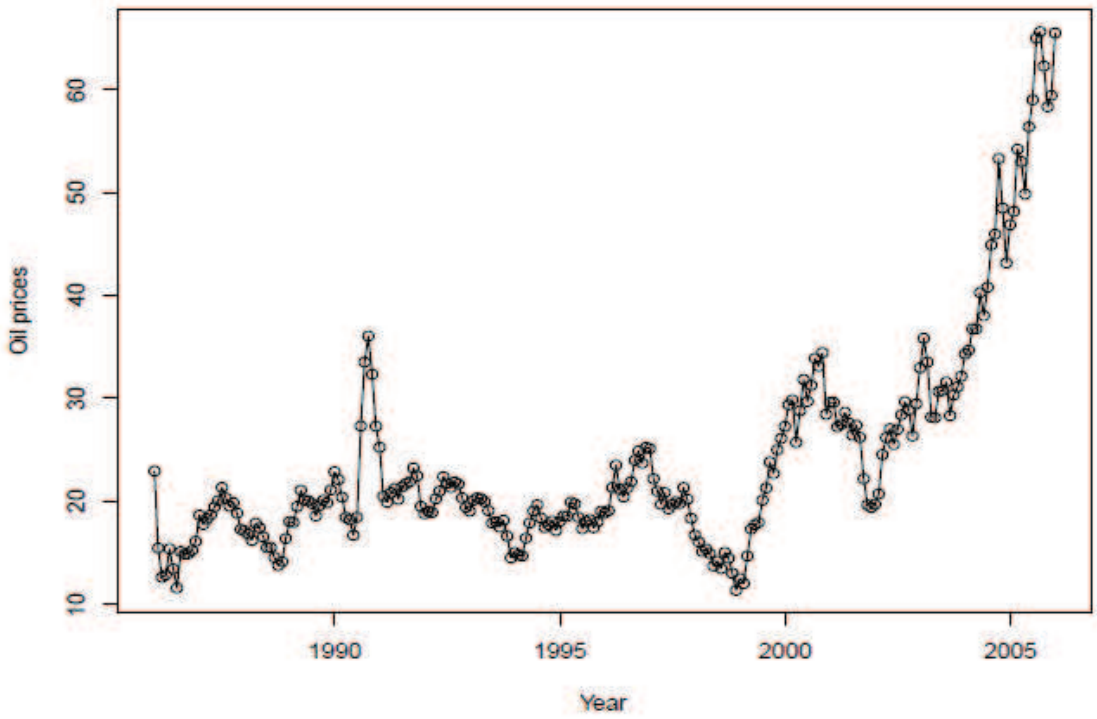


Figure 1.9: *Crude oil price data. Monthly spot prices in dollars from Cushing, OK, from 1/1986 to 1/2006.*

Example 1.10. Los Angeles rainfall data

Los Angeles averages 15 inches of precipitation annually, which mainly occurs during the winter and spring (November through April) with generally light rain showers, but sometimes as heavy rainfall and thunderstorms. The data in Figure 1.10 are annual rainfall totals for Los Angeles during 1878-1992.

- There are $n = 115$ observations.
- Measurements are taken **each year**.
- What are the noticeable patterns?
- Predictions?

