Prof. Michael Friger

Block 2
Time Series

“Whoever wishes to investigate medicine properly should proceed

thus: In the first place to consider the seasons of the year and what effect each of them produces“.

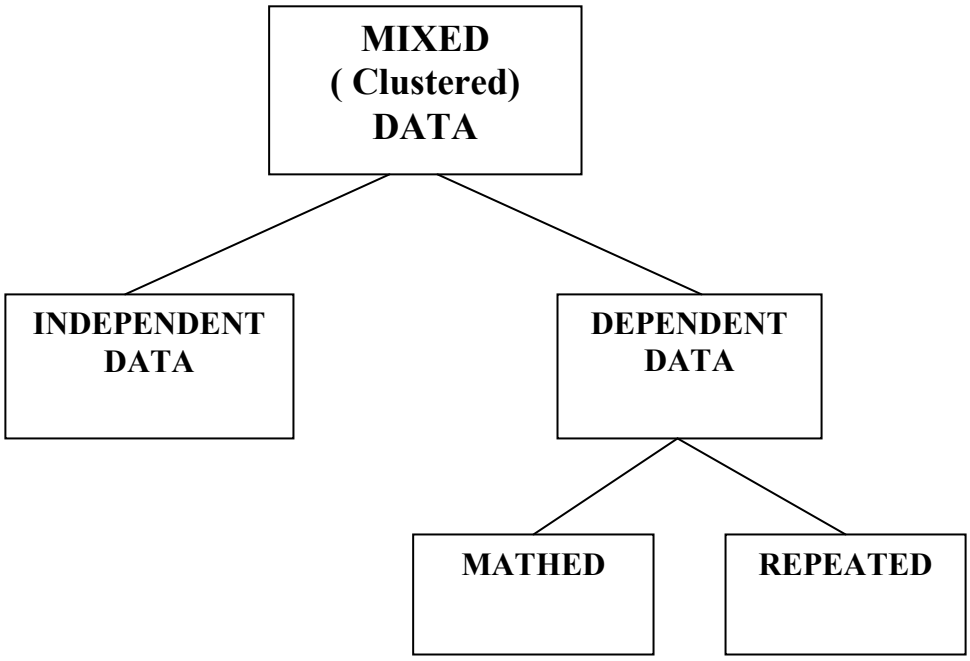
(Hippocrates, at least 400 BC)

Data structure

In matrix form:

$$\begin{array}{cccccc} \textit{case ID} & \text{var}1 & \text{var}2 & \ldots & \text{var}\,n \\ 1 & a_{11} & a_{12} & \ldots & a_{1n} \\ 2 & a_{21} & a_{22} & \ldots & a_{2n} \\ \ldots & \ldots & \ldots & \ldots & \ldots \\ k & a_{k1} & a_{k2} & \ldots & a_{kn} \end{array}$$

General types of data structure



Let we have model

$$M_{b_1,b_2,\ldots b_s}(x_1\, x_2,\ldots,x_s)$$

In general the estimation of model parameter b_i **= test of the following hypothesis:**

$$H_0: \quad b_i = 0$$

$$H_1: \quad b_i \neq 0$$

For this testing the so called Wald statistics:

$$Z = \frac{\hat{b}_i}{\sqrt{V\hat{a}r(\hat{b}_i)}}.$$

Problem of testing:
Conventional statistical approach based on Central limit Theorem.
Central limit Theorem requests random (independent) sample, because only in this case $\sigma^2_{\bar{X}} = \frac{\sigma^2_X}{n}$ **.**

<u>Note.</u> In general $Var_{ind} \geq Var_{dep}$ **,**

so

$$Z_{ind} \leq Z_{dep}$$

Longitudinal data =describes within-individual changes in variables over time.

What is *Time Series?*

Definition of Time Series: *An ordered sequence of values of a variable at equally spaced time intervals.*

Time Series is a set of data:

$$\{y_t : t = 1, 2, \ldots, n\}$$

where t indicates time at which the observation y_t was observed.

Other notation:

$$\{y(t_i) : t = 1, 2, \ldots, n\}$$

Important :

Time sequence is

$$\{t = 1, 2, \ldots, n\}$$

is homogeneous(Days, Months, Years - without jumps)

TABLE 1

Population of England and Wales at Ten-Yearly Intervals from 1811 to 1931.

(Data from the Registrar-General's *Statistical Review*, 1933, Part II.)

Year.	Population (millions).
1811	10·16
21	12·00
31	13·90
41	15·91
51	17·93
61	20·07
71	22·71
81	25·97
91	29·00
1901	32·53
11	36·07
21	37·89
31	39·95

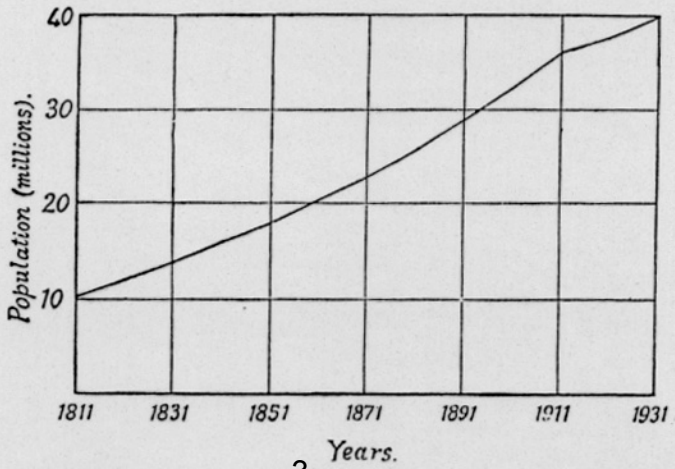


FIG. 1.—Graph of the Data of Table 1 (Population of England and Wales).

TABLE 2

Sheep Population of England and Wales for each Year from 1867 to 1939.

(Data from the *Agricultural Statistics.*)

Year.	Population (10,000).	Year.	Population (10,000).	Year.	Population (10,000).	Year.	Population (10,000).
1867	2203	1886	1892	1905	1823	1924	1484
68	2360	87	1919	06	1843	25	1597
69	2254	88	1853	07	1880	26	1686
70	2165	89	1868	08	1968	27	1707
71	2024	90	1991	09	2029	28	1640
72	2078	91	2111	10	1996	29	1611
73	2214	92	2119	11	1933	30	1632
74	2292	93	1991	12	1805	31	1775
75	2207	94	1859	13	1713	32	1850
76	2119	95	1856	14	1726	33	1809
77	2119	96	1924	15	1752	34	1653
78	2137	97	1892	16	1795	35	1648
79	2132	98	1916	17	1717	36	1665
80	1955	99	1968	18	1648	37	1627
81	1785	1900	1928	19	1512	38	1791
82	1747	01	1898	20	1338	39	1797
83	1818	02	1850	21	1383		
84	1909	03	1841	22	1344		
85	1958	04	1824	23	1384		

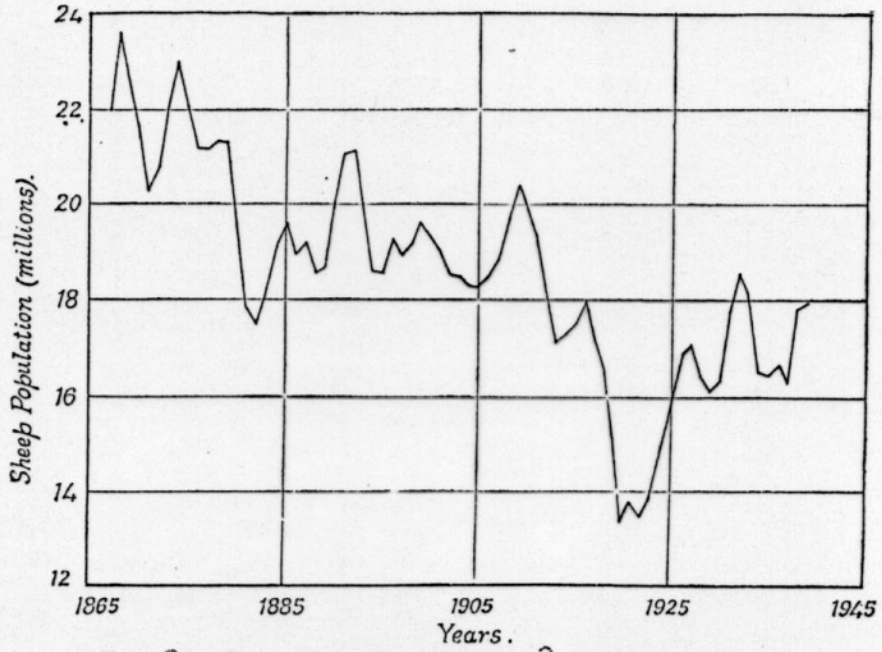


FIG. 2 —Graph of the Data of Table 2. (Sheep Population).

TABLE 3

Annual Yields per Acre of Barley in England and Wales from 1884 to 1939.

(Data from the *Agricultural Statistics.*)

Year.	Yield per acre (cwts.).	Year.	Yield per acre (cwts.).	Year.	Yield per acre (cwts.).	Year.	Yield per acre (cwts.).
1884	15·2	1898	16·9	1912	14·2	1926	16·0
85	16·9	99	16·4	13	15·8	27	16·4
86	15·3	1900	14·9	14	15·7	28	17·2
87	14·9	01	14·5	15	14·1	29	17·8
88	15·7	02	16·6	16	14·8	30	14·4
89	15·1	03	15·1	17	14·4	31	15·0
90	16·7	04	14·6	18	15·6	32	16·0
91	16·3	05	16·0	19	13·9	33	16·8
92	16·5	06	16·8	20	14·7	34	16·9
93	13·3	07	16·8	21	14·3	35	16·6
94	16·5	08	15·5	22	14·0	36	16·2
95	15·0	09	17·3	23	14·5	37	14·0
96	15·9	10	15·5	24	15·4	38	18·1
97	15·5	11	15·5	25	15·3	39	17·5

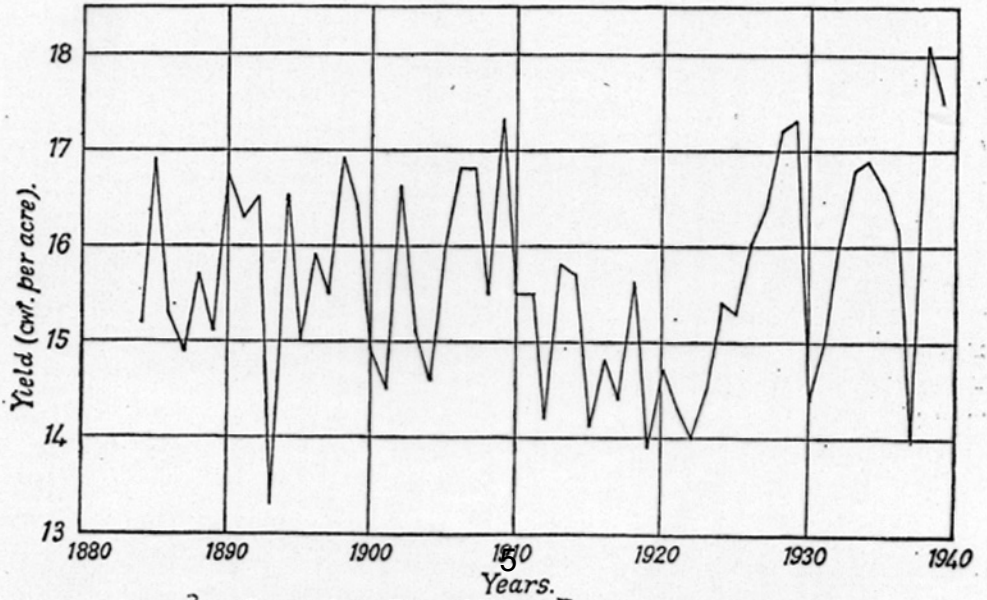


FIG. 3 —Graph of the Data of Table 3. (Barley Yields per Acre).

TABLE 4

Total Annual Rainfall at London in Inches, for each Year from 1813 to 1912.

(Data from D. Brunt, *Phil. Trans.* A, **225**, 247, 1925.)

Year.	Rainfall (inches).	Year.	Rainfall (inches).	Year.	Rainfall (inches).	Year.	Rainfall (inches).
1813	23·56	1838	21·63	1863	21·59	1888	27·74
14	26·07	39	27·49	64	16·93	89	23·85
15	21·86	40	19·43	65	29·48	90	21·23
16	31·24	41	31·13	66	31·60	91	28·15
17	23·65	42	23·09	67	26·25	92	22·61
18	23·88	43	25·85	68	23·40	93	19·80
19	26·41	44	22·65	69	25·42	94	27·94
20	22·67	45	22·75	70	21·32	95	21·47
21	31·69	46	26·36	71	25·02	96	23·52
22	23·86	47	17·70	72	38·86	97	22·86
23	24·11	48	29·81	73	22·67	98	17·69
24	32·43	49	22·93	74	18·82	99	22·54
25	23·26	50	19·22	75	28·44	1900	23·28
26	22·57	51	20·63	76	26·16	01	22·17
27	23·00	52	35·34	77	28·17	02	20·84
28	27·88	53	25·89	78	34·08	03	38·10
29	25·32	54	18·65	79	33·82	04	20·65
30	25·08	55	23·06	80	30·28	05	22·97
31	27·76	56	22·21	81	27·92	06	24·26
32	19·82	57	22·18	82	27·14	07	23·01
33	24·78	58	18·77	83	24·40	08	23·67
34	20·12	59	28·21	84	20·35	09	20·75
35	24·34	60	32·24	85	26·64	10	25·36
36	27·42	61	22·27	86	27·01	11	24·79
37	19·44	62	27·57	87	19·21	12	27·88

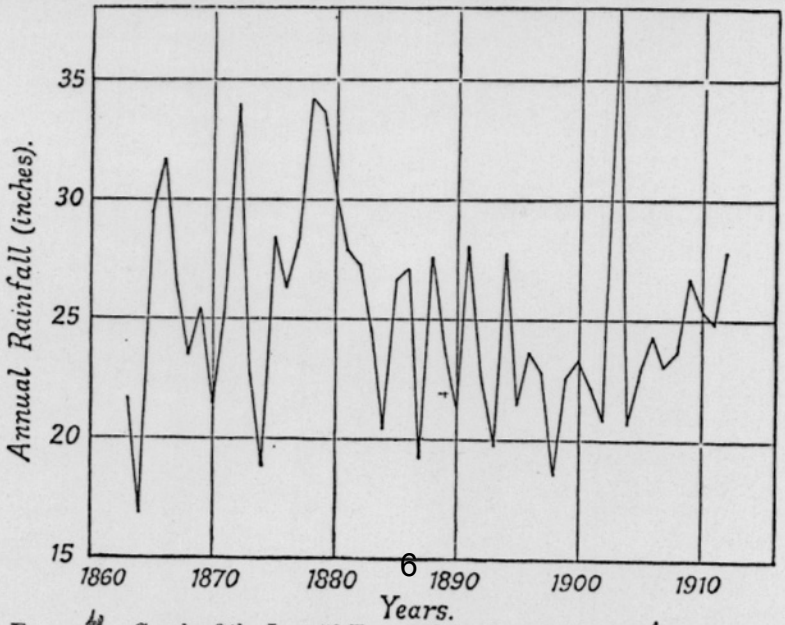


FIG. 4.—Graph of the Last 50 Terms of the Data of Table 4 (Rainfall).

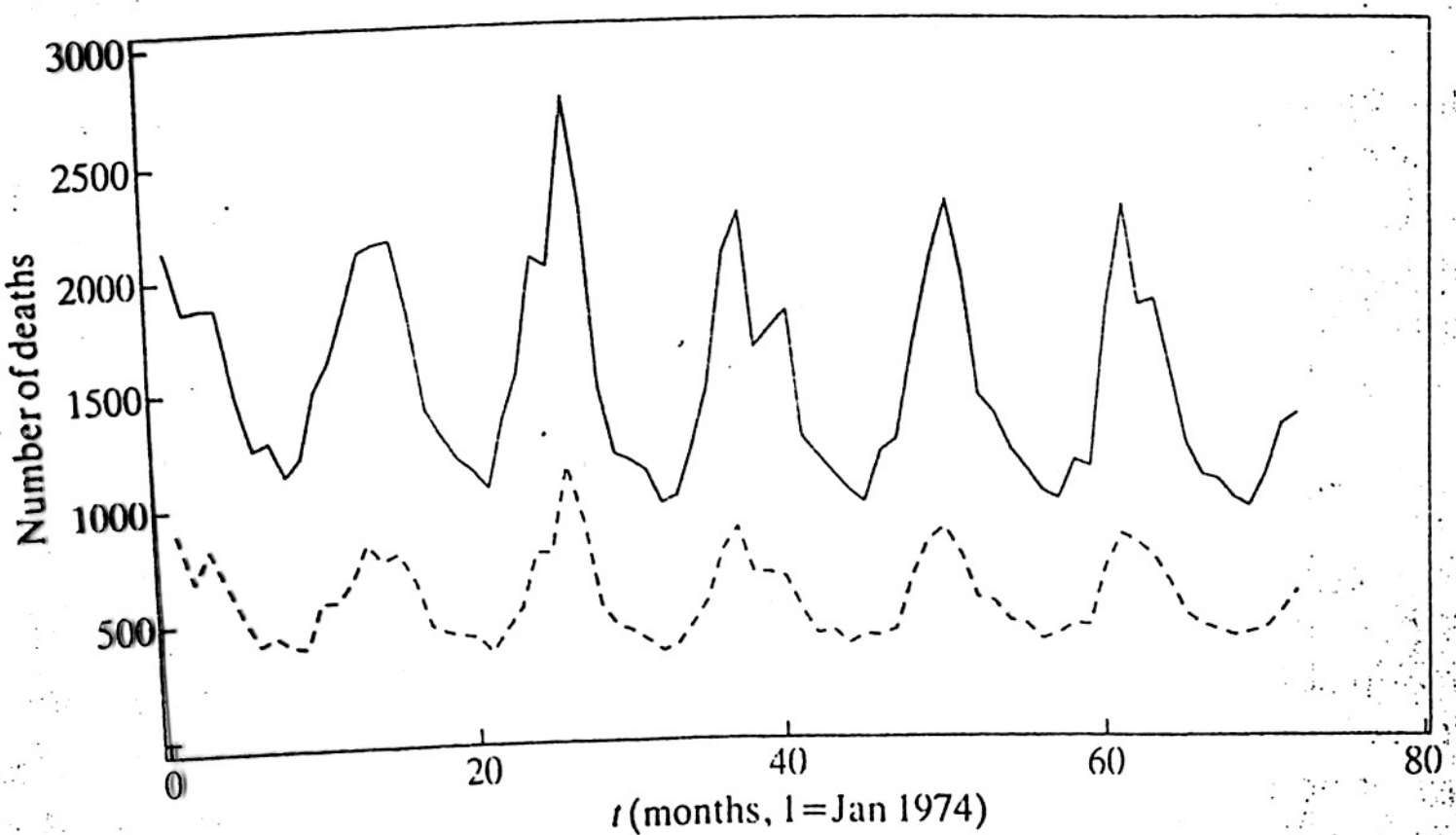


Fig. 5 Two time series of monthly returns of deaths in the United Kingdom attributed to bronchitis, emphysema, and asthma over the years 1974 to 1979. —, Males; – – –, females.

