MSDS 596 Regression & Time Series

Lecture 09 Time Series Exploratory Analysis

Department of Statistics
Rutgers University

Oct 29, 2020

Schedule

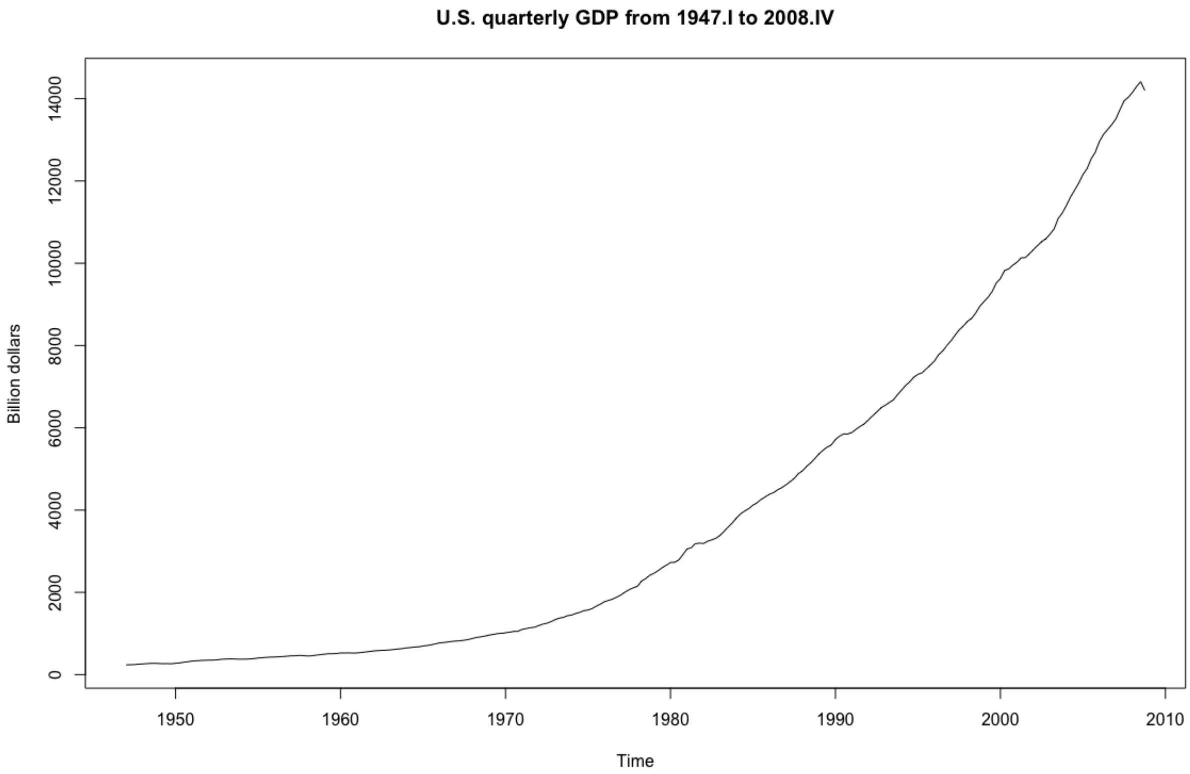
Week	Date	Topic
1	9/3	Intro to linear regression (JF1,2)
2	9/10	Estimation (JF2)
3	9/17	Inference I (JF3)
4	9/24	Inference II (JF3)
5	10/1	Inference and prediction (JF4,5)
6	10/8	Explanation; model diagnostics (JF6-8)
7	10/15	Transformation and model selection (JF9-10)
8	10/22	Shrinkage methods (JF11)
9	10/29	Time series exploratory analysis (CC2,3)
10	11/5	Linear time series: ARIMA models (CC4-5)
11	11/12	Model specification and estimation (CC6,7)
12	11/19	Diagnostics and forecasting (CC8,9)
	11/26	(no class)
13	12/3	Seasonal models (CC10)
14	week of 12/7	Project & final evaluation

Time Series

- Time series: a collection of observations generated sequentially through time. Denoted by $\{x_1, x_2, \ldots, x_T\}$.
  - Can be thought of as a *realization* of a stochastic process $\{X_t,\ t \in \mathcal{T}\}$;
- Data are ordered with respect to time, and successive observations are usually expected to be dependent.
- Time series analysis studies the dependence among adjacent observations.
- Time series emerge from a wide range of applications:
  - Business, finance and economics: stock prices, sales, GDP growth, unemployment rate;
  - Engineering: signal processing, communication;
  - Social sciences: birth rates, divorce rates, school enrollments;
  - Medicine and epidemiology: longitudinal data, neural spike-train;
  - Earth and environmental sciences: earthquake, precipitation, etc.

