

Stat 544: Time Series Analysis

The University of Montana

Chapter 4: Exploratory Data Analysis

Fall 2011

Main Objective of the Chapter

- ▶ In time series, it is important to measure the dependence between the values of the series.
- ▶ That is, we must be able to estimate the autocorrelation with precision.
- ▶ It would be difficult to measure (estimate) the dependence if the dependence structure is not regular or is changing at every time point.
- ▶ To achieve a meaningful statistical analysis of time series data, it will be crucial that the mean and the autocovariance functions to satisfy the conditions of stationarity.
- ▶ In the following, we study methods that can be used to down play the effects of Non-stationarity.

Trend Stationary Series

- ▶ This type of model may be written as

$$X_t = \mu_t + Y_t$$

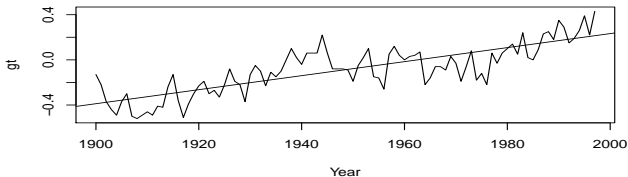
where μ_t denotes the trend and $\{Y_t\}$ is stationary process.

- ▶ Suppose the trend can be estimated using the technique studied in the previous chapter.
- ▶ Then we compute the residual (de-trended series)

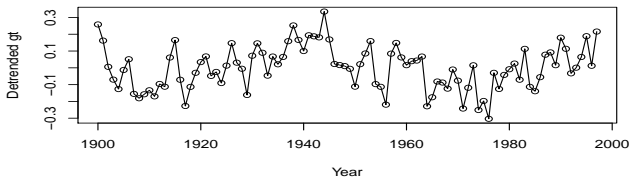
$$\hat{Y}_t = X_t - \hat{\mu}_t$$

Example: Global Warming

Average Temperature Deviation in centigrade



Detrended Global Temperature



Example: R-Code (Global Warming)

```
gt=scan("globtemp.dat") #scan can be replaced by read.table
X=gt[45:142] #use only 1900-1997
t=1900:1997
fit=lm(X~t)
gt=ts(X,start=1900,frequency=1) #formats the data as a time
par(mfrow=c(2,1))
plot(gt,main="Average
Temperature Deviation in centigrade", xlab="Year")
abline(fit)
plot(t,fit$resid,type="o",main="Detrended Global
Temprature",ylab="Detrended gt",xlab="Year")
```

Stochastic Trend

- ▶ Notice in the above example that de-trending has not removed the trend satisfactorily.
- ▶ It is possible to model a trend as a stochastic component.

$$X_t = \mu_t + Y_t$$
$$\mu_t = \delta + \mu_{t-1} + W_t.$$

- ▶ $\{X_t - X_{t-1}\}$ is a stationary series. How?

Definitions

- ▶ Difference Operator

$$\nabla X_t = X_t - X_{t-1}.$$

- ▶ Backshift Operator

$$BX_t = X_{t-1}.$$

- ▶ Powers of B

$$B^2 X_t = X_{t-2}, \dots, B^d X_t = X_{t-d}.$$

- ▶ It can be seen that

$$\nabla X_t = X_t - X_{t-1} = X_t - BX_t = (1 - B)X_t.$$

- ▶ Powers of ∇

$$\nabla^2 X_t = (1-B)^2 X_t = (1-2B+B^2)X_t = X_t - 2X_{t-1} + X_{t-2}. \quad \text{Why?}$$

Definitions Cont'd...

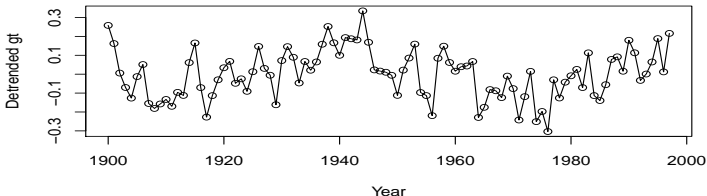
- ▶ Differences of order d are defined as

$$\nabla^d = (1 - B)^d.$$

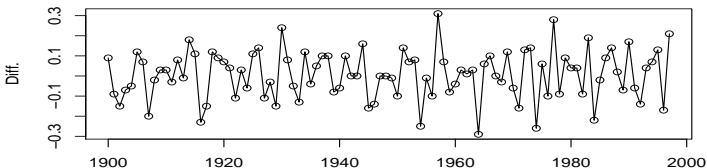
- ▶ We may expand the operator $(1 - B)^d$ algebraically to evaluate for higher integer value d .
- ▶ ∇^d removes a polynomial trend of degree d .
- ▶ For example, $X_t = a + bt + ct^2 + Y_t$ where $\{Y_t\}$ is a stationary process. Compute $\nabla^2 X_t$.
- ▶ Differencing can also remove stochastic trends.