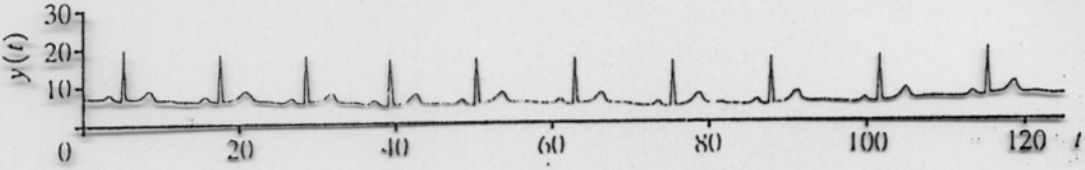


Fig. 6 An ECG trace from a healthy adult female. The inset shows an enlargement of a single heartbeat.

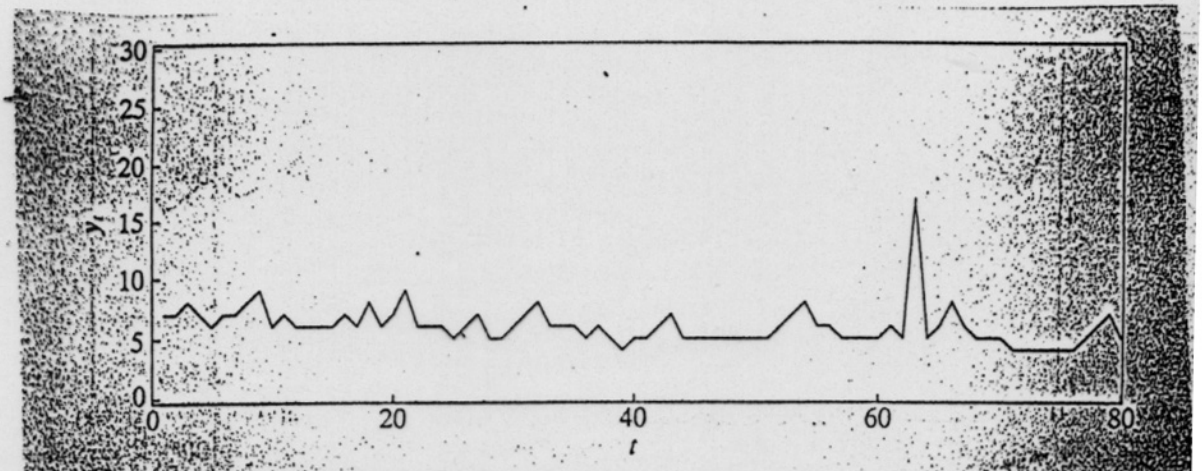


Fig. 7. A time series of 80 values obtained by sampling the intitial portion of the ECG trace in Fig. 6 at unit time intervals.

Main objectives

Two main aims:

a) Identifying and explanation of the phenomenon represented by the sequence observation as function of time and other independent variables;

b) Forecasting(predicting future values of the time series)

Time Series decomposition

Typical time series may be regard as composed of three parts:

a) A trend;

b) An oscillation about the trend of greater or less regularity;

c) "White noise"(a “random”, “irregular" or "unsystematic” component).

(M. Kendall, The advanced theory of statistics, v 2, 1948, p.369).

More detailed decomposition

a) Trends components (T_t);

b) Seasonal components(S_t);

c) Cycle components(C_t);

d) Quasi-cycle components(Q_t);

(e) "White noise"

(a “random”, “irregular" or "unsystematic”) component(R_t).

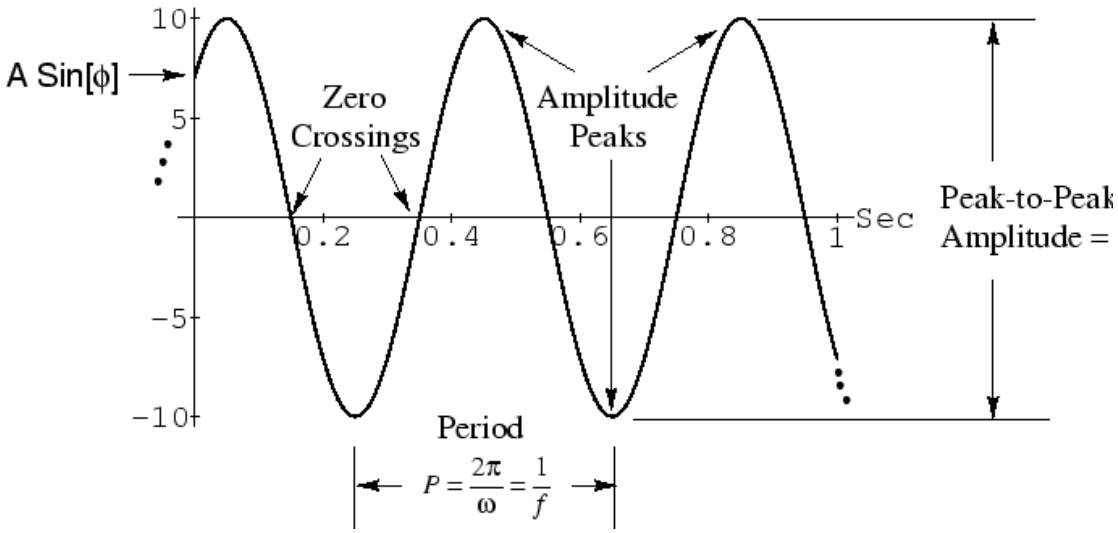
Trend Component

Trend component describes general tendency of given time series over long-term movement.

("Long" in this connection is a relative term)

Periodic oscillations

Defined by Period and Amplitude



Seasonal components

Associated with periodic oscillations with periods usually known a priori. The pattern may be fixed or slowly varying from season to season.

Cycle components

Periodic oscillation(waves) with stable amplitudes and periods (unknown a priori)

Quasi-cycle components

Periodic oscillation(waves) with fluctuating amplitudes and periods

White noise

Irregular statistical fluctuation and swings about the overall mean or trend, due to observational error, sampling variability or other reasons.

Model realization of Time Series decomposition

There are some straightforward possibilities for model realization of Time Series decomposition:

1. Additive

$$Y_t = T_t + S_t + C_t + Q_t + R_t$$

2. Multiplicative

$$Y_t = T_t * S_t * C_t * Q_t * R_t$$

3. Combined

$$Y_t = T_t * C_t + S_t + Q_t + R_t$$

(Sometimes the trend and cyclic component are customarily combined into a trend-cycle component).

There are exists other designs for realization.

Two time series models

$$Y = F(t)$$

$$Y = F(t, X_1(t), \cdots X_2(t), \ldots, X_n(t))$$

Let us consider the model:

$$Y = F(t) .$$

Identifying patterns in Time Series Data.

Trend

1. It is an essential part of concept of trend that the movement over fairly long periods is SMOOTH. This means that we could present the trend component, at least locally, by polynomial of time variable *t.*

(M. Kendall)

The consequences of this fact is:

2. We could present the trend component, at least locally, by polynomial of time variable *t.* Thus

$$T_t = a_0 + a_1 t + a_2 t^2 + \ldots + a_s t^s$$

3. In order to better understand behavior of trend we must use smoothing procedures.

In general, the model to represent trend is polynomial regression model:

$$T_t = a_0 + a_1 t + a_2 t^2 + \ldots + a_s t^s + \varepsilon \ ,$$

where $a_0, a_1, a_2, \ldots, a_s$ are regression coefficients, ε is a random error.

The trend could be classified by the following models:

1. Stationary Model

$$T_t = C + \varepsilon$$

where C is a constant, ε is an error.

2. Linear trend model.

$$T_t = C + A \cdot t + \varepsilon$$

3. Non linear trend model.

$$T_t = a_0 + a_1 t + a_2 t^2 + \ldots + a_s t^s + \varepsilon \ .$$

In practice.

For SPSS. Let Y be a time series variable,

t be a time variable

```
compute t2=t*t.
compute t3=t2*t
compute t4=t3*t
……
compute tn=t(n-1)*t
regression variables=Y
                    t t2 t3 t4  …tn/  statistics=end/
                  dependent=Y/ method=stepwise.
```

Example of Polynomial regression

$$Temp_t = a_0 + a_1 t + a_2 t^2 + \ldots + a_{30} t^{30} + \varepsilon$$

```
regression variables=temp
          t t2 t3 t4 t5 t6 t7 t8 t9 t10
           t11 t12 t13 t14 t15 t16 t17 t18 t19 t20
           t21 t22 t23 t24 t25 t26 t27 t28 t29 t30
          statistics=end
         dependent=temp method=stepwise
```

Step	Variable	MultR	Rsq
1	T	0.25	0.06
2	t2	0.70	0.49
3	t19	0.72	0.52
4	t3	0.74	0.55
5	t26	0.77	0.59

* * * * M U L T I P L E R E G R E S S I O N * * * *

Multiple R .76677
R Square .58794
Adjusted R Square .58539
Standard Error 4.19118
Analysis of Variance

	DF	Sum of Squares	Mean Square
Regression	4	16190.90575	4047.72644
Residual	646	11347.60946	17.56596

F = 230.43014 Signif F = .0000

---------- Variables in the Equation----------

Variable	B	SE B	Beta	Sig T
t2	-.002803	9.8838E-05	-6.791530	.0000
t3	1.45681E-05	5.4813E-07	7.643629	.0000
t19	-2.16038E-43	2.0332E-44	-3.311227	.0000
t26	5.26988E-60	6.1379E-61	2.262627	.0000
Constant	17.45	.351398		.0000

------ Variables not in the Equation----------

Variable	Beta In	Sig T
T	-.331299	.3230
T4	-.755457	.6452
T5	-.090677	.9151
T6	.095782	.8820
T7	.174662	.7620
T8	.211399	.7089
T9	.225415	.7034
T10	.223545	.7306
T11	.207455	.7794
T12	.175418	.8415
T13	.121877	.9101
T14	.034799	.9800
T15	-.111876	.9527
T16	-.381583	.8904
T17	-.969457	.8347
T18	-2.8650	.7861
T20	5.3842	.7100
T21	3.626413	.6816
T22	3.314459	.6588
T23	3.572223	.6409
T24	4.557832	.6273
T25	7.990577	.6173
T27	-6.471556	.6062
T28	-2.961289	. 6042
T29	-1.820692	.6040
T30	-1.267168	.6053

Differencing

Differencing provide possibility to remove trends in time-series data.

Differencing is converting each element of time series by into its difference from consequent element.
Let we have time series:

$$\{Y_t\},$$

then *the first difference* of this time series is time series

$$\{D(Y_t)\}, \text{ where } DY_t = Y_t - Y_{t-1} .$$

Example. $Y_t : 2, 5, 9, 10, 25 \quad DY_t = (5-2), (9-5), (10-9), (25-10) = 3, 4, 1, 15.$
Two-order difference

$$\{D^2(Y_t)\}$$

is defined by

$$D^2(Y_t) = D(D(Y_t)) = D(Y_t) - D(Y_{t-1}) = Y_t - 2Y_{t-1} + y_{t-2} .$$

Example. $Y_t : 2, 5, 9, 10, 25 \quad D^2 Y_t = (4-3), (1-4), (15-1) = 1, -3, \ 14.$

Let we take time series: $\{Y_t\}$, which defined by $Y_t = a + bt$, then
$DY_t = (a+bt) - (a+b(t-1)) = b.$ That is "new series" is stationary.
If $\{Y_t\}$ is defined by $Y_t = a + bt + ct^2$, then
$DY_t = (a+bt+ct^2) - (a+b(t-1)+c(t-1)^2) = b - 1 + 2ct.$ and $D\ Y_t = 2c.$
We obtain stationery by using *Two-order difference*.
By the similar way we can define
Two-order difference

$$\{D^k(Y_t)\}$$

is defined by

$$D^k(Y_t) = D(D^{k-1}(Y_t)) = D^{k-1}(Y_t) - D^{k-1}(Y_{t-1}).. $$

We can obtain the same result by using the polynomial regression.
Let $\{\hat{Y}_t\}$ is estimators of time series $\{Y_t\}$, then time series $\{\hat{Y}_t - Y_t\}$ is stationary.

Autocorrelation

One of tools to learn behavior of time series.

Standard Correlation Coefficient (Pearson)
For two variables X and Y:

$$\rho = \frac{\mathrm{cov}(X, Y)}{\sqrt{VarX \cdot VarY}}$$

where $\mathrm{cov}(X, Y) = \sum((X_i - \overline{X})(Y_i - \overline{Y}))$.

Autocorrelation:
Autocorrelation coefficient of order *k:*

$$\rho_k = \frac{\mathrm{cov}(y_t, y_{t+k})}{\sqrt{Var(y_t) \cdot Var(y_{t+k})}} = \frac{\mathrm{cov}(y_t, y_{t+k})}{\sqrt{Var(y_t) Var(y_t)}}$$

$$\rho_k == \frac{\mathrm{cov}(y_t, y_{t+k})}{Var(y_t)}$$

The order k is sometimes called *lag*.

Example.

Y_t :3, 4, 5, 3, 6, 3, 4, 4.

Order 1.

T	1	2	3	4	5	6	7	8	
Y_t	3	4	5	3	6	3	4	4	
	⇳	⇳	⇳	⇳	⇳	⇳	⇳		ρ_1= -0.67
Y_{t+1}	4	5	3	6	3	4	4		

Order 2.

T	1	2	3	4	5	6	7	8	
Y_t	3	4	5	3	6	3	4	4	
	⇳	⇳	⇳	⇳	⇳	⇳			ρ_2= 0.27
Y_{t+2}	5	3	6	3	4	4			

Example 2.

Time series of Daily mortality(without cancer mortality) in Philadelphia:

D_t: 44, 42, 53, 46, 49, 41, 55, 47, 47, 52, …

$\rho_1 = 0.37, \quad \rho_2 = 0.33, \quad \rho_4 = 0.31, \quad \rho_7 = 0.27$.

In practice,
For SPSS
For calculating k order autocorrelation for variable X
1. compute lag_k_X=lag(X, k).
2. correlations X lag_k_X.
Example
Compute lag_4_Dt=lag(D_t,4).
Correlations lag_4_Dt D_t.

Smoothing

החלקה

Main goals:

1. Understanding behavior and phenomena of trend.
2. Prediction of time series data.

To remove affects of random picks.

More generally, we use term smoothing to mean a decomposition of a time series Y_t into a "smooth" component S_t and a "rough" component r_t , so that

$$Y_t = S_t + r_t$$

Popular types of smoothing techniques:

1. Simple average.
   (ממוצע פשוט)
2. Moving average.
   (ממוצע נע)
3. Exponential smoothing.
   (החלקה אקספוננציאלית)
4. Polynomial regression.
   (רגרסיה פולינומיאלית)
5. Spline regression.
   (רגרסיה של ספליין)

Simple average

Let

$$Y_t : y_1, y_2, \ldots, y_n$$

be a time series.

$$F_s = \frac{1}{s-1}(y_{s-1} + y_{s-2} + \ldots y_2 + y_1) = \frac{1}{s-1}\sum_{i=1}^{s-1} y_i$$

Example.

Time series of Daily mortality(without cancer mortality) in Philadelphia:

D_t: 44, 42, 53, 46, 49, 41, 55, 47, 47, 52,…

$$F_3 = \frac{1}{3-1}(y_2 + y_1) = \frac{1}{2}(42 + 44) = 43$$

$$F_5 = \frac{1}{5-1}(y_4 + y_3 + y_2 + y_1) = \frac{1}{4}(46 + 53 + 42 + 44) = 46.25$$

Moving average

Moving average MA(k) is average of k closest observations.
Number k is called "window" of moving average.
Let

$$Y_t : y_1, y_2, \ldots, y_n$$

be a time series.