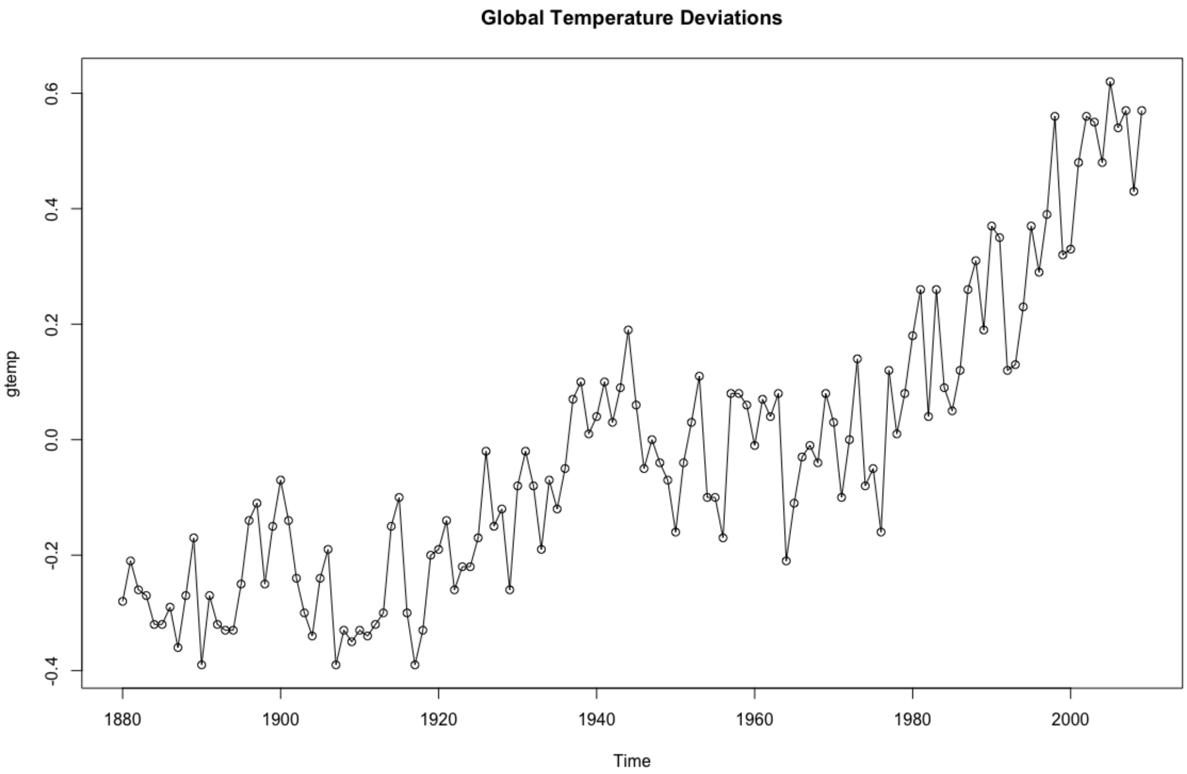
Quarterly US real GDP

The series exhibits a exponential trend, showing the growth of the U.S. economy.



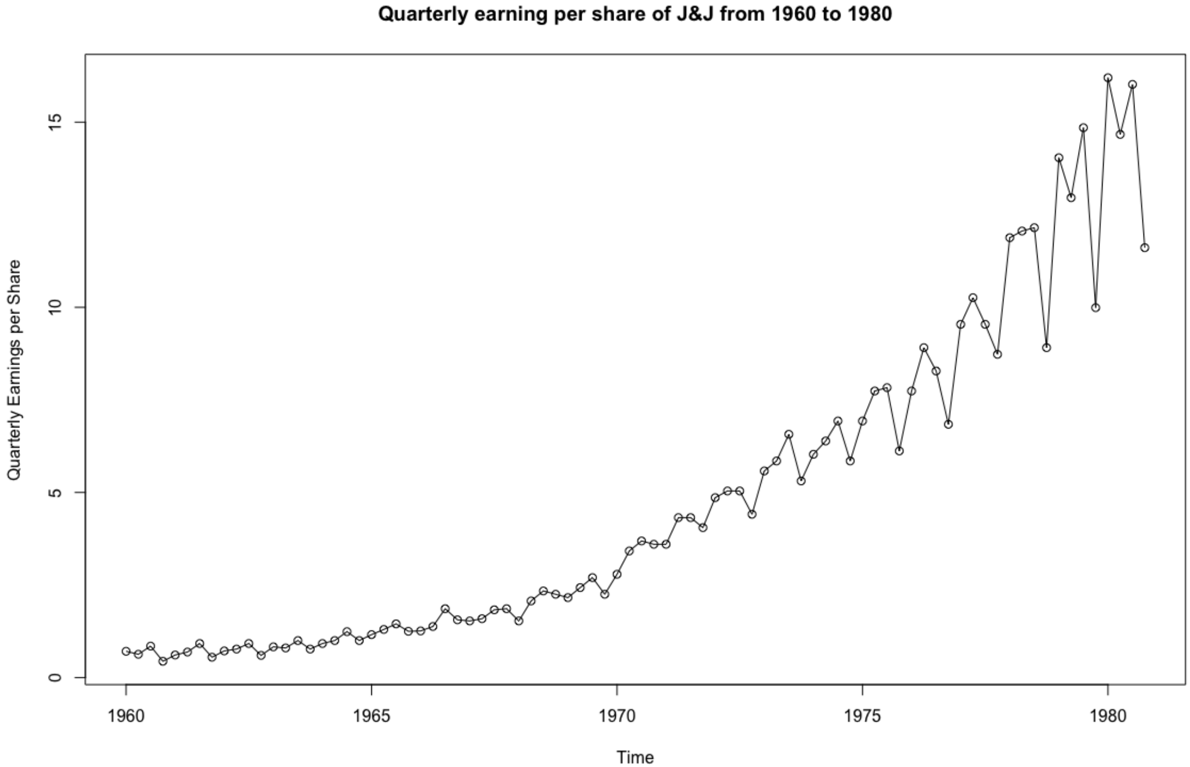
Global warming

Global mean land-ocean temperature index from 1880 to 2009, with the base period 1951–1980. The data are deviations, measured in degrees (C), from the 1951-1980 average.