Example: Global Temperature

Detrended Global Temperature



Differenced Global Temperature



Differencing versus Detrending

- ▶ Differencing involves no parameter estimation.
- ▶ If an estimate of Y_t is desired, differencing will not quite do the job for us because if, for example,

$$X_t = \beta_0 + \beta_1 t + Y_t$$

then

$$\nabla X_t = X_t - X_{t-1} = \beta_1 + Y_t - Y_{t-1}$$

- ▶ That is, estimate of $\beta_1 + Y_t - Y_{t-1}$ is obtained (not of Y_t).
- ▶ Bottom Line: If the goal is to obtain estimate of Y_t then use curve fitting. But if the goal is to coerce the data to stationarity then use differencing.
- ▶ Differencing removes both fixed and stochastic trends.

The Classical Decomposition Model

- ▶ Classical Decomposition Model:

$$X_t = M_t + S_t + Y_t$$

where M_t is global trend and S_t seasonal trend.

- ▶ If the seasonal period (in number of data points) is d , then $S_t = S_{t-d}$.
- ▶ We can use dummy variables or harmonic components to model S_t (as seen in the previous chapter).
- ▶ We can also eliminate the periodic trends by differencing.

The lag- d Difference Operator

- ▶ The lag- d differencing operator

$$\nabla_d X_t = X_t - X_{t-d} = (1 - B^d)X_t.$$

- ▶ This operator should not be confused with $\nabla^d = (1 - B)^d$.
- ▶ This removes both fixed as well as stochastic periodic components.
- ▶ Example: Let $\{Y_t\}$ be a stationary process with mean 0 and let a and b be constants.
 1. If $X_t = (a + bt) + S_t + Y_t$ where S_t is a seasonal component of period 12, show that $\{\nabla \nabla_{12} X_t\}$ is a stationary process and express its autocovariance function in terms of that of $\{Y_t\}$.
 2. If $X_t = (a + bt)S_t + Y_t$ where S_t is a seasonal component of period 12, show that $\{\nabla_{12}^2 X_t\}$ is a stationary process and express its autocovariance function in terms of that of $\{Y_t\}$.

ACF for Non-Stationary Time Series

- ▶ For data containing a trend, $\rho(h)$ will exhibit slow decay as h increases.
- ▶ For a data with a substantial deterministic periodic component, $\rho(h)$ will exhibit similar behavior with the same periodicity.
- ▶ Thus, $\rho(\cdot)$ can be useful as an indicator of nonstationarity.

Example: ACF Global Temperature Original, Detrended and Differenced

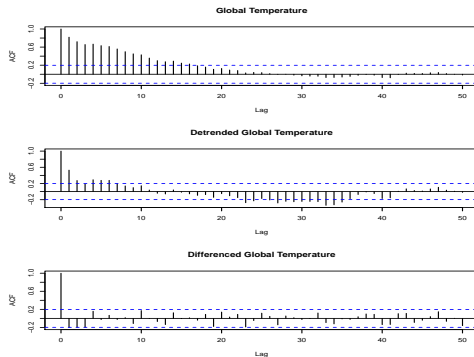


Figure: The last one was plotted by `acf(diff(Y),50,...)`

Main Objectives

- ▶ Often aberrations are present that can contribute to nonstationarity and nonlinearity.
- ▶ In such cases transformations may be useful to stabilize variance, achieve linearity and achieve normality.
- ▶ A family of power transformation defined by

$$X_t^{(\lambda)} = \begin{cases} \frac{X_t^\lambda - 1}{\lambda} & \lambda \neq 0 \\ \ln X_t & \lambda = 0 \end{cases}$$

- ▶ The transformation

$$Y_t = \ln X_t$$

tends to suppress larger fluctuation over portions of the series where the underlying values are large.

- ▶ Choice of λ depends on the purpose of the transformation.