Step-forward moving average.

$$F_s = \frac{1}{k}(y_{s-1} + y_{s-2} + \ldots + y_{s-k}) = \frac{1}{k}\sum_{i=s-k}^{s-1} y_i$$

or

$$F_{s+1} = \frac{1}{k}(y_s + y_{s-1} + \ldots + y_{s-k+1}) = \frac{1}{k}\sum_{i=s-k+1}^{s} y_i$$

Example.

Time series of Daily mortality(without cancer mortality) in Philadelphia:

D_t: 44, 42, 53, 46, 49, 41, 55, 47, 47, 52,…

k=3.

$$F_6 = \frac{1}{3}(y_5 + y_4 + y_3) = \frac{1}{3}(49 + 46 + 53) = 49.3$$

k=4.

$$F_6 = \frac{1}{4}(y_5 + y_4 + y_3 + y_2) = \frac{1}{4}(49 + 46 + 53 + 42) =$$

$$= 47.5$$

Centered moving average.

Let $k = 2p + 1$.

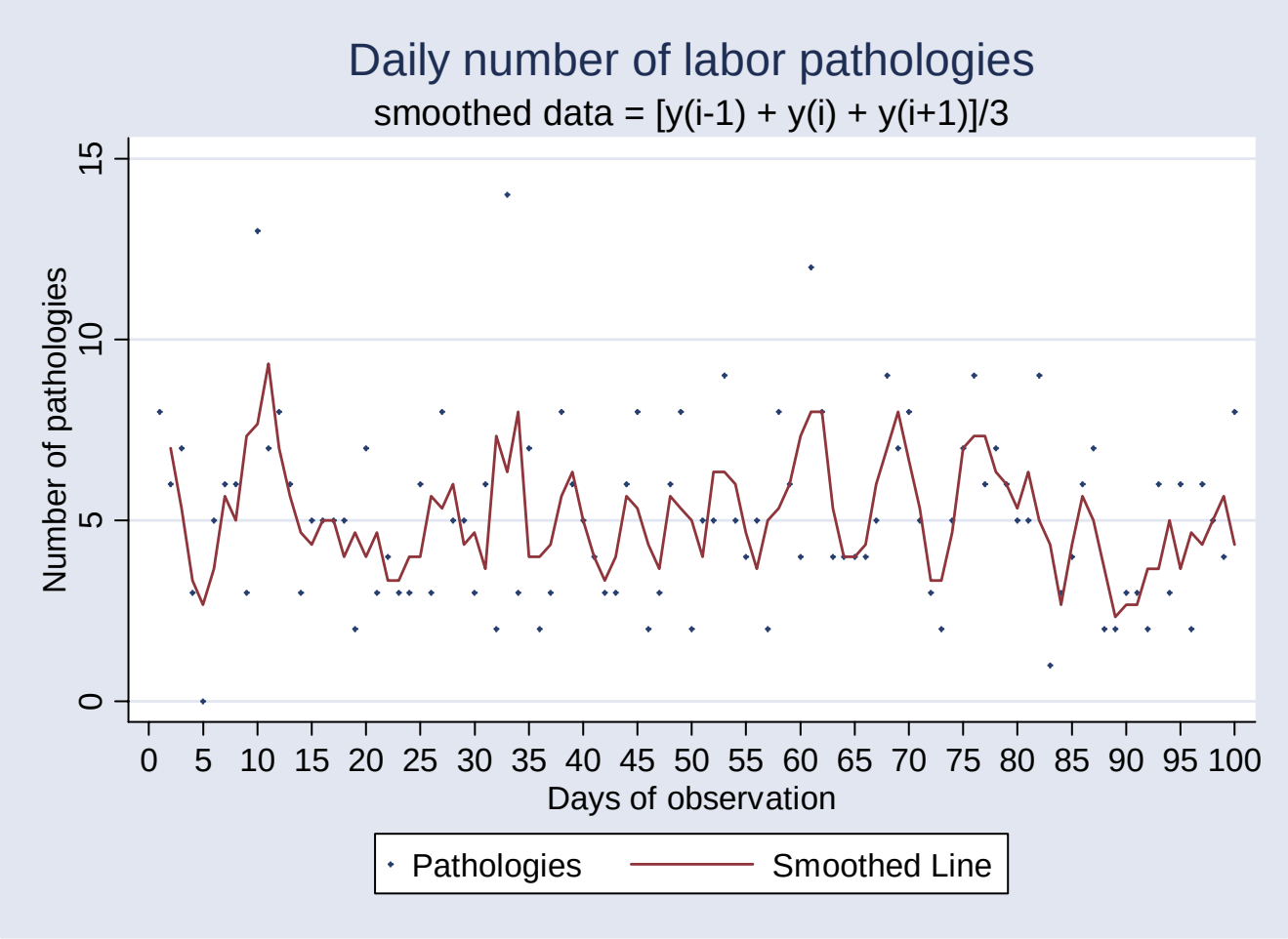
$$F_s = \frac{1}{k}(y_{s-p} \ldots + y_{s-1} + y_s + y_{s+1} \ldots + y_{s+p}) = \frac{1}{k}\sum_{i=-p}^{p} y_{s+i}$$

Example.
Time series of Daily mortality(without cancer mortality) in Philadelphia:

D_t: 44, 42, 53, 46, 49, 41, 55, 47, 47, 52,…

Let k=3.

$$F_s = (y_{s-1} + y_s + y_{s+1})/3$$
$$F_6 = (49 + 41 + 55)/3 = 48.3$$



Numbers of pathologies during labor caused by different reasons (full amount per , shown as pluses, and smoothed data , y_i day) in Soroka Hospital. Raw data, , shown as connected line segments. $s_i = (y_{i-1} + y_i + y_{i+1})/3$

In general, moving average of time series

$$Y_t : y_1, y_2, \ldots, y_n$$

is time series

$F_t : F_1, F_2, \ldots, F_n$ defined by

$$F_s = \sum_{-p}^{p} w_i y_{s+i}$$

where window k=2p+1 and

$$w_1 + w_2 + \ldots w_k = 1$$

(w_i are positive number, weights of moving average).

Exponential smoothing.

Let

$$Y_t : y_1, y_2, \ldots, y_n$$

be a time series

For any time period t the smoothed value F_t is found by computing:

$$F_t = \alpha y_{t-1} + (1-\alpha) F_{t-1}$$

where $t \geq 3 \qquad 0 \leq \alpha \leq 1 \qquad F_2 = y_1$.

α is called the smoothing constant.

How to choose α?

In practice, the smoothing parameter is often chosen by search different solution for α. The different solution for α are tried starting, for example, with α =0.1 to α=0.9, with increments of 0.1. Then α is chosen so as to produce the smallest sums of squares(or mean of squares) for residuals.

Gardner(1985) reports that among practitioners, an α smaller than 0.3 frequently recommended.

However, Makridakis *et al*(1982) show that α Values above 0.3 frequently yield the best forecast.

Why is it called "Exponential"?

Let us consider

$$F_t = \alpha y_{t-1} + (1-\alpha)F_{t-1}$$

Let t=3.

$$F_3 = \alpha y_2 + (1-\alpha)F_2 .$$

Let t=4.

$$F_4 = \alpha y_3 + (1-\alpha)F_3 =$$

$$= \alpha y_3 + (1-\alpha)(\alpha y_2 + (1-\alpha)F_2) =$$

$$= \alpha y_3 + \alpha(1-\alpha)y_2 + (1-\alpha)^2 F_2 =$$

$$= \alpha y_3 + \alpha(1-\alpha)y_2 + (1-\alpha)^2 y_1 .$$

In general,

$$F_t = \alpha \sum_{i=1}^{t-1} (1-\alpha)^{i-1} y_{t-i} + (1-\alpha)^{t-2} F_2$$

Example.
Time series of Daily mortality(without cancer mortality) in Philadelphia:

D_t: 44, 42, 53, 46, 49, 41, 55, 47, 47, 52,…
α=0.1

time	y_i	F_i	Error	Er^2
1	44			
2	42	44	-2	4
3	53	43.8	9.2	84.64
4	46	44.72	1.28	1.63
5	49	49.32	-0.32	0.10
6	49	49.28	-0.28	0.08
7	33	49.26	-16.26	264.39
8	47	47.63	-0.63	0.40
9	47	47.56	-0.56	0.31
10	52	47.50	4.5	20.25

$$\sum Er^2 = 375.8$$

$\alpha=0.3$

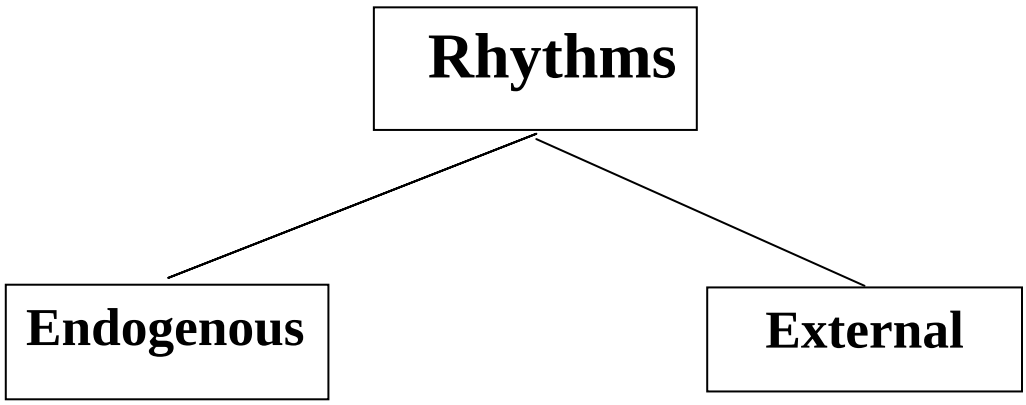
Time	y_i	F_i	Error	Er^2
1	44			
2	42	44	2	4
3	53	43.4	9.6	92.16
4	46	46.28	0.28	0.08
5	49	46.19	2.81	7.89
6	49	47.03	1.97	3.88
7	33	47.62	-14.62	213.74
8	47	43.23	3.77	14.21
9	47	44.36	2.64	6.97
10	52	45.15	6.85	46.92

$$\sum Er^2 = 385.85$$

Periodic components

Seasonality and Periodicity

Seasonality and Periodicity are description of different Rhythms.



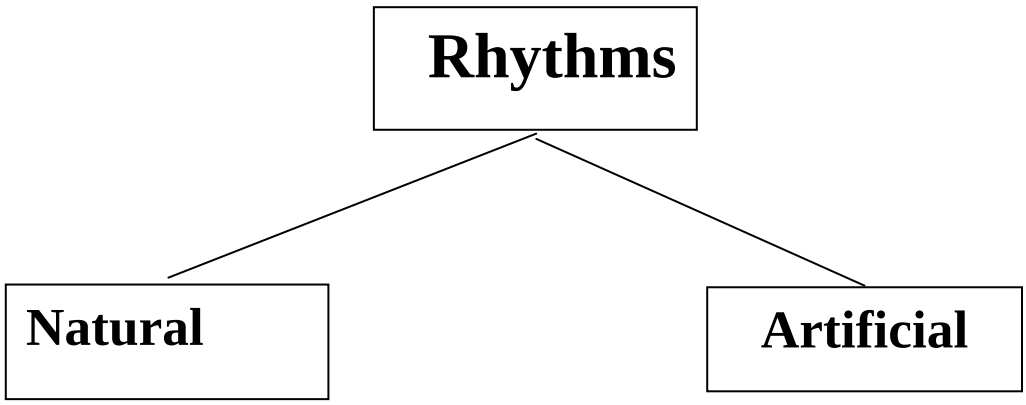


Table 1. The range of periods of biological rhythms of different frequency (Haus et al. 1980, Hyman 1990).

Term	Definition for rhythm	Range of period
Ultradian rhythms		< 20 hours
Circadian [a]		24 ± 4 hours
Dian	Daily rhythm (about 24 hours)	24 ± 0.2 hours
Infradian rhythms		> 28 hours
Circaseptan	Weekly rhythm (about a week)	7 ± 3 days
Circadiseptan	Rhythm about two weeks	14 ± 3 days
Circavigintan	Rhythm about three weeks	21 ± 3 days
Circatrigintan	Monthly rhythm (about a month)	30 ± 5 days
Circannual	Yearly (annual) rhythm (a solar year)	1 year ± 2 months
Circalunar	Lunar rhythm	about 29.5 days

[a] The Latin *circa* means “approximately“ and *dies* means a “day” (Hyman 1990).

Fundamental Period(Cycle) –

base for all other cycles.

Harmonic Analysis

Main goal: to describe, explain and predict periodic oscillations. To establish and check periods of such oscillations.

Basic model for harmonic analysis to study wave with specific period :

$$Y_t = a_0 + a \cdot Cos(t \cdot \omega) + b \cdot Sin(t \cdot \omega) + \varepsilon_t$$

where ω is frequency of wave.

$$\omega = \frac{2\pi}{P},$$

P is a period of wave.
To learn periodic behavior of time series usually periodograms are used.
Periodogram is a summary description based on a representation of an observed time series as a superposition of sinusoidal waves of various frequencies.

We could examine existing cyclic oscillations with a known a-priori period by model

$$Y_t = a_0 + \sum_{i=1}^{m} (a_i \cdot Cos(t \cdot \omega_i) + b_i \cdot Sin(t \cdot \omega_i)) + \varepsilon_t$$

where

$$\omega_i = \frac{2\pi}{P_i}$$

is a is frequency of wave number i,
P_i is a period of wave number i.
Usually all periods are based on few fundamental cycle.
Frequently fundamental cycles are:

1. Yearly cycle - 365 days(P_Y).
2. Weekly cycle – 7 days(P_W).

All other cycles are functions of fundamental cycles.

For example.

Seasonal cycle is P_Y /4.
Moon-month cycle is approximately 4*P_W .

Example 1

$$Temp_t = a_0 + a \cdot Cos(t \cdot \frac{2\pi}{365}) + b \cdot Sin(t \cdot \frac{2\pi}{365}) + \varepsilon_t$$

Model Summary[b]

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate	Change Statistics				
					R Square Change	F Change	df1	df2	Sig. F Change
1	.910[a]	.827	.827	7.42148	.827	6984.323	2	2918	.000

a. Predictors: (Constant), YSIN, YCOS

b. Dependent Variable: TEMP

ANOVA[b]

Model		Sum of Squares	df	Mean Square	F	Sig.
1	Regression	769370.5	2	384685.250	6984.323	.000[a]
	Residual	160718.7	2918	55.078		
	Total	930089.2	2920			

a. Predictors: (Constant), YSIN, YCOS

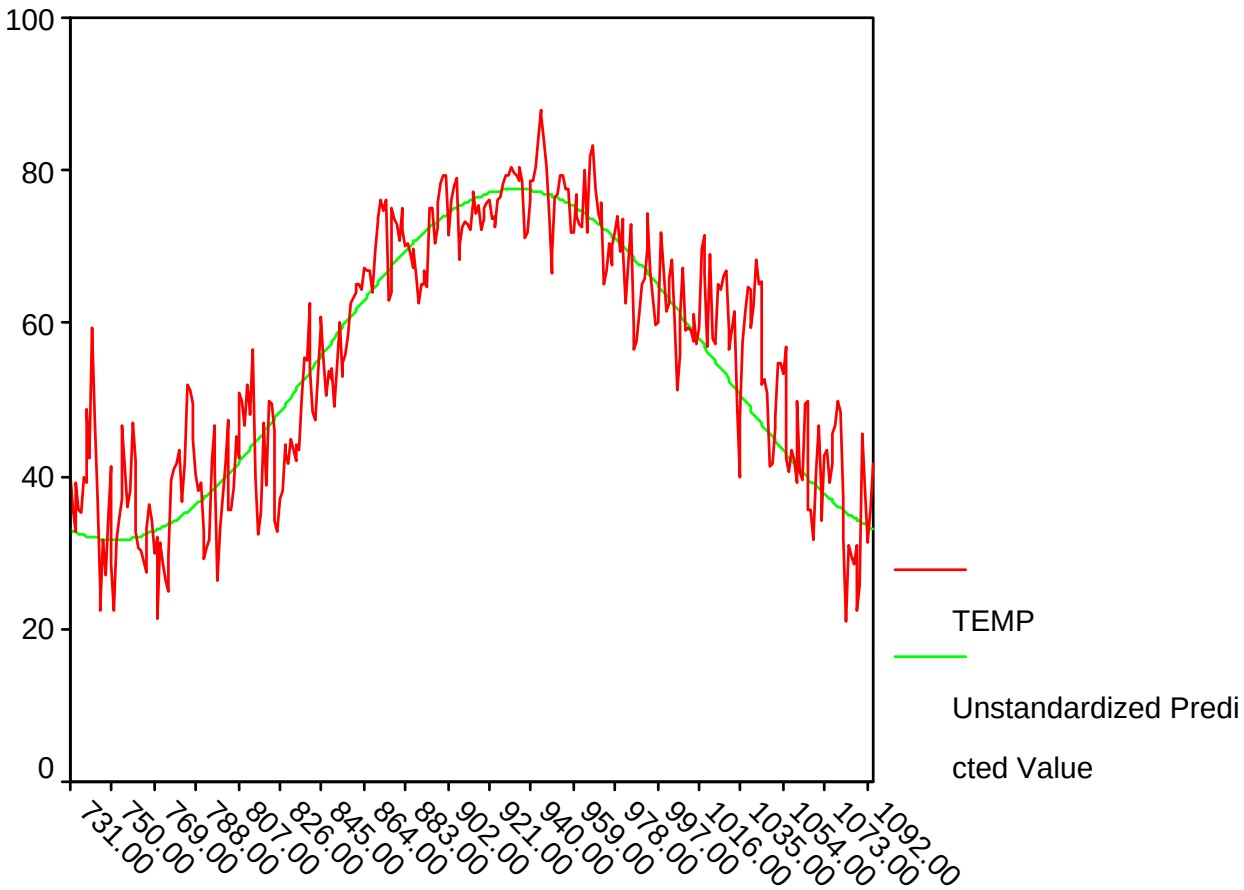
b. Dependent Variable: TEMP

Coefficients[a]

Model		Unstandardized Coefficients		Standardized Coefficients	t	Sig.	95% Confidence Interval for B	
		B	Std. Error	Beta			Lower Bound	Upper Bound
1	(Constant)	54.610	.137		397.694	.000	54.341	54.879
	YCOS	-21.573	.194	-.855	-111.103	.000	-21.954	-21.193
	YSIN	-7.836	.194	-.310	-40.346	.000	-8.217	-7.455

a. Dependent Variable: TEMP

$$Temp_t = 54.6 - 21.6 \cdot Cos(t \cdot \frac{2\pi}{365}) - 7.8 \cdot Sin(t \cdot \frac{2\pi}{365}) + \varepsilon_t$$


100
80
60
40
20
0
731.00
750.00
769.00
788.00
807.00
826.00
845.00
864.00
883.00
902.00
921.00
940.00
959.00
978.00
997.00
1016.00
1035.00
1054.00
1073.00
1092.00
SNO
TEMP
Unstandardized Predicted Value

Example 2

$$Temp_t = a_0 + a_1 \cdot Cos(t \cdot \frac{2\pi}{P}) + b_1 \cdot Sin(t \cdot \frac{2\pi}{P}) +$$

$$+ a_2 \cdot Cos(t \cdot \frac{2\pi}{P_1}) + b_2 \cdot Sin(t \cdot \frac{2\pi}{P_1}) +$$

$$+ a_3 \cdot Cos(t \cdot \frac{2\pi}{P_2}) + b_3 \cdot Sin(t \cdot \frac{2\pi}{P_2}) + \varepsilon_t$$

where P is a yearly cycle (fundamental cycle).
P_1 is half year cycle, $P_1 = P/2$.
P_2 is seasonal cycle, $P_2 = P/4$.