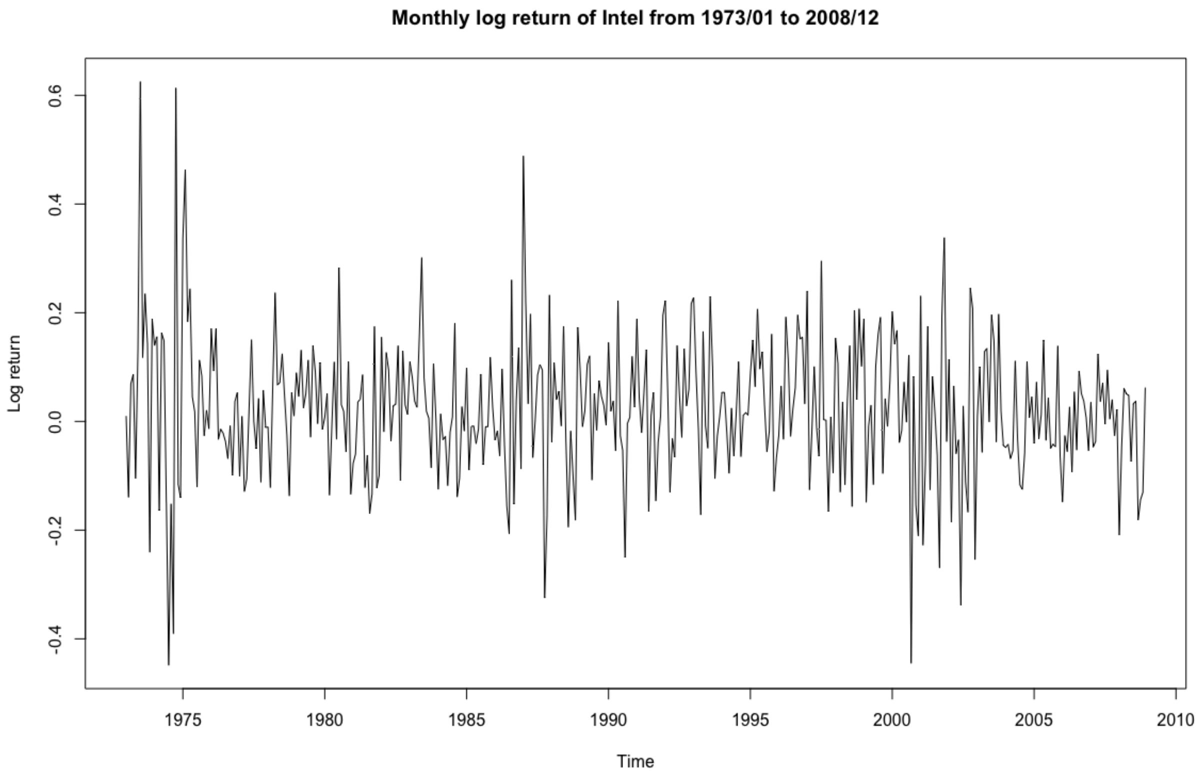
J&J quarterly earnings

A slowly increasing underlying trend, superimposed by what seems to be quarterly regular variations.