Box-Cox Transformation: Achieving Near Normality

- ▶ To achieve (near) normality, choose λ that maximizes

$$\ell(\lambda) = -\frac{n}{2} \ln \left[\frac{1}{n} \sum_{j=1}^n (X_j^{(\lambda)} - \overline{X^{(\lambda)}})^2 \right] + (\lambda - 1) \sum_{j=1}^n \ln X_j$$

$$\text{where } \overline{X^{(\lambda)}} = \frac{1}{n} \sum_{j=1}^n X_j^{(\lambda)}.$$

- ▶ The calculation of $\ell(\lambda)$ is easy task for a computer.
- ▶ Graph $\ell(\lambda)$ versus λ as well as make a tabular display of $(\lambda, \ell(\lambda))$ in order to study the behavior of the maximizing value $\hat{\lambda}$.
- ▶ For instance, if $\hat{\lambda} \approx 0$ or $\hat{\lambda} \approx \frac{1}{2}$, log or $\sqrt{\cdot}$ is preferable.

Example: J & J Quarterly Earnings per Share

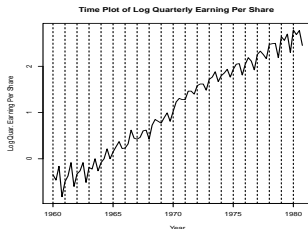


Figure: Notice that variance has been fairly stabilized and linearity has been achieved. The seasonal fluctuation in the data is preserved.

Example: Paleoclimatic Glacial Varves

- ▶ Plot of the objective function

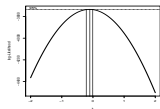


Figure: Profile Likelihood plot for the Box Cox Transformation of Glacial Varve Thickness Data with 95% Confidence Interval for λ .

- ▶ R-code

```
library(MASS)
varve=scan("varve.dat")
boxcox(varve~rep(1,length(varve)))
```

- ▶ The function is written in the regression context. However, it is tricked to work in the single variable context.

Example: Paleoclimatic Glacial Varves Cont'd...

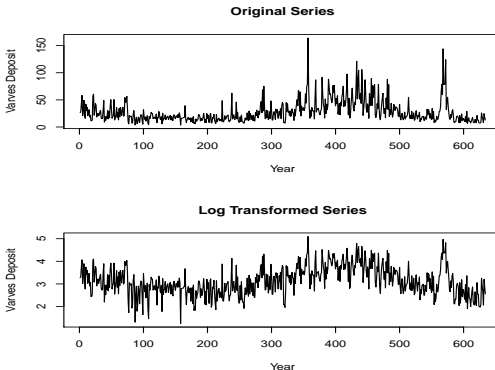
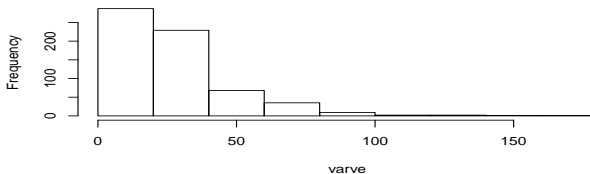


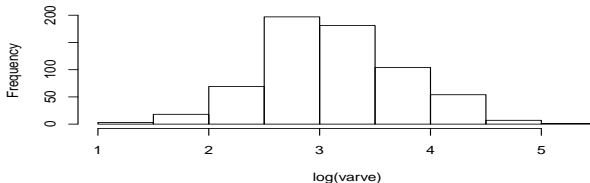
Figure: Glacial Varve Thickness ($n=634$) and Log Transformed Series

Histograms of Paleoclimatic Glacial Varves

Histogram of varve



Histogram of log(varve)



Homework

- ▶ The Monthly Federal Reserve Board Production Index for 372 months is contained in the R external data object `tsa3.rda` under the name `prodn`. Remove the trend and season components using trend estimation method and differencing. How do the two methods compare for this data set? Use time plot and acf plot to determine the order of differencing needed.
- ▶ Shumway and Stoffer, Problem 2.8. You don't have to do part of (e) which asks you to prove a formula for the autocovariance function.

Frequency of Oscillation Known?

- ▶ For example: The Johnson and Johnson quarterly earnings and the weekly cardiovascular mortality in LA make one cycle every year.
- ▶ In this case, the problem reduces to a regression analysis.
- ▶ Assume the data was generated by the signal in noise model:

$$X_t = A \cos(2\pi\omega t + \phi) + W_t$$

where $W_t \sim WN(0, \sigma^2)$

- ▶ A is amplitude, ω is frequency, ϕ is phase angle
- ▶ ω is assumed known.
- ▶ For example for Johnson and Johnson $\omega = 1/4$
- ▶ A and ϕ are parameters to be estimated from the data.

Frequency of Oscillation Known Cont'd...