Model Summary[d]

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate	Change Statistics				
					R Square Change	F Change	df1	df2	Sig. F Change
1	.855[a]	.731	.731	9.26143	.731	7924.484	1	2919	.000
2	.910[b]	.827	.827	7.42148	.096	1627.782	1	2918	.000
3	.910[c]	.829	.829	7.38525	.002	29.698	1	2917	.000

a. Predictors: (Constant), YCOS

b. Predictors: (Constant), YCOS, YSIN

c. Predictors: (Constant), YCOS, YSIN, YCOS2

d. Dependent Variable: TEMP

Coefficients[a]

Model		Unstandardized Coefficients		Standardized Coefficients	t	Sig.	95% Confidence Interval for B	
		B	Std. Error	Beta			Lower Bound	Upper Bound
1	(Constant)	54.612	.171		318.697	.000	54.276	54.948
	YCOS	-21.571	.242	-.855	-89.020	.000	-22.046	-21.096
2	(Constant)	54.610	.137		397.694	.000	54.341	54.879
	YCOS	-21.573	.194	-.855	-111.103	.000	-21.954	-21.193
	YSIN	-7.836	.194	-.310	-40.346	.000	-8.217	-7.455
3	(Constant)	54.610	.137		399.647	.000	54.343	54.878
	YCOS	-21.573	.193	-.855	-111.646	.000	-21.952	-21.194
	YSIN	-7.836	.193	-.310	-40.543	.000	-8.215	-7.457
	YCOS2	-1.053	.193	-.042	-5.450	.000	-1.432	-.674

a. Dependent Variable: TEMP

$$Temp_t = 54.6 - 21.6 \cdot Cos(t \cdot \frac{2\pi}{P}) - 7.8 \cdot Sin(t \cdot \frac{2\pi}{P}) -$$

$$- 1.1 \cdot Cos(t \cdot \frac{2\pi}{P_1}) + 0 \cdot Sin(t \cdot \frac{2\pi}{P_1}) +$$

$$+ 0 \cdot Cos(t \cdot \frac{2\pi}{P_2}) + 0 \cdot Sin(t \cdot \frac{2\pi}{P_2}) + \varepsilon_t$$

Some important comments

1. Periodograms(harmonic analysis based on Sine-Cosine functions) usually are used for description oscillations with a-priori known periods. Sometimes for this proposes are regressions with dummy variables which describe certain period.

For example

A. For describing week period could be used the following dummy variables. Let us take Day 7 as reference day.

$$D_1 = \begin{Bmatrix} 1 & for\ \ day\ \ 1 \\ 0 & for\ \ any\ \ other\ \ day \end{Bmatrix}$$

$$D_2 = \begin{Bmatrix} 2 & for\ \ day\ \ 2 \\ 0 & for\ \ any\ \ other\ \ day \end{Bmatrix}$$

$$D_3 = \begin{Bmatrix} 3 & for\ \ day\ \ 3 \\ 0 & for\ \ any\ \ other\ \ day \end{Bmatrix}$$

$$D_4 = \begin{Bmatrix} 4 & for\ \ day\ \ 4 \\ 0 & for\ \ any\ \ other\ \ day \end{Bmatrix}$$

$$D_5 = \begin{Bmatrix} 5 & for\ \ day\ \ 5 \\ 0 & for\ \ any\ \ other\ \ day \end{Bmatrix}$$

$$D_6 = \begin{Bmatrix} 6 & for\ \ day\ \ 6 \\ 0 & for\ \ any\ \ other\ \ day \end{Bmatrix}$$

B. For describing season period could be used the following dummy variables. Let us take "Summer" as referent season.

$$D_1 = \begin{cases} 1 & for\ winter\ day \\ 0 & for\ any\ other\ day \end{cases}$$

$$D_2 = \begin{cases} 1 & for\ spring\ day \\ 0 & for\ any\ other\ day \end{cases}$$

$$D_3 = \begin{cases} 1 & for\ autumn\ day \\ 0 & for\ any\ other\ day \end{cases}$$

Example.

```
regression variables=dtotal d1 d2 d3 d4 d5 d6
/   dependent=dtotal/method=enter
```

Multiple R .10842
R Square .01175
Adjusted R Square .00972
Standard Error 8.13036

Analysis of Variance

	DF	Sum of Squares	Mean Square
Regression	6	2291.98775	381.99796
Residual	2915	192689.29630	66.10267

F = 5.77886 Signif F = .0000

---------- Variables in the Equation ----------

Variable	B	SE B	Beta	Sig T
d1	3.137981	.562725	.134499	.0000
d2	1.704967	.562725	.073078	.0025
d3	1.499225	.562725	.064259	.0078
d4	.988010	.563062	.042305	.0794
d5	1.100719	.563062	.047132	.0507
d6	.992806	.563062	.042511	.0780
Constant	36.323741	.398145		.0000

```
   regression variables=dtotal s1 s2 s3
/   dependent=dtotal/method=enter
```

Multiple R .20638
R Square .04259
Adjusted R Square .03464
Standard Error 7.97159

Analysis of Variance

DF	Sum of Squares	Mean Square
Regression	3	1020.52066	340.17355
Residual	361	22940.18071	63.54621

F = 5.35317 Signif F = .0013

---------- Variables in the Equation ----------

Variable	B	SE B	Beta	Sig T
S1	4.572841	1.178706	.243264	.0001
S2	1.301456	1.185134	.069234	.2729
S3	1.575350	1.160247	.085604	.1754
Constant	40.7207	.835747		.0000

2. **Special technique of Spectral analysis is used for study and description Quasi-cycle components. This technique is based on learning frequency behavior of time series.**

Autoregression

Most time series consists of elements that are serially dependent in the sense that one can estimate a coefficient or a set of coefficients that describe consecutive elements of the series from specific, time-lagged previous observations. This could be summarized in the following regression equation:

$$Y_t = \alpha_0 + \alpha_1 Y_{t-1} + \alpha_2 Y_{t-2} + \ldots + \alpha_m Y_{t-m} + \varepsilon_t$$

Multiple R	.50902
R Square	.25910
Adjusted R Square	.25732
Standard Error	7.00998

Analysis of Variance

	DF	Sum of Squares	Mean Square
Regression	7	49956.49049	7136.64150
Residual	2907	142849.27829	49.13976

----------Variables in the Equation----------

Variable	B	SE B	Beta	Sig T
DLAG1	.196692	.018533	.196452	.0000
DLAG2	.122252	.018826	.122206	.0000
DLAG3	.114260	.018865	.114105	.0000
DLAG4	.099694	.018900	.099433	.0000
DLAG5	.105040	.018867	.104778	.0000
DLAG6	.075388	.018832	.075164	.0001
DLAG7	.037331	.018531	.037238	.0440
Constant	9.38914	.913370		.0000

In practice combined model $Y = F(t)$ for realizing trend and periodic components used:

$$Y_t = a_0 + a_1 t + a_2 t^2 + a_3 t^3 + \ldots + a_k t^k +$$

$$+ \sum_{i=1}^{m} (_i \cdot Cos(t \cdot \omega_i) + b_i \cdot Sin(t \cdot \omega_i)) + \varepsilon_t$$

Example

```
regression variables= dtotal
                      t t2 t3 t4 t5 t6 t7 t8 t9 10
                      ycos ysin scos ssin ycos2 ysin2
                     s120cos s120sin msin mcos wmcos wmsin

                     /statistics=end
                     /dependent=dtotal
                     /method=stepwise
```

* * * * M U L T I P L E R E G R E S S I O N * * *

Step	MultR	Rsq	F(Eqn)	SigF		Variable	BetaIn
1	.3619	.1310	385.105	.000	In:	YCOS	.3619
2	.4122	.1699	261.414	.000	In:	YSIN	.1973
3	.4359	.1900	199.622	.000	In	t	-.1426
4	.4608	.2123	171.950	.000	In:	t10	.2096
5	.4790	.2295	151.926	.000	In:	YCOS2	.1310
6	.4900	.2401	134.276	.000	In:	YSIN2	.1034
7	.4983	.2483	120.281	.000	In:	SCOS	-.0906

Multiple R	.49829
R Square	.24830
Adjusted R Square	.24623
Standard Error	6.88952

Analysis of Variance

	DF	Sum of Squares	Mean Square
Regression	7	39964.23420	5709.17631
Residual	2549	120989.38645	47.46543

---------- Variables in the Equation ----------

Variable	B	SE B	Beta	Sig T
YCOS	3.98	.192	.355	.0000
YSIN	2.21	.194	.196	.0000
SCOS	-1.01	.193	-.090	.0000
YCOS2	1.46	.192	.130	.0000
YSIN2	1.14	.193	.102	.0000
t	-.003	.001	-.284	.0000
t10	.0001	.0001	. 215	.0000
Const	41.148	.388		.0000

$$Y_t = 41.15 - 0.03 \cdot t + 0.0001 \cdot t^{10} + 3.98 \cdot Cos(\omega_y \cdot t) +$$
$$+ 2.21 \cdot Sin(\omega_y \cdot t) + 1.46 \cdot Cos(\omega_{y/2} \cdot t) + 1.14 \cdot Sin(\omega_{y/2} \cdot t) -$$
$$- 1.01 \cdot Cos(\omega_{y/4} \cdot t) + \varepsilon_t$$

ARIMA models

The modeling and forecasting procedures discussed, involved knowledge about the mathematical model of the process. However, in real-life research and practice, patterns of the data are unclear, individual observations involve considerable error, and we still need not only to uncover the hidden patterns in the data but also generate forecasts. The ARIMA methodology developed by Box and Jenkins (1976) allows us to do just that; it has gained enormous popularity in many areas and research practice confirms its power and flexibility (Hoff, 1983; Pankratz, 1983; Vandaele, 1983). However, because of its power and flexibility, ARIMA is a complex technique; it is not easy to use, it requires a great deal of experience, and although it often produces satisfactory results, those results depend on the researcher's level of expertise.

For any Time Series

$$Y_t : y_1, y_2, \ldots, y_n$$

We have three following processes:

1. Moving average process

$F_t : F_1, F_2, \ldots, F_n$ defined by

$$F_s = \sum_{-p}^{p} w_i y_{s+i} \quad .$$

2. Differencing process

$$D^k(Y_t) = D(D^{k-1}(Y_t)) = D^{k-1}(Y_t) - D^{k-1}(Y_{t-1})$$

3. Autoregressive process.

$$Y_t = \alpha_0 + \alpha_1 Y_{t-1} + \alpha_2 Y_{t-2} + \ldots + \alpha_m Y_{t-m} + \varepsilon_t$$

Let us consider Moving Average process.

Example.

Window k=1.

$$Y_1 = w_0 Y_0 + \varepsilon_1, \quad Y_0 = 0 \quad \Rightarrow Y_1 = \varepsilon_1$$

$$Y_2 = w_1 Y_1 + \varepsilon_2 = w_1 \varepsilon_1 + \varepsilon_2$$

$$Y_3 = w_2 Y_2 + \varepsilon_3 = w_2 (w_1 \varepsilon_1 + \varepsilon_2) + \varepsilon_3 =$$

$$= w_2 w_1 \varepsilon_1 + w_2 \varepsilon_2 + \varepsilon_3$$

In general, Moving Average process could be presented in the following form:

$$Y_t = \varepsilon_t + \sum_{j=1}^{q} a_j \varepsilon_{t-j} .$$

If we compare Moving Average process

$$Y_t = \varepsilon_t + \sum_{j=1}^{q} b_j \varepsilon_{t-j} \quad .$$

with Autoregressive process

$$Y_t = \alpha_0 + \alpha_1 Y_{t-1} + \alpha_2 Y_{t-2} + \ldots + \alpha_m Y_{t-m} + \varepsilon_t$$

we can see that both of equations are identical,

But roles of $\{Y_i\}$ and $\{\varepsilon_i\}$ reversed.

These equations represent some kind of duality.

Most time series consist of elements that are serially dependent in the sense that one can estimate a coefficient or a set of coefficients that describe consecutive elements of the series from specific, time-lagged (previous) elements.

Independent from the autoregressive process, each element in the series can also be affected by the past error (or random shock) that cannot be accounted for by the autoregressive component.

Now we define autoregressive moving average process ARMA(*p, q)*:

$$Y_t = \sum_{i=1}^{p} a_i Y_{i-1} + \varepsilon_t + \sum_{j=1}^{q} b_j \varepsilon_{t-j},$$

where *p* and *q* are called orders of the process.

Order *p* reflects the autoregressive parameters
Order *q* reflects moving average parameters .

There exists another form of ARMA(*p, q)*:

$$Y_t - \mu = \sum_{i=1}^{p} a_i (Y_{i-1} - \mu) + \varepsilon_t + \sum_{j=1}^{q} b_j \varepsilon_{t-j},$$

where μ is mean value.

Stationarity requirement. Note that an autoregressive process will only be stable if the parameters are within a certain range;

Autoregressive Integrated Moving Average Model(ARIMA).

The general model introduced by Box and Jenkins (1976) includes

Autoregressive ,

Moving average parameters,

Differencing

in the formulation of the model.

Specifically, the three types of parameters in the model are: the autoregressive parameters (*p*), the number of differencing passes (*d*), and moving average parameters (*q*). In the notation introduced by Box and Jenkins, models are summarized as ARIMA (*p, d, q*);

Example

1. A model described as (0, 1, 2) means that it contains 0 (zero) autoregressive (p) parameters and 2 moving average (q) parameters which were computed for the series after it was differenced once.

2. ARIMA(0,1,0) = Random Walk.

$Y_t = Y_{t-1} + \varepsilon_i$ (From differencing). Other form

$\hat{Y}_t - Y_{t-1} = \mu,$ where μ is average difference in Y.

3. ARIMA(1,0,1).

$Y_t = aY_{t-1} + b\varepsilon_{i-1} + \varepsilon_i$. Other form

$\hat{Y}_t = aY_{t-1} + b\varepsilon_{i-1}$

4. ARIMA(1,1,0) = differenced first-order autoregressive model(the same as first order autoregressive model on $D(Y_i)$.

$\hat{Y}_t - Y_{i-1} = \mu + a(Y_{t-1} - Y_{t-2})$ or

$\hat{Y}_t = \mu + Y_{t-1} + a(Y_{t-1} - Y_{t-2})$.

5. ARIMA(0,1,1) =simple exponential smoothing.

$\hat{Y}_t = Y_{t-1} + b\varepsilon_{t-1}$.

Why?

Exponential smoothing is

$F_t = \alpha y_{t-1} + (1-\alpha)F_{t-1}$.