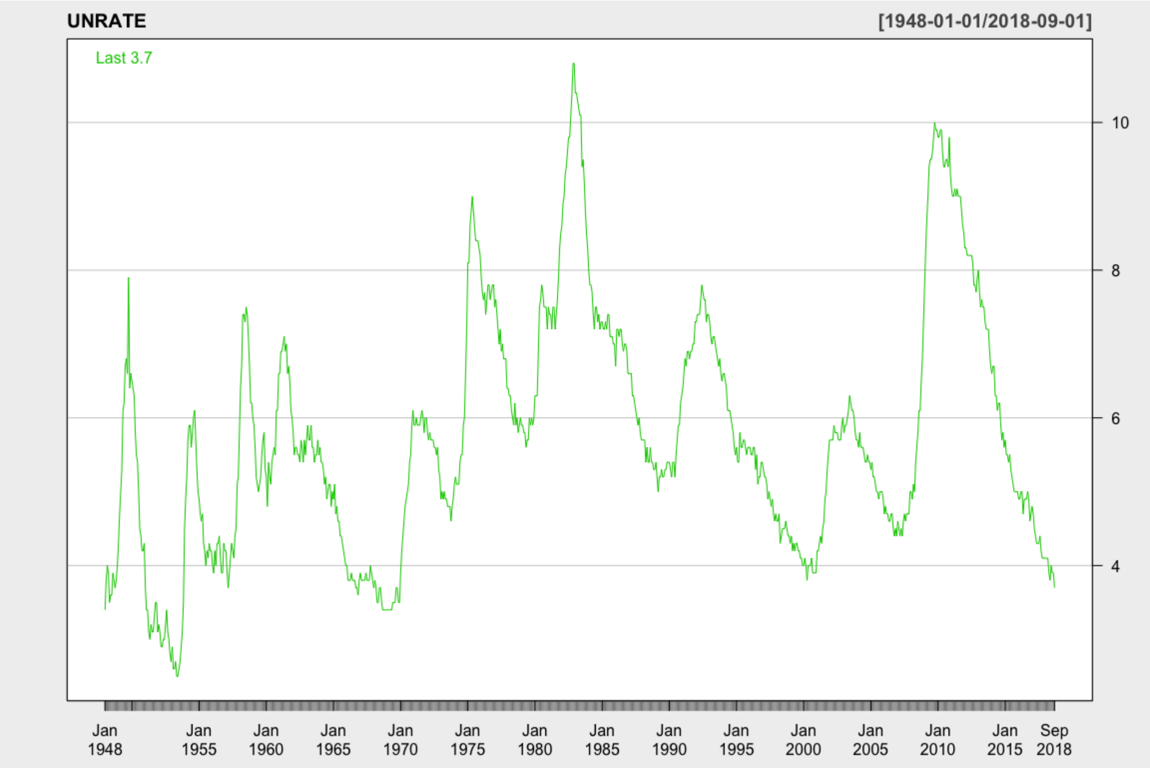
Monthly log returns of Intel

The mean of the series appears to be stable with an average return of approximately zero, however, the volatility (or variability) of data changes over time. The data shows volatility clustering.



Monthly unemployment rate (seasonally adjusted)

There appears to be a slow but upward trend in the overall unemployment rate. Also, the unemployment rate tends to increase rapidly and decrease slowly. Threshold autoregressive (TAR) models can capture this feature.



Goals of time series analysis

- Probabilistic modeling of underlying processes
    - Model Specification
    - Model Fitting
    - Model Diagnostics
- Forecasting
- Control and intervention.

Measure of dependence: autocorrelation

- Consider time series $\{x_1, x_2, \ldots, x_T\}$.
- The mean function is

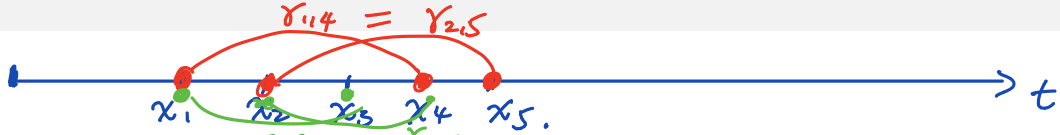
$$\mu_t = \mathbb{E}(x_t), \quad t \in \mathcal{T}.$$

- The autocovariance function is

$$\gamma_{s,t} = \mathsf{Cov}(x_s, x_t) = \mathbb{E}[(x_s - \mu_s)(x_t - \mu_t)].$$

  Note that $\gamma(t, t)$ is the variance of x_t.
- The autocorrelation function is

$$\rho_{s,t} = \mathrm{Corr}(x_s, x_t) = \frac{\gamma_{s,t}}{\sqrt{\gamma_{s,s}\gamma_{t,t}}}.$$



- A stochastic process $\{x_t : \ t \in \mathcal{T}\}$ is said to be weakly stationary or stationary, if
  - (i) μ_t is a constant over time, $\mu_t = \mu$
  - (ii) $\gamma_{s,t} = \gamma_{k,l}$ for all $s, t, k, l \in \mathcal{T}$ such that $|s-t| = |k-l|$.
- The mean function μ_t is a constant over time, which we denote by μ.
- The autocovariance function is written as

$$\gamma_{-h} = \gamma_h = \mathrm{Cov}(x_t, x_{t+h}) = \mathbb{E}[(x_t - \mu)(x_{t+h} - \mu)].$$

  Note that γ_0 is the variance of x_t. i.e. $Var(X_t) = \gamma_0 \ \forall t$
- The autocorrelation function is defined as

$$\rho_h = \mathrm{Corr}(x_t, x_{t+h}) = \frac{\gamma_h}{\gamma_0}.$$

Examples of stochastic processes

Find the mean, variance, and covariance functions of the following stochastic processes:

1. White noise: $e_t \sim [0, \sigma^2]$ i.i.d.;
2. Random walk: $x_t = x_{t-1} + e_t$;
3. A moving average: $x_t = 0.5e_t + 0.5e_{t-1}$; or $x_t = \frac{1}{5}(e_t + e_{t-1} + e_{t-2} + e_{t-3} + e_{t-4})$
4. A random cosine wave: for $t = 0, \pm 1, \pm 2, \ldots$ and $\Phi \sim Unif[0, 1]$,

$$x_t = \cos\left[2\pi\left(\frac{t}{12} + \Phi\right)\right].$$

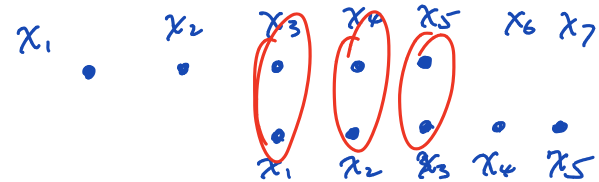
See R markdown for illustrations of each. Are they stationary processes?

Estimation of autocorrelations

For stationary TS:

- Sample mean: $\hat{\mu} = \bar{x} = \frac{1}{n}\sum_{t=1}^{T} x_t$.