- ▶ The use of a trigonometric identity leads to

$$\begin{aligned} X_t &= A \cos(\phi) \cos(2\pi\omega t) - A \sin(\phi) \sin(2\pi\omega t) + W_t \\ &= \beta_1 Z_{1t} + \beta_2 Z_{2t} + W_t \end{aligned}$$

where $\beta_1 = A \cos(\phi)$, $\beta_2 = -A \sin(\phi)$, $Z_{1t} = \cos(2\pi\omega t)$ and $Z_{2t} = \sin(2\pi\omega t)$

- ▶ A regression of this form is known as Harmonic Regression.
- ▶ Intercept term can be included if necessary or the data has to be corrected for its mean if the mean is not zero.
- ▶ OLS estimators of β_1 and β_2 (A and ϕ) are

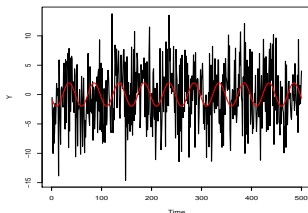
$$\hat{\beta}_1 = \frac{2}{n} \sum_{t=1}^n X_t \cos(2\pi\omega t) \quad \text{and} \quad \hat{\beta}_2 = \frac{2}{n} \sum_{t=1}^n X_t \sin(2\pi\omega t)$$

An Example: Signal in Noise

- ▶ Data generated by,

$$X_t = 2 \cos(2\pi(1/50)t + 0.6\pi) + W_t$$

$\{W_t\}$ is normal white noise with $\sigma^2 = 25$.



- ▶ The signal is obscured by the noise. Can we detect the signal by using regression?

Example Cont'd...

- ▶ The fitted equation is:

$$\hat{X}_t = -0.7933_{(0.3149)}Z_{1t} - 1.6153_{(0.3149)}Z_{2t}$$

and $\hat{\sigma}^2 = 24.79044$.

- ▶ Notice that $\beta_1 = 2 \cos(0.6\pi) = -0.62$ and $\beta_2 = -2 \sin(0.6\pi) = -1.9$.
- ▶ We have been able to detect the signal quite well. Why?
- ▶ Sample size is very large $n = 500$ and σ_W^2 is small.
- ▶ We can obtain estimates of A and ϕ by solving

$$-0.7933 = A \cos(\phi) \quad \text{and} \quad -1.6153 = -A \sin(\phi).$$

R-Code for the Example

```
t=1:500
omega=1/50
C=2*cos(2*pi*omega*t + 0.6*pi)
W=rnorm(500,0,1)
Y=C+5*W
plot.ts(Y)
lines(t,C,col="red" )
Z_1=cos(2*pi*omega*t)
Z_2=sin(2*pi*omega*t)
fit=lm(Y~0+Z_1+Z_2)
summary(fit)
```

The Periodogram

- ▶ Consider the saturated model (assuming n to be even),

$$X_t = \sum_{j=0}^{n/2} [\beta_1(j/n) \cos(2\pi(j/n)t) + \beta_2(j/n) \sin(2\pi(j/n)t)].$$

- ▶ We do not need W_t because the fit is perfect. How?
- ▶ Predictors are orthogonal, i.e.

$$\sum_{t=1}^n Z_{it}Z_{jt} = 0$$

for $i \neq j$. And $\sum_{t=1}^n Z_{it}^2 = n/2$.

The Periodogram Cont'd...

- ▶ Hence the estimates of the parameters, for $j = 1, \dots, \frac{n}{2} - 1$, are

$$\hat{\beta}_1(j/n) = \frac{2}{n} \sum_{t=1}^n X_t \cos(2\pi(j/n)t) \quad \text{and}$$

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

- ▶ For $j = 0$ and $j = n/2$,

$$\hat{\beta}_1(0) = \frac{1}{n} \sum_{t=1}^n X_t, \quad \hat{\beta}_1(1/2) = \frac{1}{n} \sum_{t=1}^n (-1)^t X_t,$$

$$\hat{\beta}_2(0) = 0 \quad \text{and} \quad \hat{\beta}_2(1/2) = 0.$$

The Periodogram Cont'd...

- ▶ If n is odd, we drop the first observation.
- ▶ The regression coefficients $\hat{\beta}_1(j/n)$ and $\hat{\beta}_2(j/n)$, for each j , are essentially measuring the correlation of the data with a sinusoid oscillating at j cycles in n time points.
- ▶ An appropriate measure of the contribution of a frequency j cycle in n time points in the data would be

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

- ▶ We will call this quantity the *Scaled Periodogram*
- ▶ A computationally simple approach is the following.

The Periodogram Cont'd...