So, we have

$\hat{Y}_t = \alpha y_{t-1} + (1-\alpha)F_{t-1}$, but

$Y_{t-1} - \hat{Y}_{t-1} = Y_{t-1} - F_{t-1} = \varepsilon_{t-1} \Rightarrow F_{t-1} = Y_{t-1} - \varepsilon_{t-1}$ and

$$\hat{Y}_t = F_t = \alpha y_{t-1} + (1-\alpha)F_{t-1} =$$
$$= \alpha Y_{t-1} + (1-\alpha)(Y_{t-1} - \varepsilon_{t-1}) =$$
$$= \alpha Y_{t-1} + (1-\alpha)Y_{t-1} - (1-\alpha)\varepsilon_{t-1} =$$
$$= Y_{t-1} + b\varepsilon_{t-1},$$

where

$$b = -(1-\alpha).$$

6. ARIMA(1,1,1) = mixed model.

$$\hat{Y}_t = \mu + Y_{t-1} + a(Y_{t-1} - Y_{t-2}) + b\varepsilon_{t-1}\,.$$

Important note. Normally, in practice, 'unmixed models" are used (either model with only-AR parameters or only-MA parameters), because

1. duality of AR and MA parameters;

2. including both kind of terms in some model sometimes leads to overfitting of the data and non-uniqueness of the coefficients.

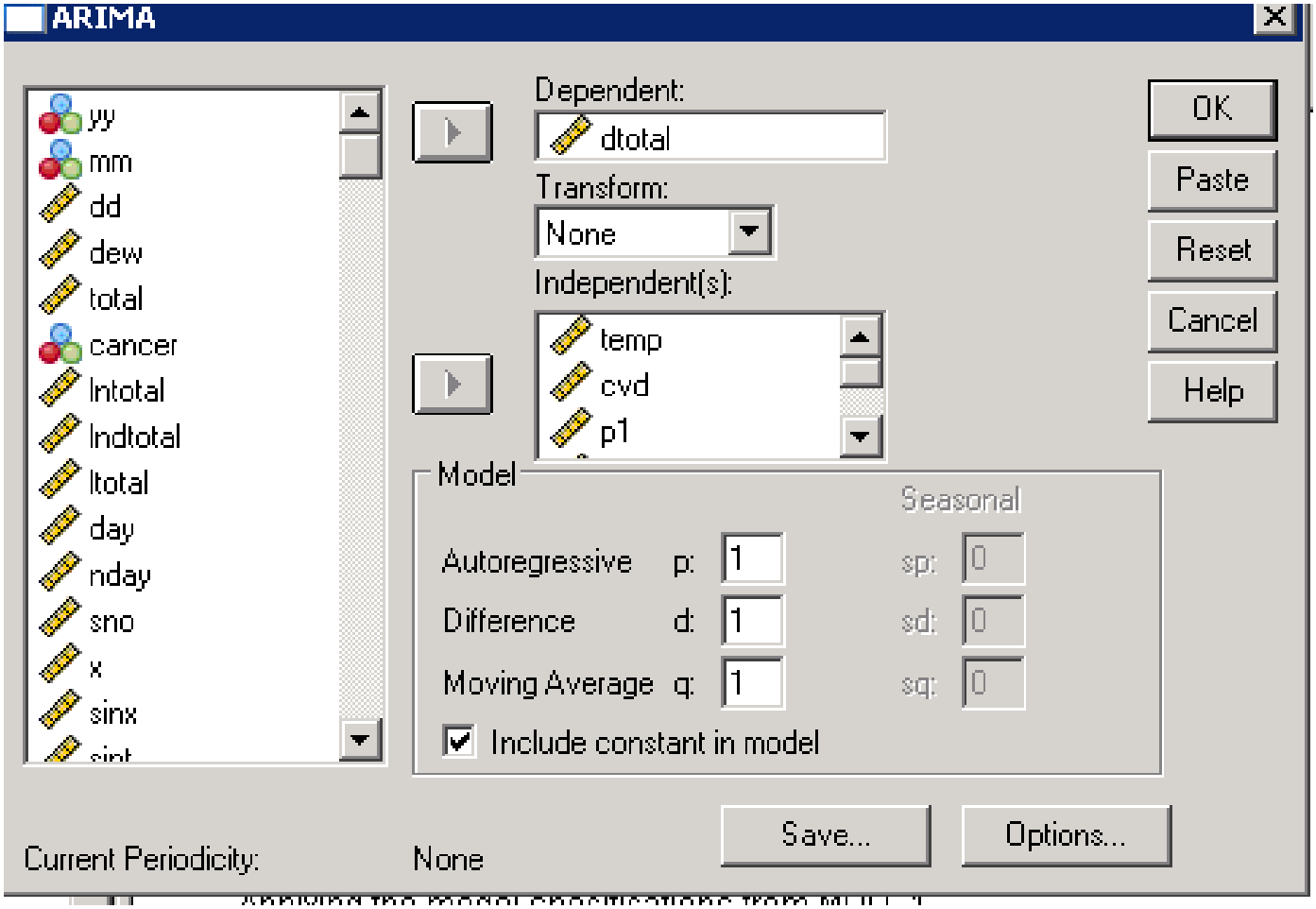
Identification.

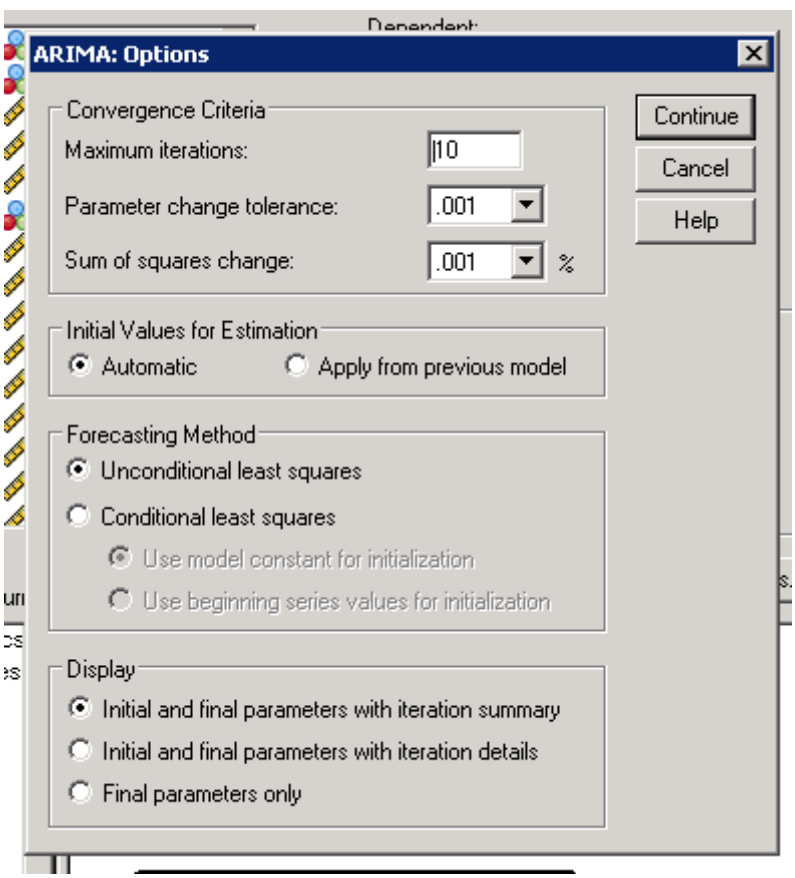
As mentioned earlier, the input series for ARIMA needs to be stationary, that is, it should have a constant mean, variance, and autocorrelation through time. Therefore, usually the series first needs to be differenced until it is stationary (this also often requires log transforming the data to stabilize the variance). The number of times the series needs to be differenced to achieve stationarity is reflected in the *d* parameter (see the previous paragraph). In order to determine the necessary level of differencing, one should examine the plot of the data and autocorrelogram. Significant changes in level (strong upward or downward changes) usually require first order non seasonal (lag=1) differencing; strong changes of slope usually require second order non seasonal differencing. Seasonal patterns require respective seasonal differencing (see below). If the estimated autocorrelation coefficients decline slowly at longer lags, first order differencing is usually needed. However, one should keep in mind that some time series may require little or no

differencing, and that *over differenced* series produce less stable coefficient estimates.

At this stage (which is usually called *Identification* phase, see below) we also need to decide how many autoregressive (p) and moving average (q) parameters are necessary to yield an effective but still *parsimonious* model of the process (*parsimonious* means that it has the fewest parameters and greatest number of degrees of freedom among all models that fit the data). In practice, the numbers of the p or q parameters very rarely need to be greater than 2 (see below for more specific recommendations).

This procedure estimates nonseasonal and seasonal univariate ARIMA (Autoregressive Integrated Moving Average) models (also known as "Box-Jenkins" models) with or without fixed regressor variables. The procedure produces maximum-likelihood estimates and can process time series with missing observations.


ARIMA
yy
mm
dd
dew
total
cancer
lntotal
lndtotal
ltotal
day
nday
sno
x
sinx
Dependent:
dtotal
Transform:
None
Independent(s):
temp
cvd
p1
OK
Paste
Reset
Cancel
Help
Model
Seasonal
Autoregressive p: 1 sp: 0
Difference d: 1 sd: 0
Moving Average q: 1 sq: 0
Include constant in model
Current Periodicity: None
Save...
Options...


ARIMA: Options
Convergence Criteria
Maximum iterations: 10
Parameter change tolerance: .001
Sum of squares change: .001 %
Continue
Cancel
Help
Initial Values for Estimation
Automatic
Apply from previous model
Forecasting Method
Unconditional least squares
Conditional least squares
Use model constant for initialization
Use beginning series values for initialization
Display
Initial and final parameters with iteration summary
Initial and final parameters with iteration details
Final parameters only

Practical aspects

How to perform ARIMA models by SPSS.

Example 1

Model Description(a)	
Model Name	MOD_1
Dependent Series	dtotal
Transformation	None
Constant	Included
AR	None
Non-Seasonal Differencing	1
MA	None

Parameter Estimates				
	Estimates	**Std Error**	**t**	**Approx Sig**
Constant	.007	.175	.040	.968

Example 2

Model Description(a)	
Model Name	MOD_2
Dependent Series	dtotal
Transformation	None
Constant	Included
AR	1
Non-Seasonal Differencing	0
MA	1

Parameter Estimates					
		Estimates	**Std Error**	**t**	**Approx Sig**
Non-Seasonal Lags	**AR1**	.981	.005	182.574	.000
	MA1	.854	.014	62.935	.000
Constant		37.484	1.024	36.607	.000
Melard's algorithm was used for estimation.					

Example 3

Model Description(a)	
Model Name	MOD_3
Dependent Series	dtotal
Transformation	None
Constant	Included
AR	1
Non-Seasonal Differencing	1
MA	None

Parameter Estimates					
		Estimates	**Std Error**	**t**	**Approx Sig**
Non-Seasonal Lags	**AR1**	-.472	.017	-27.0	.000
Constant		.006	.105	.057	.954
Melard's algorithm was used for estimation.					

Example 4

Model Description(a)	
Model Name	MOD_4
Dependent Series	dtotal
Transformation	None
Constant	Included
AR	None
Non-Seasonal Differencing	1
MA	1

Parameter Estimates					
		Estimates	**Std Error**	**t**	**Approx Sig**
Non-Seasonal Lags	**MA1**	.876	.010	90.690	.000
Constant		.004	.017	.210	.834
Melard's algorithm was used for estimation.					

Example 5

Model Description(a)	
Model Name	MOD_5
Dependent Series	dtotal
Transformation	None
Constant	Included
AR	1
Non-Seasonal Differencing	1
MA	1

Iteration History					
	Non-Seasonal Lags		Constant	Adjusted Sum of Squares	Marquardt Constant
	AR1	MA1			
0	.052	.685	.005	127759.940	.001
1	.132	.944	.001	120750.895	.001
2	.088	.913	.003	119170.166	.000
3	.066	.896	.003	118949.449	.000
4	.061	.892	.003	118941.648(a)	.000

Parameter Estimates					
		Estimates	Std Error	t	Approx Sig
Non-Seasonal Lags	**AR1**	.061	.023	2.703	.007
	MA1	.891	.010	85.670	.000
Constant		.003	.016	.207	.836

Example 6

Model Description(a)		
Model Name		MOD_6
Dependent Series		dtotal
Transformation		None
Independent Series	**1**	temp
	2	cvd
	3	p1
	4	p2
Constant		Included
AR		1
Non-Seasonal Differencing		1
MA		1

Parameter Estimates					
		Estimates	**Std Error**	**t**	**Approx Sig**
Non-Seasonal Lags	**AR1**	.060	.022	2.772	.006
	MA1	.937	.008	121.429	.000
Regression Coefficients	**temp**	-.006	.012	-.560	.575
	cvd	1.078	.018	61.194	.000
	p1	.006	.005	1.232	.218
	p2	.018	.011	1.687	.092
Constant		.002	.006	.327	.743

General Linear Model for time series analysis

Let us consider model

$$Y = F(t, X_1(t), \cdots X_2(t), \ldots, X_n(t))$$

where Y be dependent variable and $X_1, X_2, \ldots, X_n$ be list of independent variables and additional independent time variable *t*.
Y is random normal distributed variable, i.e. $Y \sim N(\mu, \sigma^2)$
$X_1, X_2, \ldots, X_n$, t are non-random variables bu all of Y, $X_1, X_2, \ldots, X_n$ are functions of T.

For building general linear model based on time series it is rather common to use “filtration” approach or “hierarchical” approach.

“Filtration” approach

We use General Linear Model in form:

$$Y_t = a_0 + a_1 t + a_2 t^2 + a_3 t^3 + \ldots + a_k t^k +$$

$$+ \sum_{i=1}^{m} \left(c_i \cdot Cos(t \cdot \omega_i) + b_i \cdot Sin(t \cdot \omega_i) \right) +$$

$$+ \Phi(X_1(t), X_2(t), \ldots, X_n(t)) + \varepsilon_t$$

In particular,

$$Y_t = a_0 + a_1 t + a_2 t^2 + a_3 t^3 + \ldots + a_k t^k +$$

$$+ \sum_{i=1}^{m} \left(c_i \cdot Cos(t \cdot \omega_i) + b_i \cdot Sin(t \cdot \omega_i) \right) +$$

$$+ d_1 \cdot X_1(t) + d_2 \cdot X_2(t) + \ldots + d_n \cdot X_n(t) + \varepsilon_t$$

Example

```
regression variables= dtotal ycos ysin scos ssin ycos2 ysin2
                      s120cos s120sin msin mcos wmcos wmsin
                      t t1 t2 t3 t4 t5 t6 t7 t8 t9 t10
                      temp dew  p1 p2
                      /statistics=end
                      /dependent=dtotal
                      /method=stepwise
```

*** * * * M U L T I P L E R E G R E S S I O N * * * ***

Step	MultR	Rsq	F(Eqn)	SigF		Variable	BetaIn
1	.3670	.1347	390.813	.000	In:	YCOS	.3670
2	.4157	.1728	262.083	.000	In:	YSIN	.1952
3	.4417	.1951	202.677	.000	In:	t	-.1506
4	.4676	.2187	175.416	.000	In:	t10	.2154
5	.4856	.2358	154.667	.000	In:	YCOS2	.1309
6	.4954	.2454	135.793	.000	In:	YSIN2	.0985
7	.5041	.2541	121.863	.000	In:	SCOS	-.0932
8	.5117	.2618	110.957	.000	In:	P1	.0897
9	.5130	.2632	99.311	.000	In:	DEW	.0733
10	.5147	.2649	90.132	.000	In:	TEMP	-.1930
11	.5159	.2662	82.436	.000	In:	P2	.0550

Multiple R .51592
R Square .26617
Adjusted R Square .26294
Standard Error 6.82489

Analysis of Variance

	DF	Sum of Squares	Mean Square
Regression	11	42237.76748	3839.79704
Residual	2500	116447.81055	46.57912

*** * * * M U L T I P L E R E G R E S S I O N * * * ***

---------- Variables in the Equation ----------

Variable	B	SE B	Beta	Sig T
YCOS	3.7	.48	.332	.0000
YSIN	2.2	.25	.198	.0000
SCOS	-1.0	.19	-.095	.0000
YCOS2	1.3	.20	.117	.0000
YSIN2	.9	.19	.086	.0000
t	-.003	0.0002	-.277	.0000
t10	0.0001	0.0001	.224	.0000
TEMP	-.08	.035	-.181	.0237
DEW	.07	.024	.190	.0017
P1	.014	.007	.052	.0357
P2	.034	.016	.055	.0385
Const	0.33	1.272		.0000

“Hierarchical” approach

We use General Linear Model in form:

$$Y_t = a_0 + a_1 t + a_2 t^2 + a_3 t^3 + \ldots + a_k t^k +$$
$$+ \sum_{i=1}^{m} (c_i \cdot Cos(t \cdot \omega_i) + b_i \cdot Sin(t \cdot \omega_i)) +$$
$$+ \Phi(X_1(t), X_2(t), \ldots, X_n(t)) +$$
$$\sum_{i=1}^{m} (c_i \cdot Cos(t \cdot \omega_i) \varphi(X_1(t), X_2(t), \ldots, X_n(t))$$
$$+ b_i \cdot Sin(t \cdot \omega_i) \varphi(X_1(t), X_2(t), \ldots, X_n(t))) + \varepsilon_t$$

Example

Let us expand (decompose) this "one-to-many" structural model into series of "one-to-one" models. "One-to-one" relationship means both of the related components have unique indexes. We can reveal the following "one-to-one" models:

1).G→H

2).S→H

3).M→H

4).G→M→H

5).G→M→P→H

6).S→P→H

7).P→H

8).I→H.