$h=2$, $T=5$



- Sample autocovariance function:

$$\hat{\gamma}_h = \frac{1}{T}\sum_{t=|h|+1}^{T}(x_t - \bar{x})(x_{t-|h|} - \bar{x}), \quad 1 - T \leq h \leq T - 1.$$

- Sample autocorrelation function: $\hat{\rho}_h = \hat{\gamma}_h/\hat{\gamma}_0$.
- If x_t are i.i.d. with finite fourth moment, then for any fixed $h \neq 0$

$$\sqrt{T} \cdot \hat{\rho}_h \Rightarrow N(0, 1).$$

Time Series Exploratory Analysis

- The first step in time series analysis is to plot the data.
- Type of nonstationarity.
  - Mean is not constant (trend).
  - Variance is not constant.
  - Seasonal pattern.
- Variance stabilization – data transformaton.
  - If the variance increases as the mean level, log transformation is often used: $\log x_t$.
  - Power transformation x_t^λ.

Time Series Modeling Approaches

Two general approaches to handling trend and seasonality in time series:

- **Decomposition method**-assumes trend and seasonality to be non-stochastic (Ch. 3). Today we discuss estimation/removal of trend and seasonality by the decomposition method.
- Box-Jenkin's approach (stochastic trends) (Ch. 4 and on).

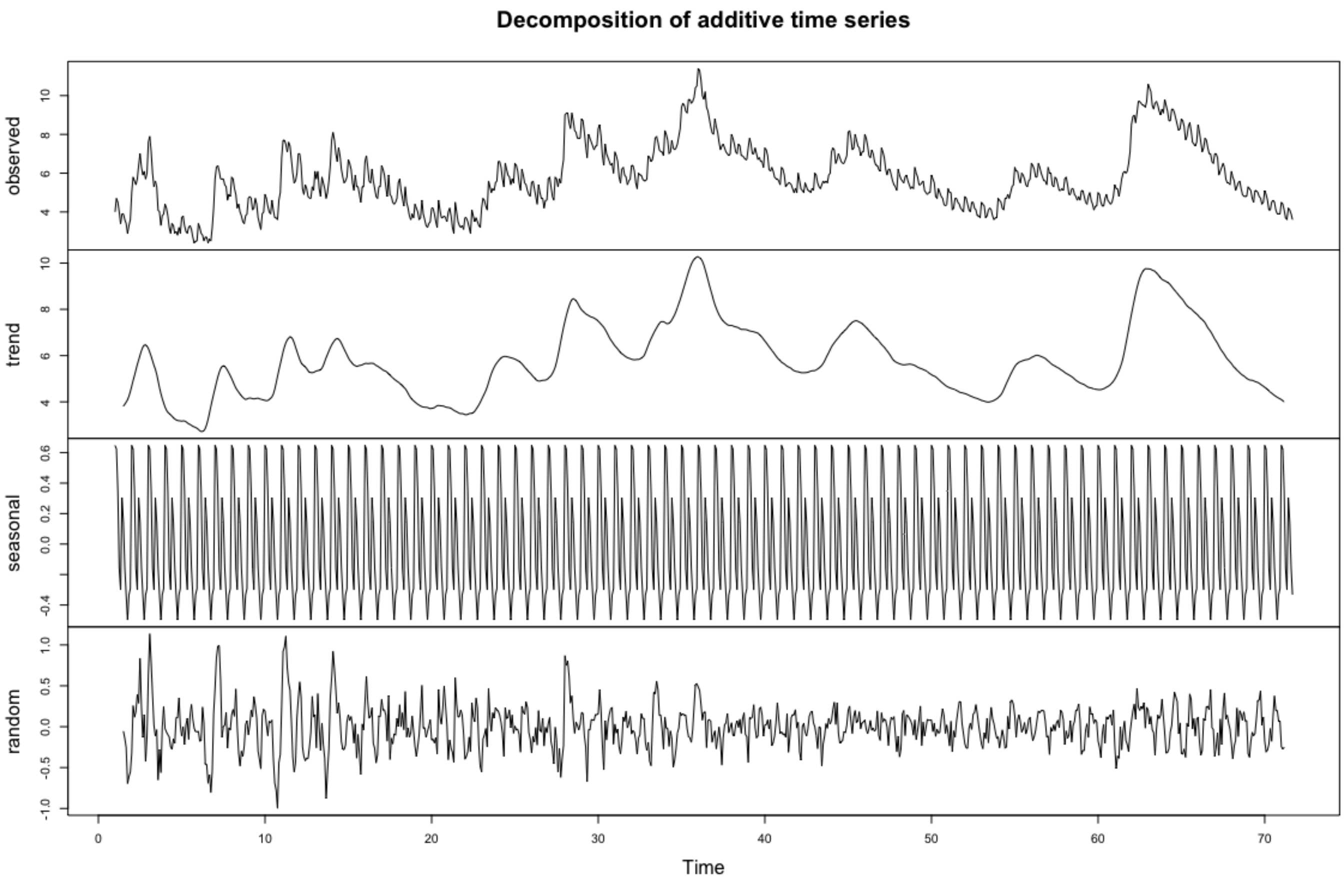
Classical decomposition model

To deal with the trend and the seasonal pattern, we introduce the classical decomposition model:

$$x_t = m_t + s_t + e_t.$$

- m_t is a slowly changing function know as a *trend component*;
- s_t is a function with known period d: referred to as a *seasonal component*;
- e_t is a *random noise component*, which is stationary.