- ▶ Consider the following complex-valued weighted average of the data given by

$$d(j/n) = \frac{1}{\sqrt{n}} \sum_{t=1}^n X_t e^{-2\pi i(j/n)t}.$$

- ▶ This quantity is known as Discrete Fourier Transform (DFT).
- ▶ The values j/n are called the Fourier or fundamental frequencies.
- ▶ The DFT can be easily computed using the Fast Fourier Transform (FFT) which is available in R .

The Periodogram Cont'd...

- ▶ It can be shown that

$$|d(j/n)|^2 = \frac{n}{4} \left(\hat{\beta}_1^2(j/n) + \hat{\beta}_2^2(j/n) \right) \\ := I(j/n).$$

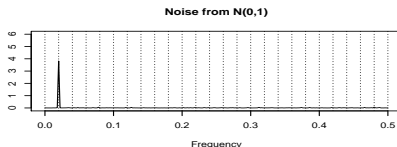
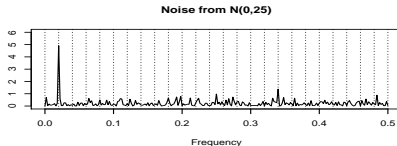
- ▶ $I(j/n)$ is called the Periodogram.
- ▶ Obviously, the scaled Periodogram is given by

$$P(j/n) = \frac{4}{n} I(j/n).$$

- ▶ Plot $P(j/n)$ against j/n to identify frequencies of oscillations.

Example: Signal in Noise

- Data generated by: $X_t = 2 \cos(2\pi(1/50)t + 0.6\pi) + W_t$
 where $\{W_t\}$ is normal white noise with $\sigma^2 = 25$.



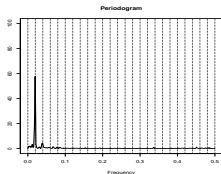
- The Periodograms say there is a dominant frequency $\omega = 0.02$ which is $1/50$.

R-Code for the Signal in Noise Example

```
# Periodogram Analysis
t=1:500
n=500
x=2*cos(2*pi*(1/50)*t+0.6*pi)+rnorm(500,0,5)
I=abs(fft(x)/sqrt(n))^2
P=(4/500)*I
f=0:250/500
plot(f,P[1:251],type="l",main="Periodogram",
     xlab="Frequency", ylab=" ",ylim=c(0,6))
abline(v=seq(0,.5,.02),lty="dotted")
```

Example: Weekly Cardiovascular Mortality in Los Angeles ($n = 508$ weeks)

- ▶ Periodogram is constructed after adjusting for trend.
- ▶ If there was no trend, we have to correct the series for its mean to get rid of an intercept term.

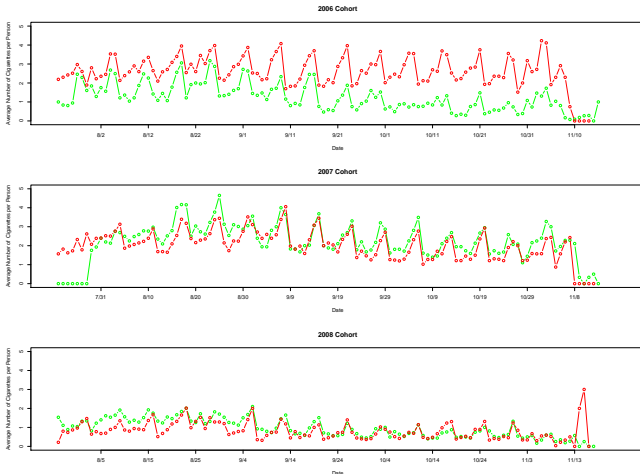


- ▶ The periodogram above shows that once a year cycle ($1.04 = 0.02 \times 52$).

R-Code for the Cardiovascular Example

```
# Periodogram Analysis of Cardio Vascular Mortality Data
mort=scan("cmort.dat")
n=length(mort)
t=1:n
fit=lm(mort~t)
x=fit$resid
I=abs(fft(x)/sqrt(n))^2
P=(4/n)*I
f=0:(n/2)/n
plot(f,P[1:(n/2+1)],type="l",main="Periodogram",
     xlab="Frequency", ylab=" ",ylim=c(0,100))
abline(v=seq(0,.5,.02),lty="dotted")
```

Homework: Smoking Patterns in Greek Societies at UMC



Homework Cont'd...

Shumway and Stoffer, Page 82, Problem 2.9

Purpose of Smoothing

- ▶ uses to bring out (discover)
 - ▶ long-term trend
 - ▶ seasonal (periodic) behavior
- ▶ General set up for a time plot is:

$$X_t = f_t + Y_t$$

where f_t is some smooth function of t and $\{Y_t\}$ is a stationary process.