Figure1-10. Los Angeles rainfall data

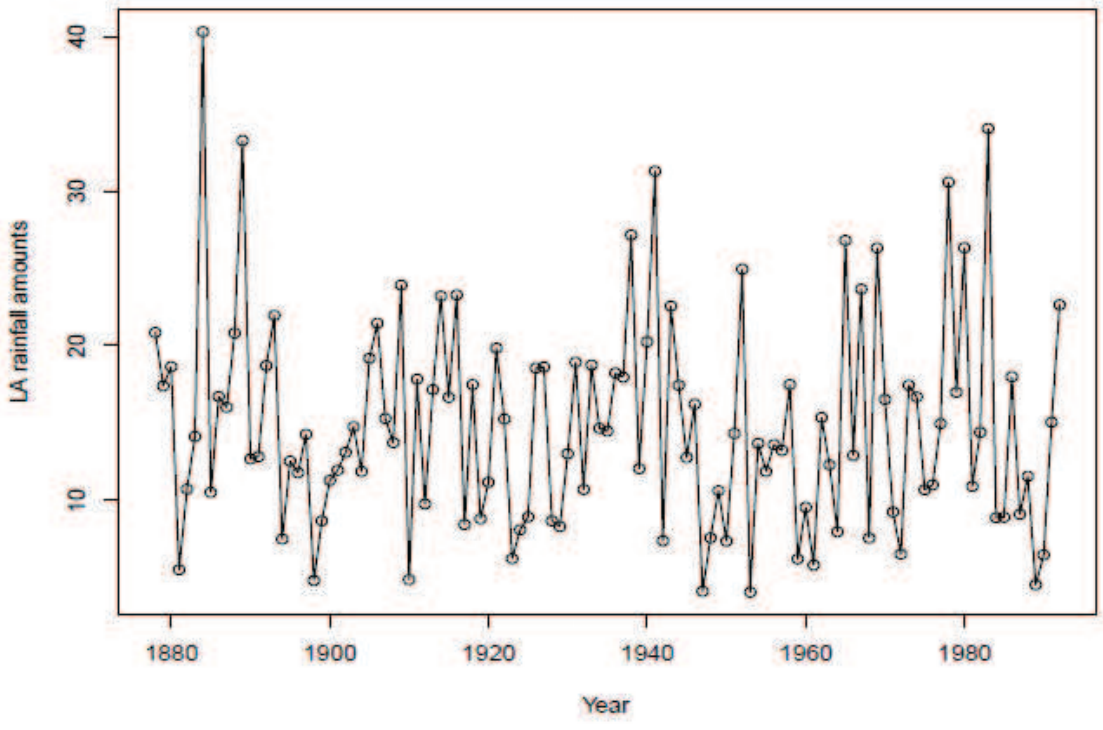


Figure 1.10: *Los Angeles rainfall data. Annual precipitation measurements, in inches, during 1878-1992.*

A sequence of ordered data: Time series

A **time series** is a sequence of ordered data.

- The "ordering" refers generally to time, but other orderings could be envisioned (e.g., over space, etc.).
- In this class, we will be concerned almost exclusively with time series that are
  - measured on a single continuous random variable Y
  - equally spaced in discrete time; that is, we will have a single measurement of Y at each second, hour, day, week, month, year, etc.

Time series data and other data

- Time series data have a natural temporal ordering. This makes time series analysis distinct from other common data analysis problems, in which there is no natural ordering of the observations, e.g. explaining people's wages by reference to their education level, where the individuals' data could be entered in any order.
- Time series analysis is also distinct from spatial data analysis where the observations typically relate to geographical locations, e.g. accounting for house prices by the location as well as the intrinsic characteristics of the houses.
- A time series model will generally reflect the fact that observations close together in time will be more closely related than observations further apart.
- Time series models will often make use of the natural one-way ordering of time so that values for a given period will be expressed as deriving in some way from past values, rather than from future values.

A variety of fields needs time series model to fit data

- **business** and **economics**, daily stock prices, weekly interest rates, quarterly sales, yearly earnings, monthly supply, etc.
- **agriculture**, annual yields (crop production), daily prices, annual herd sizes, etc.
- **engineering**, sound, electric signals, voltage measure^{ts}, etc.
- **natural sciences**, chemical yields, turbulence in ocean waves, earth measurements, etc.
- **medicine**, EEG and EKG measurements on patients, drug concentrations, blood pressure readings, etc.
- **epidemiology**, the number of H1N1 cases per day, the number of health-care clinic visits per month, etc.
- **meteorology**, daily high temperatures, annual rainfall, hourly wind speeds, etc.
- **social sciences**, annual birth and death rates, accident frequencies, crime rates, school enrollments, etc.

Purpose of time series analysis

The purpose of time series analysis is twofold:

1. to model the stochastic (random) mechanism that gives rise to the series of data
2. to predict (forecast) the future values of the series based on the previous history.

The analysis of time series data calls for a "new way of thinking" when compared to other statistical methods courses. Essentially, we get to see only a single measurement from a population (at time t) instead of a sample of measurements at a fixed point in time (cross-sectional data).

Features of times series

- The special feature of time series data is that they are dependent! Instead, obs$^n s$ are correlated through time.
  - Correlated data are generally more difficult to analyze.
  - Statistical theory in the absence of independence becomes markedly more difficult.
- Most classical statistical methods (e.g., regression, analysis of variance, etc.) assume that observations are statistically independent. For example, in the general linear model

$$\mathbf{Y} = X\boldsymbol{\beta} + \mathcal{E},$$

  typically assume that $\mathcal{E} \sim N(\mathbf{0}, \sigma_e^2 I)$.
- Complications are further exacerbated by additional trends or seasonal variation (seasonality) that may be difficult to identify and remove.
- The data may also be further contaminated by outliers or missing observations.