Example: classical decomposition of non-seasonally adjusted monthly unemployment rate



Classical decomposition model

$$x_t = m_t + s_t + e_t.$$

Two approaches:

- View m_t and s_t as deterministic, and estimate them. Once the components m_t and s_t are identified, we can use the theory of stationary processes to fit a probabilistic model for e_t to analyze its properties.
- Apply difference operators repeatedly to the data x_t until the differenced observations resemble a realization of some stationary process.

Elimination of a trend: least squares

Select a model for the trend component m_t:

- Linear trend: $m_t = \beta_0 + \beta_1 t$.
- Quadratic trend: $m_t = \beta_0 + \beta_1 t + \beta_2 t^2$.
- Polynomial trend: $m_t = \beta_0 + \beta_1 t + \cdots + \beta_k t^k$.
- Other deterministic function, e.g. $m_t = \beta_0 + \beta_1 \sin[2\pi(t - t_0)/k]$.

Proceed with fitting the model:

- Estimate the parameters β_j using least squares.
- Check the stationarity of the residuals.
- Can be used to predict future values – proceed with caution (extrapolation).
- Do not over fit – use model selection.

Example: U.S. Population

U.S. population from 1790 to 1980. We attempt to fit a quadratic trend:

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

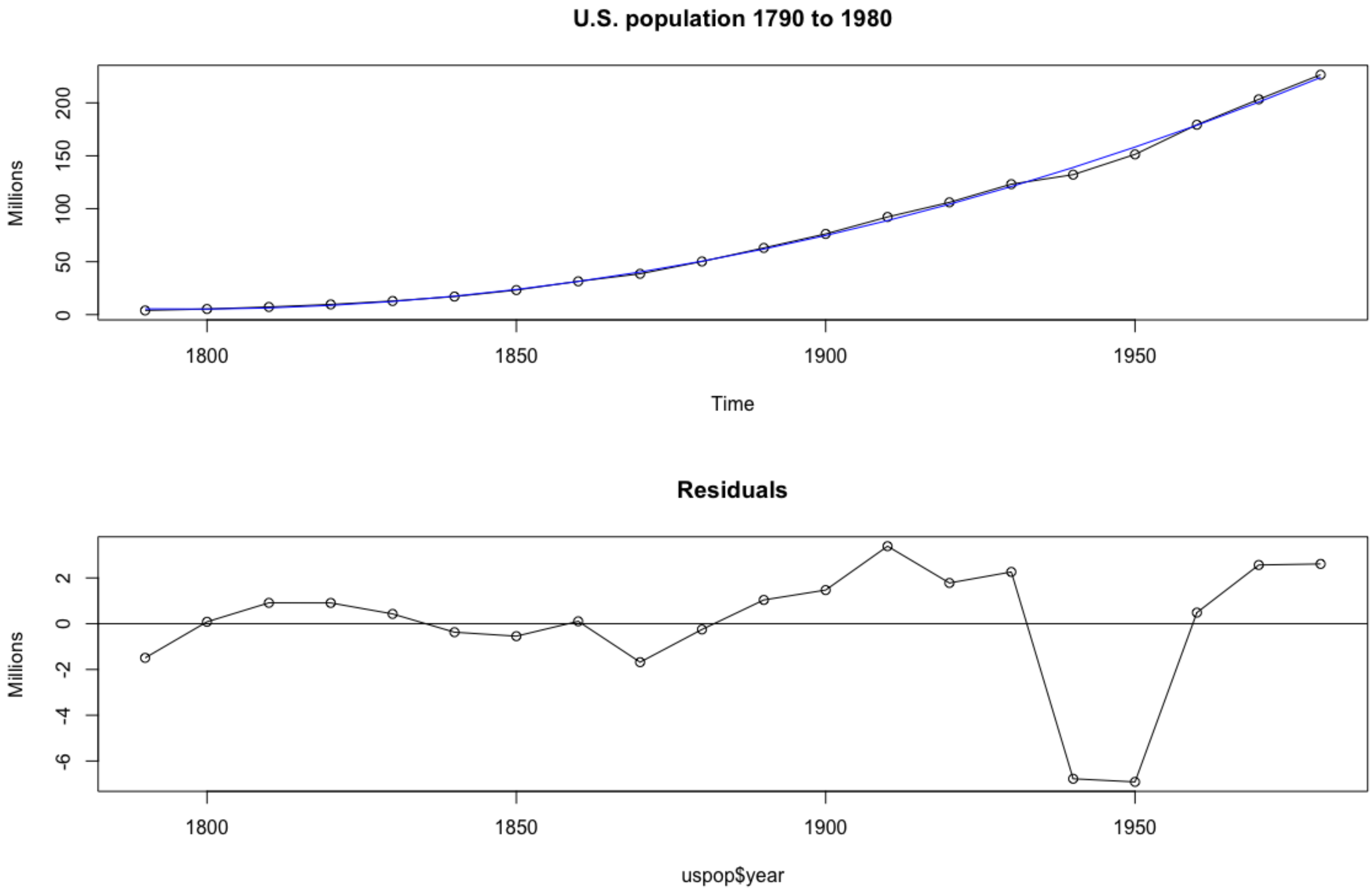
```
> uspop=read.table("us_pop.txt",header=T)
> head(uspop)
  year population
1 1790    3929214
2 1800    5308483
3 1810    7239881

> uspop.lm=lm(population~year+I(year^2),data=uspop)
> summary(uspop.lm)
Coefficients:
              Estimate Std. Error t value Pr(>\mid t\mid )
(Intercept)  2.098e+10  7.629e+08   27.50 1.56e-15 ***
year        -2.335e+07  8.100e+05  -28.83 7.11e-16 ***
I(year^2)    6.499e+03  2.148e+02   30.25 3.18e-16 ***
```