Regression Smoothing

- ▶ f_t is the same function over the range of time
- ▶ Polynomial Regression of degree q

$$f_t = b_0 + b_1 t + b_2 t^2 + \cdots + b_q t^q$$

- ▶ Periodic (Harmonic) Regression

$$f_t = \alpha_0 + \alpha_1 \cos(2\pi\omega_1 t) + \beta_1 \sin(2\pi\omega_1 t) + \\ \cdots + \alpha_p \cos(2\pi\omega_p t) + \beta_p \sin(2\pi\omega_p t)$$

where $\omega_1, \omega_2, \dots, \omega_p$ are distinct known frequencies.

Regression Smoothing Cont'd...

- ▶ A combination

$$\begin{aligned} f_t = & b_0 + b_1 t + b_2 t^2 + \dots + b_q t^q \\ & + \alpha_1 \cos(2\pi\omega_1 t) + \beta_1 \sin(2\pi\omega_1 t) + \\ & \dots + \alpha_p \cos(2\pi\omega_p t) + \beta_p \sin(2\pi\omega_p t) \end{aligned}$$

Example: Weekly Cardiovascular Mortality in LA ($n = 508$)

Choosing $q = 3$, $p = 1$ and $\omega_1 = 1/52$, we get

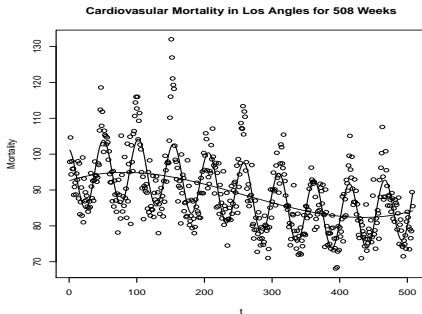


Figure: Polynomial Regression and Polynomial–Harmonic Regression
Global Smoothing

R-Codes

```
mort=scan("cmort.dat")
t=1:length(mort)
t2=t^2
t3=t^3
c=cos(2*pi*(1/52)*t)
s=sin(2*pi*(1/52)*t)
fit1=lm(mort~t+t2+t3)
fit2=lm(mort~t+t2+t3+c+s)
plot(t,mort, main="Cardiovasular Mortality in Los Angles fo
      ylab="Mortality")
lines(fit1$fit)
lines(fit2$fit)
```

Main Idea

- ▶ f_t is allowed to be different function over time
- ▶ The estimator \hat{f}_t has the general form

$$\hat{f}_t = \sum_{i=1}^n w_t(i) X_i.$$

where $w_t(i)$ are weights that reflect the contribution of X_i .

- ▶ Approaches to derive $w_t(i)$ are
 1. Moving-Average (Median) Smoothing
 2. Kernel Smoothing
 3. Nearest-Neighbor Regression
 4. Smoothing Splines
 5. Locally-Weighted Regression

Moving Averages Smoothing

- ▶ In general moving averages takes the form

$$\hat{f}_t = \sum_{j=-q}^q a_j X_{t-j}.$$

- ▶ The average of $(2q + 1)$ observations are used to obtain \hat{f}_t .
- ▶ This is a two sided $2q + 1$ points moving average
- ▶ If $a_j = a_{-j}$ then we have a symmetric moving average.
- ▶ Smoothness of \hat{f}_t increases as q increases
- ▶ Moving averages are *Linear Filters*.

Weekly Cardiovascular Mortality in Los Angeles

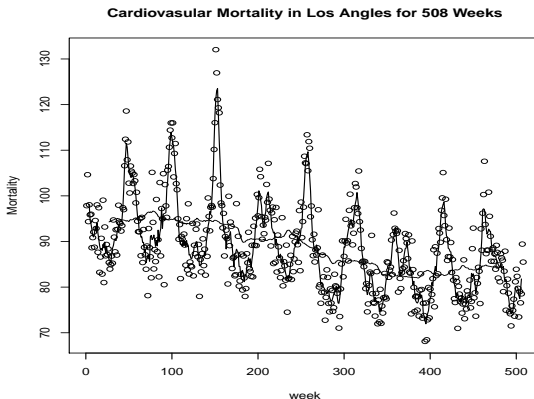


Figure: 5-points and 53-points Moving Averages

R-Codes

```
mort=scan("cmort.dat")
t=1:length(mort)
ma5=filter(mort,sides=2,rep(1,5)/5)
ma53=filter(mort,sides=2,rep(1,53)/53)
plot(t,mort, main="Cardiovascular Mortality in Los Angles for 508
Weeks", ylab="Mortality",xlab="week")
lines(ma5)
lines(ma53)
```

Kernel Smoothing

- ▶ The Nadaraya–Watson Estimator

$$\hat{f}_t = \sum_{i=1}^n w_t(i) X_i \quad \text{where} \quad w_t(i) = K\left(\frac{t-i}{b}\right) / \sum_{j=1}^n K\left(\frac{t-j}{b}\right),$$

the function $K(\cdot)$ is known as the kernel function and b is known as the bandwidth.