In mathematic form:

$$
\begin{aligned}
H_t = {} & F_1^{GH}(\mathbf{G_t}, t) + \\
& + F_2^{SH}(\mathbf{S_t}, t) + \\
& + F_3^{MH}(\mathbf{M_t}, t) + \\
& + F_4^{GM}(\mathbf{G_t}, t) F_4^{MH}(\mathbf{M_t}, t) + \\
& + F_5^{GH}(\mathbf{G_t}, t) F_5^{GH}(\mathbf{M_t}, t) F_5^{PH}(\mathbf{P_t}, t) + \\
& + F_6^{SP}(\mathbf{S_t}, t) F_6^{PH}(\mathbf{P_t}, t) + \\
& + F_7^{PH}(\mathbf{P_t}, t) + \\
& + F_8^{IH}(\mathbf{I_t}, t)
\end{aligned}
\qquad (1)
$$

where:

- t is the time variable;
- $\mathbf{G_t}$, $\mathbf{S_t}$, $\mathbf{M_t}$, $\mathbf{P_t}$, $\mathbf{I_t}$ are the vectors of geophysical, socio-cultural, meteorological, pollution and individual variables, respectively;
- H_t is the "health" outcome (dependent variable);
- the terms F_1^{GH}, F_2^{SH}, F_3^{MH}, F_7^{PH}, F_8^{PH} are the functions expressing the "direct" effects of geophysical G, socio-cultural S, meteorological M, pollution P and individual I components on "health" H.

The terms describing "indirect effects" are:

- F_4^{GM} - geophysical component G effect on "health" H (via meteorology M);
- F_4^{GMH} - meteorological component M effect on "health" H (after removing influence of geophysical component G on meteorology M);
- F_5^{GM} - geophysical component G effect on "health" H (via meteorology M and pollution P);
- F_5^{GMP} - meteorological component M effect on "health" H (via pollution P and after removing influence of geophysical component G on meteorology M);
- F_5^{GMPH} - pollution P effect on "health" H (after removing influence of geophysical component G on pollution P);
- F_6^{SH} - socio-cultural component S effect on "health" H (via pollution P);
- F_6^{SPH} - pollution P effects on "health" H (after removing influence of socio-cultural component S on pollution P).

The next step in model building is determining the type of all of functions F_i**** in the model.

Let's consider the functions which are related to component G.
We hypothesize that the periodic seasonal changes (caused by geophysical processes) have an effect on the weather, pollution P and "health" H. Weather, in its turn, also influences pollution P and health outcome H. So, the relationship between P and H has both direct and indirect links.
Hence it follows that all the functions of G can be presented by pairs of sinusoidal and co-sinusoidal functions with various periods. We pick up just 4 periods built on the twelve-monthly cycle.
Periodical changes in socio-cultural component S are described by pairs of sinusoidal and co-sinusoidal functions with 7-days period. Furthermore, periodic changes in S can be described by a set of 6 dummy variables representing each day of the week.
Other model functions can be constructed using polynomial and/or logarithmic functions.
Bellow we formulate the model which realizes the described approach.
Let be variables $I = (I_1, I_2, \ldots, I_{r_l})$ and $P = (P_1, P_2, \ldots, P_{r_k})$, $M = (M_1, M_2, \ldots, M_{r_m})$ representing meteorological M, pollution P and individual I components, respectively.
Then we get

$$
\begin{aligned}
H_t &= q_0 + q^0 + \sum_{j=1}^{r_m} q_j M_j + \sum_{k=1}^{r_p} q_k P_k + \sum_{l=1}^{r_i} q_l I_l + \\
&+ \sum_{i=1}^{g} \left(a_i \cos\left(\mathrm{T}_i \cdot \omega t\right) + b_i \sin\left(\mathrm{T}_i \cdot \omega t\right)\right) + \\
&+ \sum_{j=1}^{r_p} \left(c_j P_j \cos\left(7\omega t\right) + d_j P_j \sin\left(7\omega t\right)\right) + \\
&+ \sum_{j=1}^{r_m} \sum_{i=1}^{g} \left(a_{ij} M_j \cos\left(\mathrm{T}_i \cdot \omega t\right) + b_{ij} M_j \sin\left(\mathrm{T}_i \cdot \omega t\right)\right) + \\
&+ \sum_{j=1}^{r_p} \sum_{i=1}^{g} \left(c_{ij} P_j \cos\left(\mathrm{T}_i \cdot \omega t\right) + d_{ij} P_j \sin\left(\mathrm{T}_i \cdot \omega t\right)\right) + \\
&+ \sum_{k=1}^{r_p} \sum_{j=1}^{r_m} \sum_{i=1}^{g} \left(a_{ij}^k M_j P_k \cos\left(\mathrm{T}_i \cdot \omega t\right) + b_{ij}^k M_j P_k \sin\left(\mathrm{T}_i \cdot \omega t\right)\right)
\end{aligned}
\qquad (2)
$$

where:

- terms $\mathrm{T}_i \cdot \omega t$ (where $\omega = \frac{2\pi}{365.4}$) are various periods built on annual cycle (in our case: month, season, 6 months and year);
- terms $7\omega t$ represents weekly cycling of socio-cultural component S;
- *q, a, b, c, d* are the constants obtained from an appropriate multiple regression.

Fragment of program.

```
compute         mcos=cos(6.28*sno/30)
compute         msin=sin(6.28*sno/30)
compute         wcos=cos(6.28*sno/7)
compute         wsin=sin(6.28*sno/7)
compute         scos=cos(6.28*sno/90)
compute         ssin=sin(6.28*sno/90)
compute         ycos=cos(6.28*sno/365)
compute         ysin=sin(6.28*sno/365)
compute         y14cos=cos(6.28*sno/(5110))
compute         y14sin=sin(6.28*sno/5110)
compute         y11cos=cos(6.28*sno/(4015))
compute         y11sin=sin(6.28*sno/(4015))
compute         y12cos=cos(6.28*sno/(4380))
compute         y12sin=sin(6.28*sno/(4380))
compute         y13cos=cos(6.28*sno/(4745))
compute         y13sin=sin(6.28*sno/(4745))
compute         y15cos=cos(6.28*sno/(5475))
compute         y15sin=sin(6.28*sno/(5475))
compute         y16cos=cos(6.28*sno/(5840))
compute         y16sin=sin(6.28*sno/(5840))
compute         ycos2=cos(2*6.28*sno/365)
compute         ysin2=sin(2*6.28*sno/365)
compute         s120cos=cos(6.28*sno/120)
compute         s120sin=sin(6.28*sno/120)
compute         wmcos=cos(6.28*sno/28)
compute         wmsin=sin(6.28*sno/28)
compute         sno2=sno*sno
compute         sno3=sno2*sno
compute         sno4=sno3*sno
compute         sno5=sno4*sno
compute         sno6=sno5*sno
compute         sno7=sno6*sno
compute         sno8=sno7*sno
compute         sno9=sno8*sno
compute         sno10=sno9*sno
*meteo factors
*temperature
* temp for year
compute ytcos=temp*ycos
compute ytsin=temp*ysin
```

```
*temp for1/2 year
compute ytcos2=temp*ycos2
compute ytsin2=temp*ysin2
*temp for season
compute stcos=temp*scos
compute stsin=temp*ssin
*temp for 120-days season
compute s120tcos=temp*s120cos
compute s120tsin=temp*s120sin
*dew
* dew for year
compute ydewcos=dew*ycos
compute ydewsin=dew*ysin
*dew for1/2 year
compute ydewcos2=dew*ycos2
compute ydewsin2=dew*ysin2
*dew for season
compute sdewcos=dew*scos
compute sdewsin=dew*ssin
*dew for 120-days season
compute s120dcos=dew*s120cos
compute s120dsin=dew*s120sin
*pollutants
*P1
* p1 for year
compute yp1cos=p1*ycos
compute yp1sin=p1*ysin
*p1 for1/2 year
compute yp1cos2=p1*ycos2
compute yp1sin2=p1*ysin2
*p1 for season
compute sp1cos=p1*scos
compute sp1sin=p1*ssin
*p1 for 120-days season
compute s120p1cs=p1*s120cos
compute s120p1sn=p1*s120sin
compute p1d1=p1*num1
compute p1d2=p1*num2
compute p1d3=p1*num3
compute p1d4=p1*num4
compute p1d5=p1*num5
compute p1d6=p1*num6
compute p1d7=p1*num7
```

```
*P2
* p2 for year
compute yp2cos=p2*ycos
compute yp2sin=p2*ysin
*p2 for1/2 year
compute yp2cos2=p2*ycos2
compute yp2sin2=p2*ysin2
*p2 for season
compute sp2cos=p2*scos
compute sp2sin=p2*ssin
*p2 for 120-days season
compute s120p2cs=p2*s120cos
compute s120p2sn=p2*s120sin
compute p2d1=p2*num1
compute p2d2=p2*num2
compute p2d3=p2*num3
compute p2d4=p2*num4
compute p2d5=p2*num5
compute p2d6=p2*num6
compute p2d7=p2*num7
compute tempp1=temp*p1
compute tempp2=temp*p2
compute dewp1=dew*p1
compute dewp2=dew*p2
regression variables=dtotal  temp dew p1 p2 ycos ysin scos ssin ycos2
ysin2
                s120cos s120sin
                sno sno2 sno3 sno4 sno5 sno6 sno7 sno8 sno9 sno10
            ytcos ytsin ytcos2 ytsin2 stcos stsin s120tcos s120tsin
        ydewcos ydewsin ydewcos2 ydewsin2 sdewcos sdewsin s120dcos
s120dsin
        yp1cos yp1sin yp1cos2 yp1sin2 sp1cos sp1sin s120p1cs s120p1sn
        yp2cos yp2sin yp2cos2 yp2sin2 sp2cos sp2sin s120p2cs s120p2sn
  p1d1 p1d2 p1d3 p1d4 p1d5 p1d6 p1d7
  p2d1 p2d2 p2d3 p2d4 p2d5 p2d6 p2d7
              /statistics=end
          /dependent=dtotal
              /method=stepwise
```

Year 78

**** MULTIPLE REGRESSION ****

Step	MultR	Rsq	F(Eqn)	SigF		Variable	BetaIn
1	.4992	.2492	119.483	.000	In:	TEMP	-.4992
2	.5481	.3004	77.073	.000	In:	YP1COS2	.2263
3	.5731	.3284	58.350	.000	In:	P1	.1691
4	.5880	.3458	47.169	.000	In:	YCOS	.2519
5	.6439	.4146	50.427	.000	In:	SNO	-.3762
6	.6511	.4239	43.532	.000	In:	YP2SIN2	-.1135
7	.6571	.4317	38.421	.000	In:	SNO4	7.9735
8	.6565	.4310	44.818	.000	Out:	YCOS	
9	.6630	.4396	39.663	.000	In:	SDEWCOS	-.0938
10	.6712	.4506	36.184	.000	In:	SP1COS	.2287
11	.6685	.4469	40.869	.000	Out:	YP1COS2	

Multiple R .66854
R Square .44695
Adjusted R Square .43601
Standard Error 6.29450

Analysis of Variance

	DF	Sum of Squares	Mean Square
Regression	7	11334.75870	1619.25124
Residual	354	14025.74959	39.62076

F = 40.86876 Signif F = .0000

---------- Variables in the Equation ----------

Variable	B	SE B	Beta	Sig T
TEMP	.16	.042	.356	.0002
P1	.03	.013	.096	.0183
t	-1.01	.102	-12.818	.0000
t4	0.00001	0.00	12.354	.0000
SDEWCOS	-.07	.022	-.300	.0004
SP1COS	.03	.013	.236	.0062
YP2SIN2	-.06	.022	-.134	.0050
Const	1571.81	153.038		.0000

Year: 80

**** M U L T I P L E R E G R E S S I O N ****

Step	MultR	Rsq	F(Eqn)	SigF		Variable	BetaIn
1	.4084	.1668	72.653	.000	In:	TEMP	-.4084
2	.4590	.2106	48.300	.000	In:	P2	.2140
3	.4852	.2354	37.054	.000	In:	DEW	.5410
4	.5023	.2523	30.365	.000	In:	YP2SIN2	.1306
5	.5185	.2688	26.398	.000	In:	SNO10	.1493
6	.6040	.3648	34.264	.000	In:	YSIN	.5688
7	.6266	.3926	32.961	.000	In:	SNO8	-12.585
8	.6240	.3894	38.050	.000	Out:	TEMP	
9	.6327	.4003	34.047	.000	In:	SDEWCOS	-.1053
10	.6384	.4075	30.606	.000	In:	YDEWSIN	-.3364
11	.6445	.4154	28.030	.000	In:	SDEWSIN	-.0901
12	.6531	.4265	26.324	.000	In:	S120TSIN	-.1292
13	.6507	.4234	28.967	.000	Out:	SDEWCOS	
14	.6559	.4301	26.721	.000	In:	SP2COS	-.0864
15	.6630	.4396	25.175	.000	In:	S120DSIN	.6962

Multiple R .66304
R Square .43962
Adjusted R Square .42216
Standard Error 6.53713

Analysis of Variance

	DF	Sum of Squares	Mean Square
Regression	11	11834.34312	1075.84937
Residual	353	15085.12538	42.73407

F = 25.17545 Signif F = .0000

**** M U L T I P L E R E G R E S S I O N ****

---------- Variables in the Equation ----------

Variable	B	SE B	Beta	Sig T
DEW	.153	.03	.361	.0000
P2	.096	.04	.107	.0175
YSIN	13.344	2.30	1.098	.0000
t8	-.00001	0.00	-18.700	.0000
t10	.00001	0.00	19.500	.0000
S120TSIN	-.175	.06	-.824	.0037
YDEWSIN	-.119	.04	-.444	.0080
SDEWSIN	-.029	.01	-.111	.0092
S120DSIN	.188	.07	.696	.0150
YP2SIN2	.152	.03	.241	.0001
SP2COS	-.079	.02	-.122	.0062
Const	115.18	14.20	.	0000

Piecewise polynomials

A low-order polynomial may provide a poor fit to the data, and increasing the order of the polynomial may not help. Transformations of x or y may solve this problem, but sometimes we may prefer to use more flexible approaches. One such approach is to use **splines**.

- piecewise polynomials used in curve fitting
- polynomials within intervals of x that are connected acoress different intervals of x

The piecewise linear spline function is given by

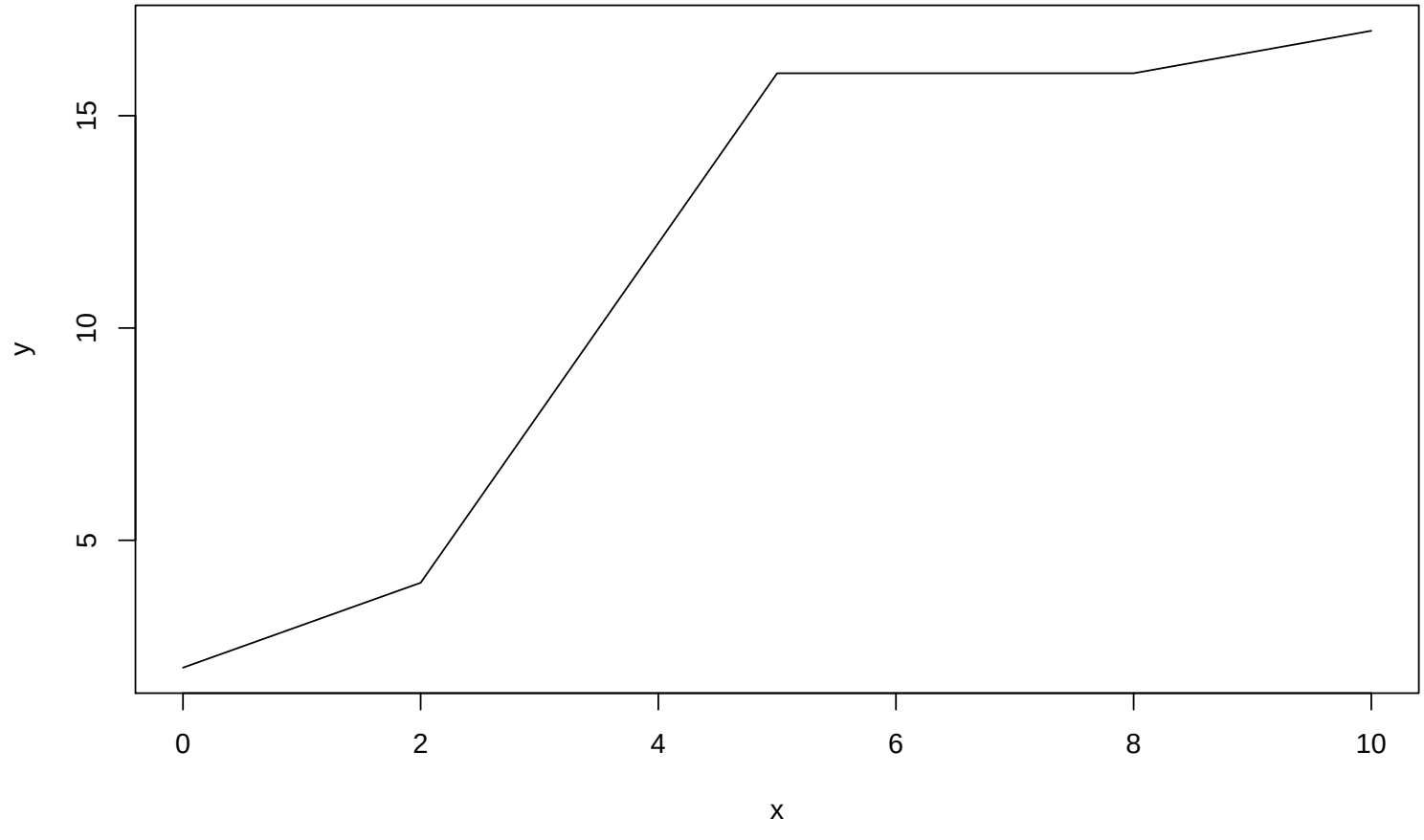
$$f(x) = \beta_0 + \beta_1 x + \beta_2 (x-a)_+ + \beta_3 (x-b)_+ + \beta_4 (x-c)_+,$$

where

$$(u)_+ = \begin{cases} u, & u > 0, \\ 0, & u \leq 0 \end{cases}$$

and a, b and c are referred to as knots.