Estimates of parameters are

$$\hat{\beta}_0 = 2.098 \times 10^{10}, \quad \hat{\beta}_1 = -2.335 \times 10^{7}, \quad \hat{\beta}_2 = 6.499 \times 10^{3}.$$

Example: U.S. Population



Deterministic seasonal component

$$x_t = m_t + s_t + e_t$$

- The seasonal component s_t is a deterministic function of period d, i.e. $s_t = s_{t+d}$.
- Usually there is an intercept term in the trend component m_t. Two ways to impose constraints for identifiability.
  - First constraint: $s_1 + \cdots + s_d = 0$.
  - Second constraint: one of the seasonal term is zero, say $s_1 = 0$.

Deterministic seasonal component

For example, consider a quarterly data with a linear trend. Let Q_{it} be the indicator function for the i-th quarter, i.e. $Q_{it} = 1$ if t corresponds to the i-th quarter, and $Q_{it} = 0$ otherwise.

overall mean

- First way: $x_t = m_t + \gamma_1 Q_{1t} + \gamma_2 Q_{2t} + \gamma_3 Q_{3t} + \gamma_4 Q_{4t} + e_t$, where $\gamma_1 + \gamma_2 + \gamma_3 + \gamma_4 = 0$, and e_t is stationary.
- Second way: $x_t = m_t + \gamma_2 Q_{2t} + \gamma_3 Q_{3t} + \gamma_4 Q_{4t} + e_t$, where we have assumed $\gamma_1 = 0$.

1st quarter as baseline.

Elimination of both trend and seasonality: least squares

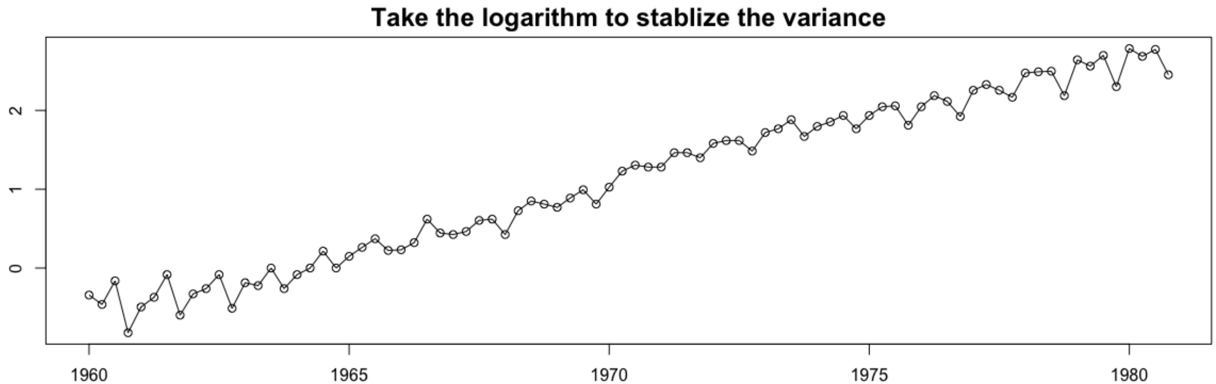
- For a quarterly data, consider the model

$$x_t = \underbrace{\beta_0 + \beta_1 t + \cdots + \beta_k t^k}_{m_t} + \underbrace{\gamma_2 Q_{2t} + \gamma_3 Q_{3t} + \gamma_4 Q_{4t}}_{s_t} + e_t.$$

- Parameters can be estimated by least squares.
- Need to perform variable selection to determine the order k, as well as identifying the seasonal components.

Example: J&J quarterly earnings

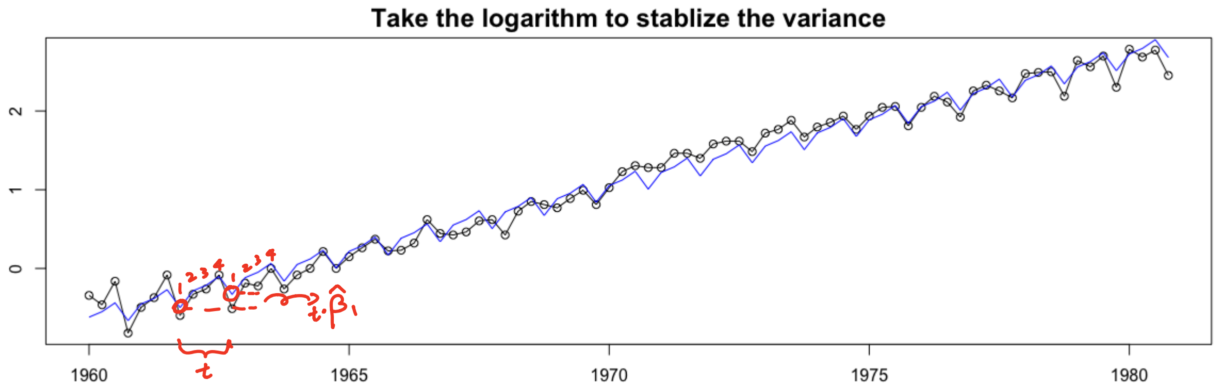
Consider the quarterly earnings of J&J from 1960 to 1980. We first take the logarithm to linearize the trend and stabilize the variance.