- ▶ b provides a scale for the relative weighting of nearby observations.
- ▶ The wider the bandwidth, the smoother the result.
- ▶ There are different choices for kernel functions.

Kernel Functions

1. Uniform Kernel

$$K(z) = \begin{cases} 1 & \text{if } |z| \leq 0.5 \\ 0 & \text{Otherwise} \end{cases}$$

2. Gaussian Kernel

$$K(z) = (2\pi)^{-1/2} e^{-\frac{1}{2}z^2}$$

3. Triangular Kernel

$$K(z) = \begin{cases} 1 - |z| & \text{if } |z| \leq 1 \\ 0 & \text{Otherwise} \end{cases}$$

4. Quadratic (Epanechnikov) Kernel

$$K(z) = 0.75(1 - z^2) \quad \text{if } |z| \leq 1.$$

Kernel Functions

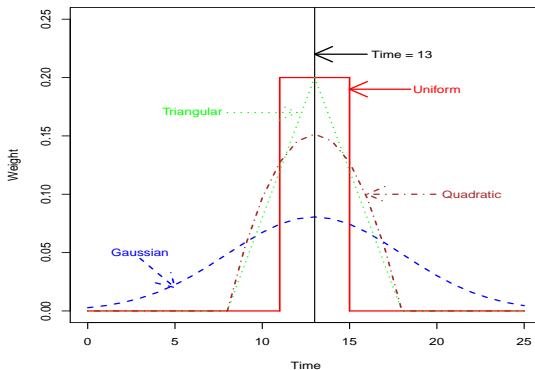


Figure: weights associated with the four kernel functions, centered at time point $t = 13$ and bandwidth $b = 5$

Example: Weekly # of Pneumonia and Influenza Deaths in the US (1999-2000)

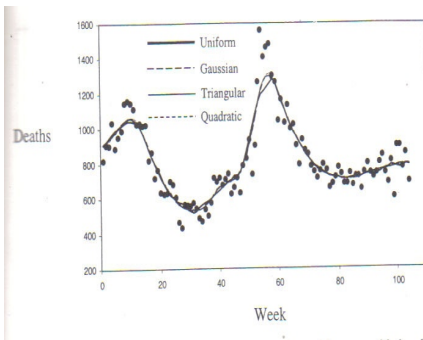


Figure: Smoothers produced using four Kernel functions with comparable bandwidths.

Selection of Kernel Function

- ▶ No rule of thumb exists.
- ▶ Theoretical arguments favor the quadratic kernel.
- ▶ In general, differences attributable to the form of kernel function are outweighed by the influence of the selection of bandwidth.

Selection of Bandwidth

- ▶ Represents implied trade-off between
 - (a) fidelity of the estimated model to the data
 - (b) reduction of random variance
- ▶ In other words, criteria for optimal bandwidth selection looks for optimum balance between bias and variance.
- ▶ Notice that minimization of bias implies smoothness of the fitted line and minimization of variance implies roughness of the fitted line.
- ▶ These are opposing minimization goals.
- ▶ What is more, the minimization must be based on observed data.

Example: Weekly Cardiovascular Mortality in Los Angeles

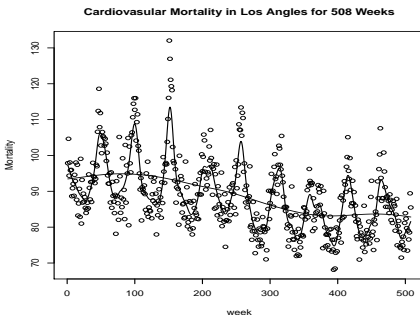


Figure: Kernel smoothers with bandwidths $b = 0.036$ and $b = 0.75$.