Example of piecewise linear spline with knots at 2, 5 and 8.



```
if sno ge 2000 sno2000=sno-2000.
if sno <2000 sno2000=0.
REGRESSION
  /MISSING LISTWISE
  /STATISTICS COEFF OUTS CI R ANOVA CHANGE
  /CRITERIA=PIN(.05) POUT(.10)
  /NOORIGIN
  /DEPENDENT dtotal
  /METHOD=enter sno sno2000/save pred(dtotal2).
REGRESSION
  /MISSING LISTWISE
  /STATISTICS COEFF OUTS CI R ANOVA CHANGE
  /CRITERIA=PIN(.05) POUT(.10)
  /NOORIGIN
  /DEPENDENT dtotal
  /METHOD=enter sno /save pred(dtotal_sno).
REGRESSION
  /MISSING LISTWISE
  /STATISTICS COEFF OUTS CI R ANOVA CHANGE
  /CRITERIA=PIN(.05) POUT(.10)
  /NOORIGIN
  /DEPENDENT dtotal
  /METHOD=enter sno2000 /save pred(dtotal_sno2).
GRAPH
  /LINE(SIMPLE)=MEAN( totaltrend  dtotal2 dtotal_sno dtotal_sno2) BY sno.
```

Coefficients[a]

Model		Unstandardized Coefficients		Standardized Coefficients	t	Sig.
		B	Std. Error	Beta		
1	(Constant)	41.907	.481		87.128	.000
	sno	-.004	.000	-.340	-10.519	.000
	sno2000	.006	.001	.222	6.863	.000

a. Dependent Variable: dtotal

Coefficients[a]

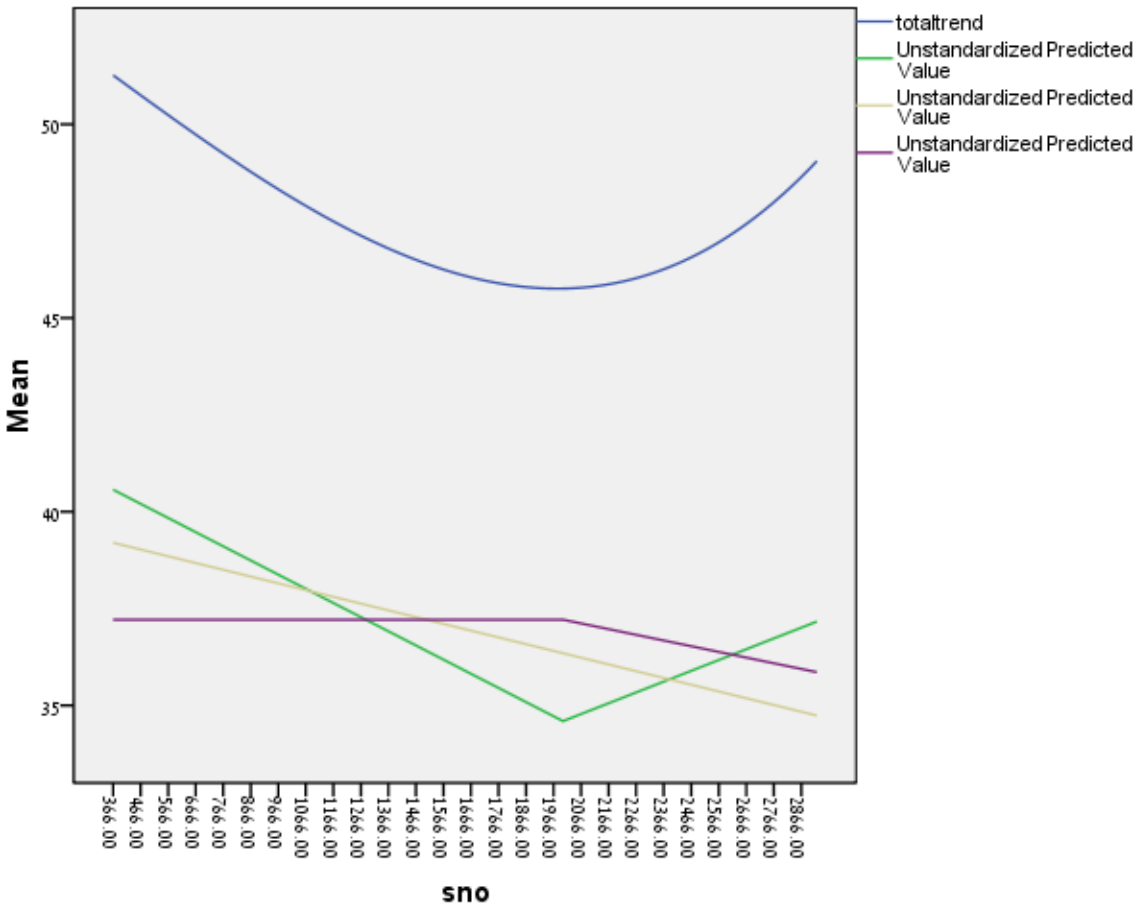
Model		Unstandardized Coefficients		Standardized Coefficients	t	Sig.
		B	Std. Error	Beta		
1	(Constant)	39.838	.378		105.357	.000
	sno	-.002	.000	-.162	-8.316	.000

a. Dependent Variable: dtotal

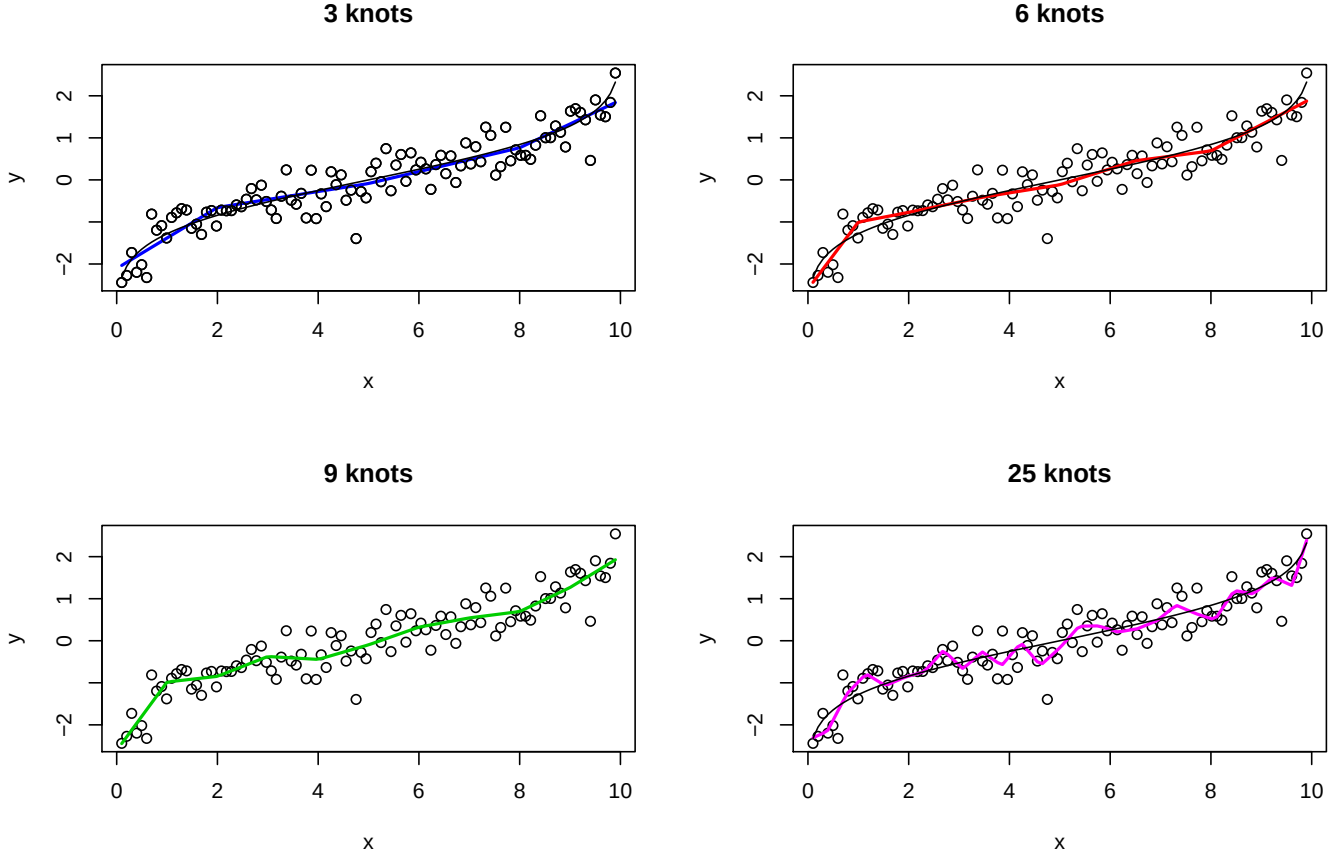
Coefficients[a]

Model		Unstandardized Coefficients		Standardized Coefficients	t	Sig.
		B	Std. Error	Beta		
1	(Constant)	37.214	.184		202.751	.000
	sno2000	-.001	.001	-.051	-2.561	.010

a. Dependent Variable: dtotal


totaltrend
Unstandardized Predicted Value
Unstandardized Predicted Value
Unstandardized Predicted Value
Mean
50
45
40
35
366.00
466.00
566.00
666.00
766.00
866.00
966.00
1066.00
1166.00
1266.00
1366.00
1466.00
1566.00
1666.00
1766.00
1866.00
1966.00
2066.00
2166.00
2266.00
2366.00
2466.00
2566.00
2666.00
2766.00
2866.00
sno

As we increase the number of knots, the piecewise linear polynomial more closely resembles a continuous line.



Cubic splines

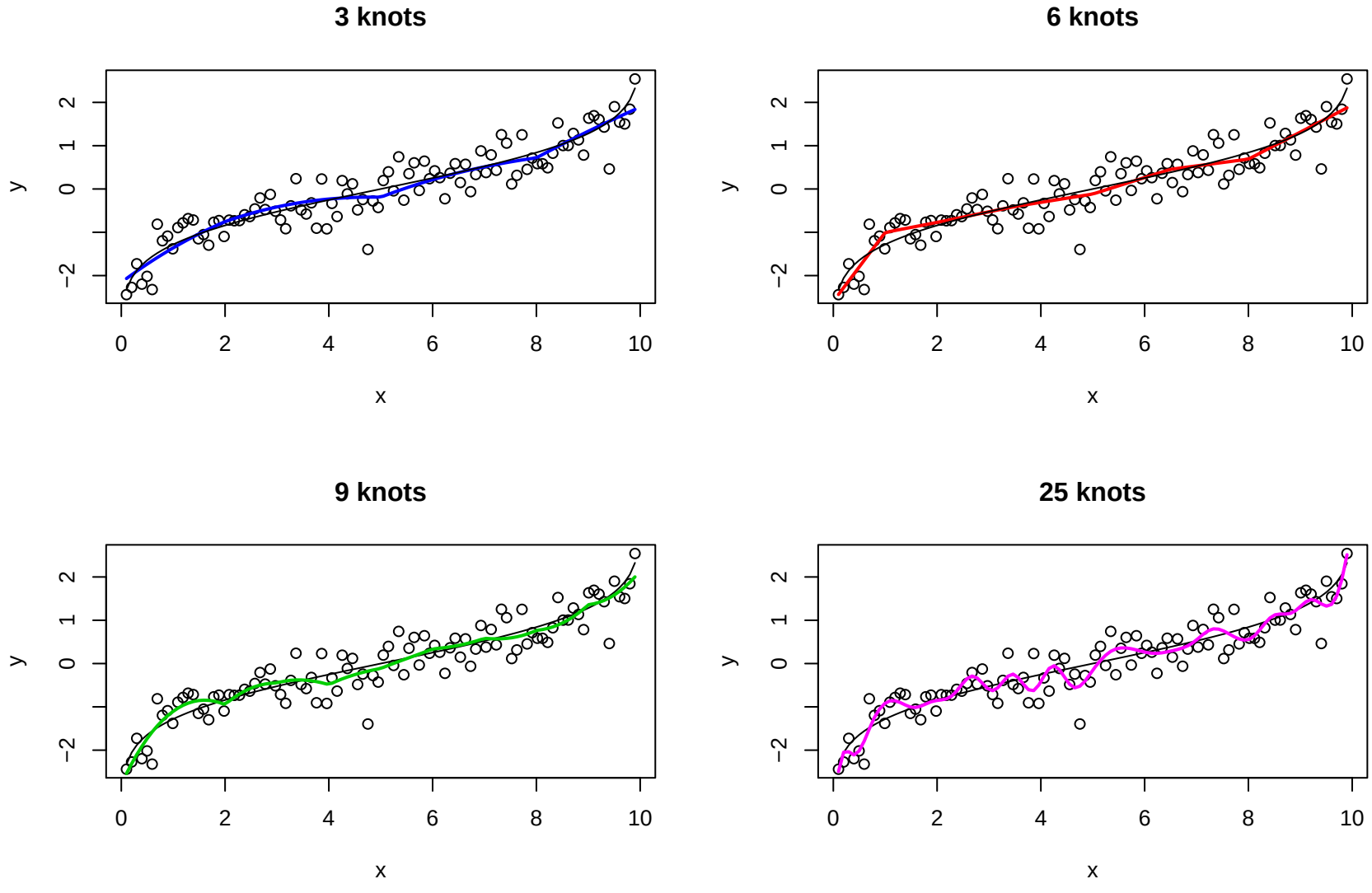
Although, linear splines may work well, they are not smooth and will not fit highly curved functions well (unless many knots are used - which requires a lot of data).

It is more common for cubic splines to be used in practice. A cubic spline function with k knots is given by

$$f(x) = \sum_{j=0}^{3} \beta_{0j} x^j + \sum_{l=1}^{k} \beta_i (x - t_l)_+^3,$$

where $t_l,\ l = 1, \ldots, k$ are the k knots. We relate x to the outcome as

$$y_i = f(x_i) + \epsilon_i.$$


3 knots
6 knots
9 knots
25 knots
y
x
−2
0
1
2
0
2
4
6
8
10

For estimation purposes, we assume that both the locations and the number of knots are fixed. Although there are methods that allow the number and/or position of the knots to be random; these models are too complex to be fit using least squares.

The piecewise cubic splines may give us a more flexible model, but they still may be discontinuous at the knots.

Continuous cubic splines

Cubic B splines

- Given k knots at $t_1, \ldots, t_k$, a cubic B spline function is a cubic polynomial on the interval $[t_j, t_{j+1}]$,

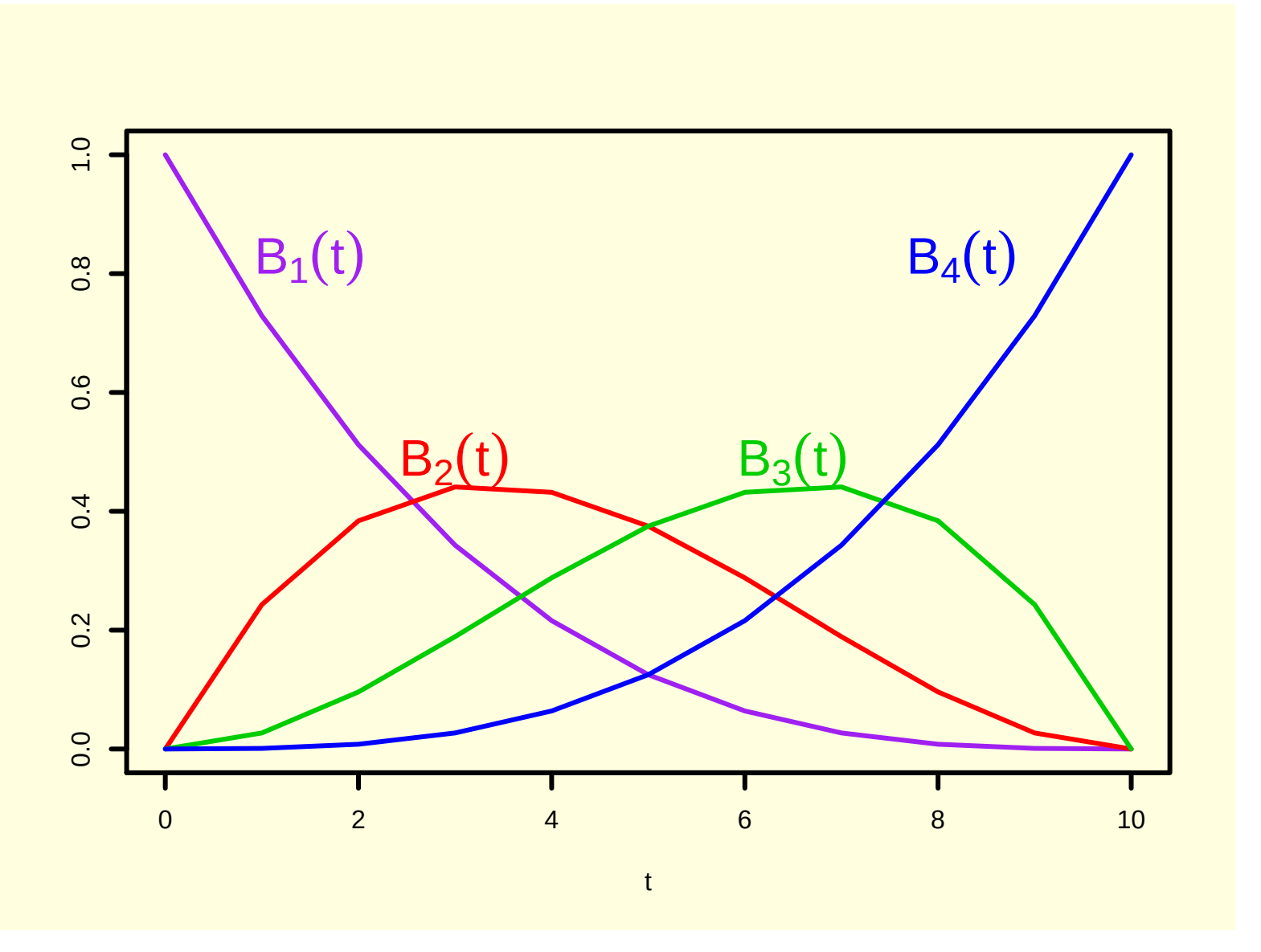
- It has continuous first and second derivatives, imposing 3 conditions at each knot.