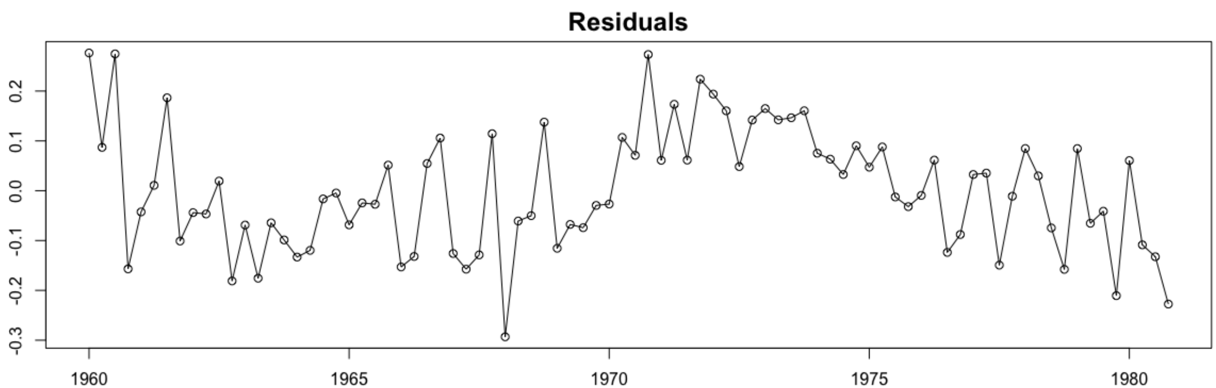


Example: J&J quarterly earnings

Fit a model with a linear trend

$$x_t = \beta_0 + \beta_1 t + \gamma_2 Q_{2t} + \gamma_3 Q_{3t} + \gamma_4 Q_{4t} + e_t.$$





Example: J&J quarterly earnings

$$x_t = \underbrace{\beta_0 + \beta_1 t}_{m_t} + \underbrace{\gamma_2 Q_{2t} + \gamma_3 Q_{3t} + \gamma_4 Q_{4t}}_{s_t} + e_t.$$

```
log.jj=log(jj) ## log transform of the J&J earnings
ss=as.factor(rep(1:4,n/4))
tt=1:n
jj.lm=lm(log.jj~tt+ss)
```

log earnings in 1971 vs 1970:
same quarter: $\approx 4 \cdot \hat{\beta}_1$
(assuming t is
of 1-increments).

The estimates are give by

$$\hat{\beta}_0 = -.6607, \qquad \hat{\beta}_1 = .0418,$$

$$\hat{\gamma}_2 = .0281, \qquad \hat{\gamma}_3 = .098, \qquad \hat{\gamma}_4 = -.1705.$$

How to interpret these estimates?

Residual analysis

Suppose we fit regression model to the time series

$$y_t = \mu_t + e_t,$$

where e.g. $\mu_t = \beta_0 + \beta_1 t$ or other deterministic trend. Properties of the regression output depend heavily on the usual assumption that the unobserved stochastic component $\{e_t\}$ is white noise, and some depend on the further assumption that is approximately normally distributed.

Residual analysis

$$y_t = \mu_t + e_t,$$

- **Estimated trend**: e.g. $\hat{\mu}_t = \hat{\beta}_0 + \hat{\beta}_1 t$.
- **Residual**: $\hat{e}_t = y_t - \hat{\mu}_t$
- **Residual standard deviation**:

$$s = \sqrt{\frac{1}{n-p}\sum_{t=1}^{n}(y_t - \underline{\hat{\mu}_t})^2}$$

$\rightsquigarrow \hat{y}_t$

where p is the number of parameters used to estimate $\hat{\mu}_t$, and $n - p$ is the residual degrees of freedom.

- Concepts such as *TSS*, *RSS* and R^2 defined analogously.

Residual analysis

Residual analysis of fitted time series model utilizes the following items:

- Sample autocorrelation function (ACF), given by

$$\hat{\rho}_k = r_k = \frac{\sum_{t=k+1}^{n}(y_t - \bar{y})(y_{t-k} - \bar{y})}{\sum_{t=1}^{n}(y_t - \bar{y})^2}.$$

  Result. Let r_k denote the sample ACF of white noise obtained from n observations. Then

$$r_k \overset{approx}{\sim} N(0, 1/n) \;\text{ for }\; k > 0 \;\text{ and large }\; n.$$

  Hence, 95% confidence interval is $\pm 1.96/\sqrt{n}$.

- Normality: Shapiro-Wilk test and QQ plot;
- Independence: runs test.

See R markdown for illustration.