R-codes

```
plot(cmort,type="p",,ylab="Mortality")  
lines(ksmooth(time(cmort),mort,"normal",bandwidth=0.1))  
lines(ksmooth(time(cmort),mort,"normal",bandwidth=2))
```

Nearest Neighbor Regression Smoothing

- ▶ Based on k -nearest neighbors linear regression where one uses $\{X_{t-\frac{k}{2}}, \dots, X_t, \dots, X_{t+\frac{k}{2}}\}$ to predict X_t using linear regression.
- ▶ Example: Weekly Cardiovascular Mortality in Los Angeles

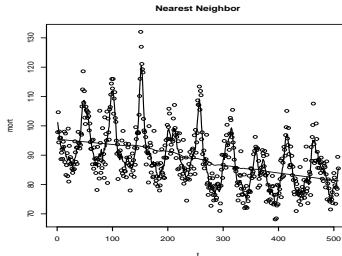


Figure: Nearest Neighbor Smoothing with $k = n/2$ and $k = n/100$

R-Codes

```
plot(t,mort,main="Nearest Neighbor")  
lines(supsmu(t,mort,span=0.5))  
lines(supsmu(t,mort,span=0.01))
```

Locally Weighted Regression Smoothing (Lowess)

- ▶ The basic idea is close to nearest neighbor regression.
- ▶ A certain portion of nearest neighbors to X_t are included in a weighting scheme.
- ▶ Values closer to X_t in time are given more weight.
- ▶ Then a robust weighted regression is used to to predict X_t and obtain a smoothed estimate \hat{f}_t of f_t .
- ▶ The larger the fraction of nearest neighbors included, the smoother \hat{f} is going to be.

Example: Weekly Cardiovascular Mortality in Los Angeles

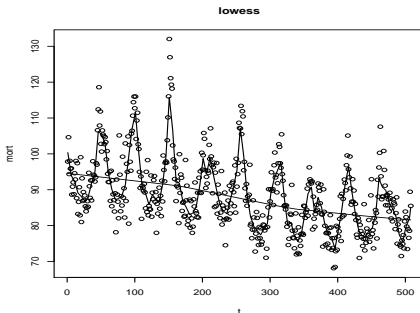


Figure: Lowess Smoothing using $2/3$ and 2% proportions

R-Codes

```
plot(t,mort, main="lowess")  
lines(lowess(t,mort,0.02))  
lines(lowess(t,mort,2/3))
```

Smoothing Splines

- ▶ Piecewise polynomial regression.
- ▶ Divide the time range into k -intervals as

$$[t_0 = 1, t_1], [t_1 + 1, t_2], \dots, [t_{k-1} + 1, n = t_k].$$

- ▶ t_0, t_1, \dots, t_k are called knots.
- ▶ In Each interval a polynomial regression of degree q is fitted.
- ▶ For example, when $q = 1$ we get what is known as *linear spline* and $q = 3$ we get *cubic spline*.

Smoothing Splines Cont'd...

- ▶ Choose \hat{f}_t that minimizes a compromise between the fit and the degree of smoothness given by

$$Q(\lambda) = \sum_{t=1}^n (X_t - f_t)^2 + \lambda \int (f_t'')^2 dt.$$

- ▶ The second term is a penalty term.
- ▶ This criterion trades off fidelity to the data (measured by the residual sum-of-squares) versus roughness of the function (measured by the penalty term).

Smoothing Splines Cont'd...

- ▶ $\lambda > 0$ represents a trade-off between closeness of fit to the data and smoothness of fit
- ▶ small λ implies a lot of roughness and large λ implies a relatively smooth f_t and allow less close fit to the data
- ▶ The minimizer for a given λ is a cubic spline.
- ▶ Silverman (1984) has shown that Smoothing Splines is asymptotically equivalent to kernel smoothing based on the kernel function

$$K(z) = 0.5 \exp(-\sqrt{2}z) \sin\left(\frac{|z|}{\sqrt{2}} + \frac{\pi}{4}\right).$$

- ▶ $GCV(\lambda)$ can be used to select optimal value of λ .

Example: Weekly Cardiovascular Mortality in Los Angeles

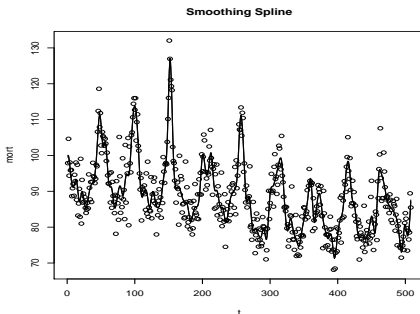


Figure: Smoothing Spline with $\lambda = 0.1$ and $\lambda =$ determined by GCV

R-Codes

```
plot(t,mort,main="Smoothing Spline")  
lines(smooth.spline(t,mort))  
lines(smooth.spline(t,mort,spar=0.1))
```

Homework

Shumway and Stoffer, 2.11(e) and 2.12