Box-Jenkins modeling approach

Our overarching goal in this course is to build (and use) time series models for data. This breaks down into different parts.

1. Model specification (identification)
   - Consider different classes of time series models for stationary processes.
   - Use descriptive statistics, graphical displays, subject matter knowledge, etc. to make sensible candidate selections.
   - Abide by the Principle of Parsimony.
2. Model fitting
   - Once a candidate model is chosen, estimate the parameters in the model.
   - We will use least squares and/or the method of maximum likelihood to do this.
3. Model diagnostics
   - Use statistical inference and graphical displays to check how well the model fits the data.
   - This part of the analysis may suggest the candidate model is inadequate and may point to more appropriate models.

Time series plot

The time series plot (or sequence plot) is the most basic graphical display in the analysis of time series data. The plot is a basically a scatterplot of Y_t versus t, with straight lines connecting the points. Notationally,

$$Y_t = \text{value of the variable } Y \text{ at time } t,$$

for $t \in \{0, 1, 2, \cdots, \infty\}$. The subscript t tells us which time point the measurement Y_t corresponds to. Note that in the sequence $Y_1, Y_2, \cdots, Y_n$, the subscripts are very important because they correspond to a particular ordering of the data.

Time series plot

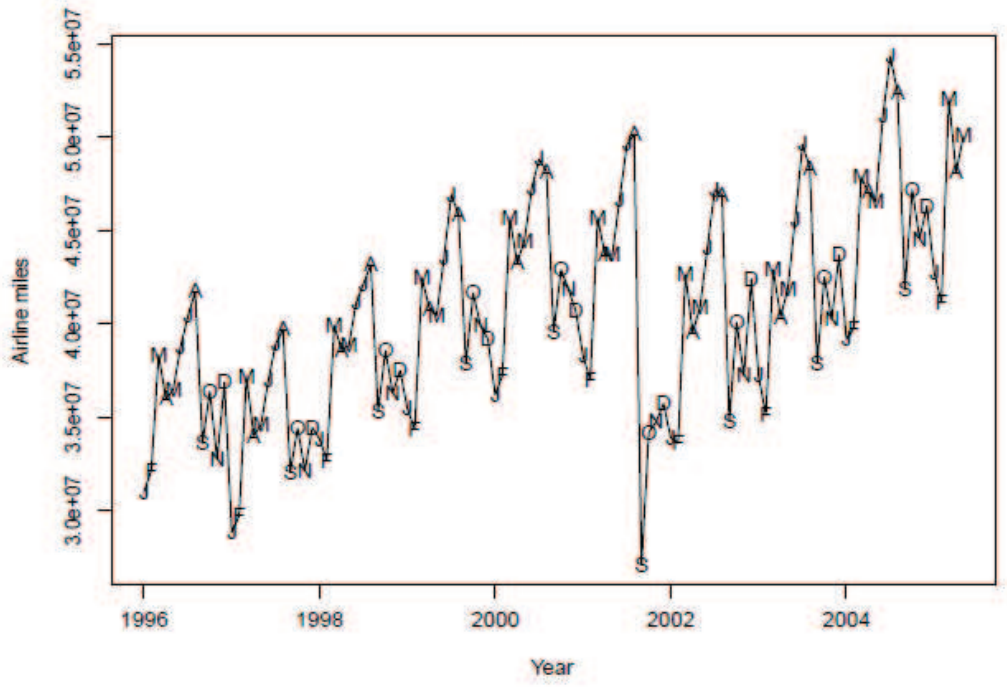


Figure 1.11: *Airline passenger mile data. The number of miles, in thousands, traveled by passengers in the United States from January, 1996 to May, 2005. Monthly plotting symbols have been added.*

Time series plot

The time series plot is vital, both to describe the data and to help formulating a sensible model. Here are some simple, but important, guidelines when constructing these plots.

- Give a clear, self-explanatory title or figure caption.
- State the units of measurement in the axis labels or figure caption.
- Choose the scales carefully (including the size of the intercept). Default settings from software may be sufficient.
- Label axes clearly.
- Use special plotting symbols where appropriate; e.g., months of the year, days of the week, actual numerical values for outlying values, etc.

Thank You for Your Attention !

Have a nice day !