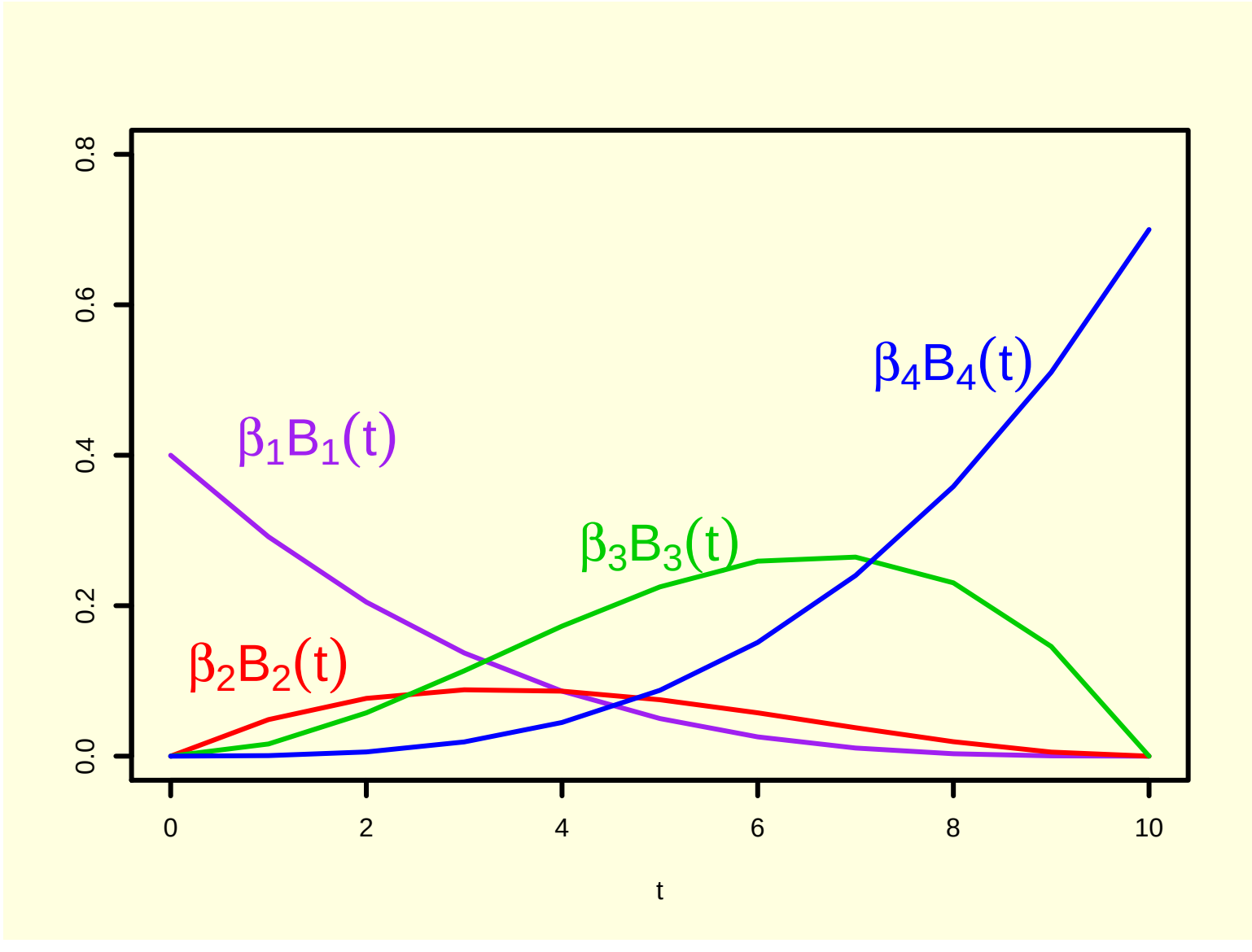
- With k knots, $k + 1$ parameters are needed to represent the cubic spline.

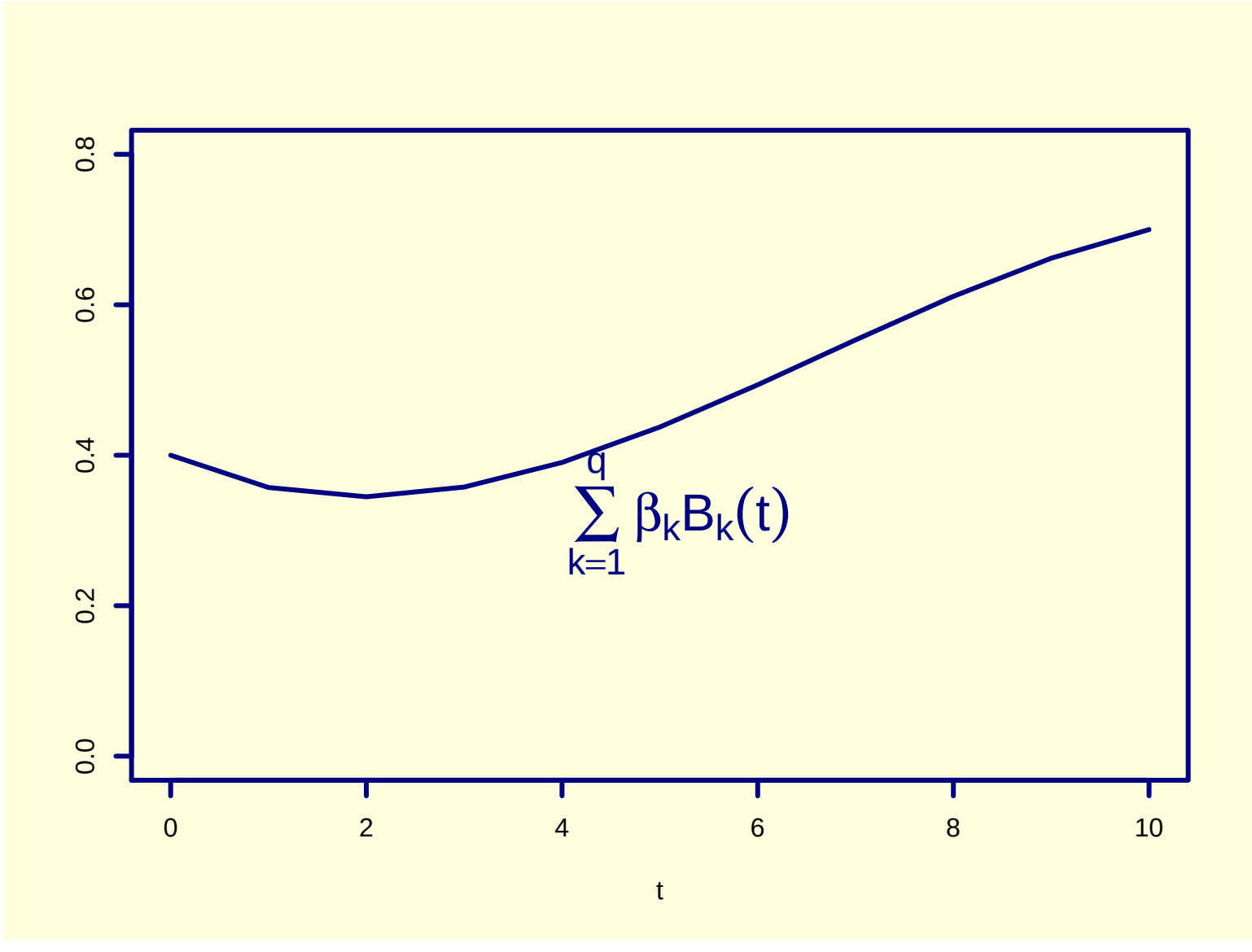
A cubic B-spline function with k knots is given by

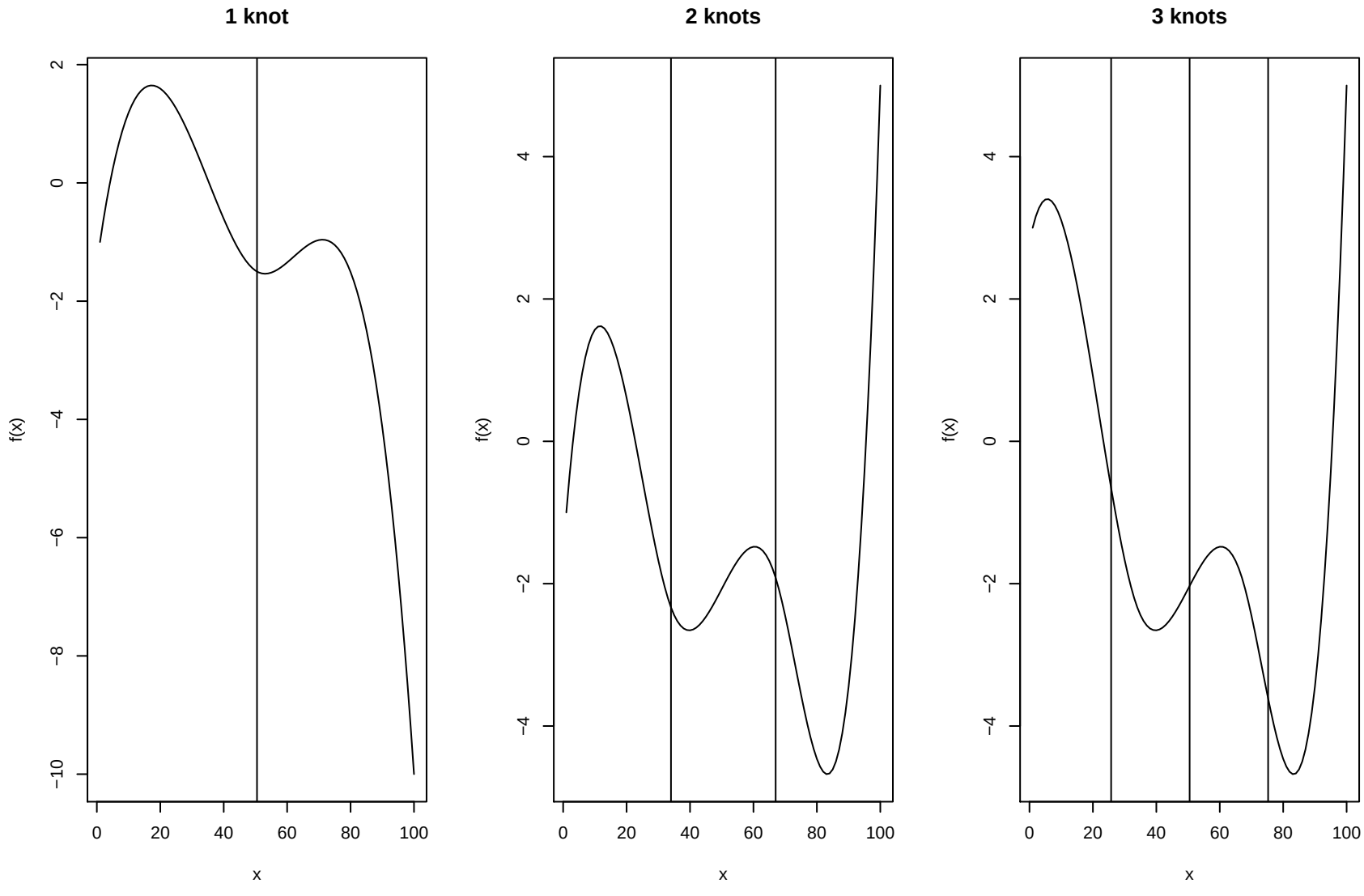
$$f(x) = \sum_{i=1}^{k+4} \beta_k B_k(x),$$

where $B_k(x)$ is the kth B-spline basis function.


B1(t)
B2(t)
B3(t)
B4(t)
0.0
0.2
0.4
0.6
0.8
1.0
0
2
4
6
8
10
t


0.0
0.2
0.4
0.6
0.8
0
2
4
6
8
10
t
β1B1(t)
β2B2(t)
β3B3(t)
β4B4(t)

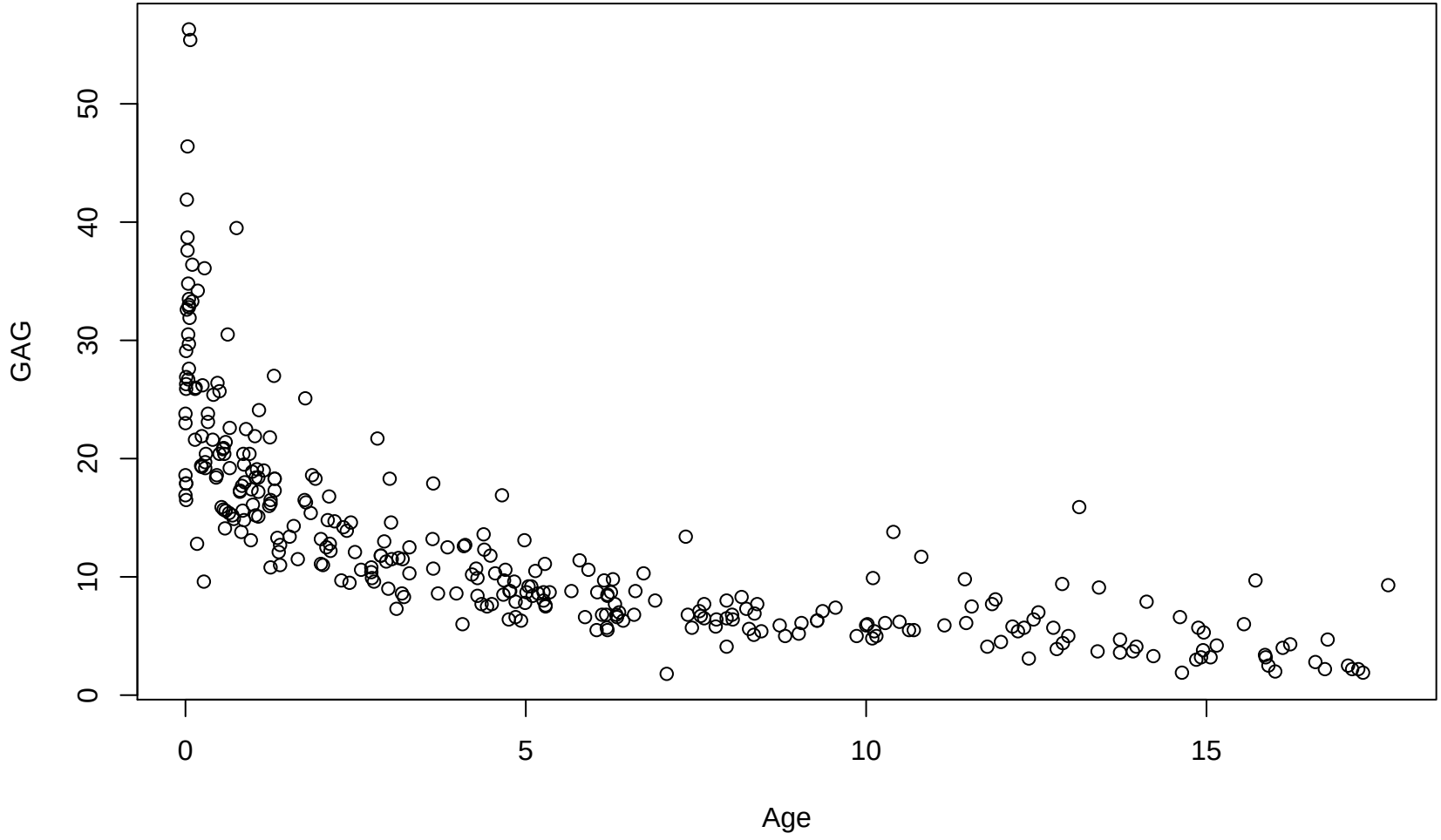

0.0
0.2
0.4
0.6
0.8
0
2
4
6
8
10
t
$\sum_{k=1}^{q} \beta_k B_k(t)$

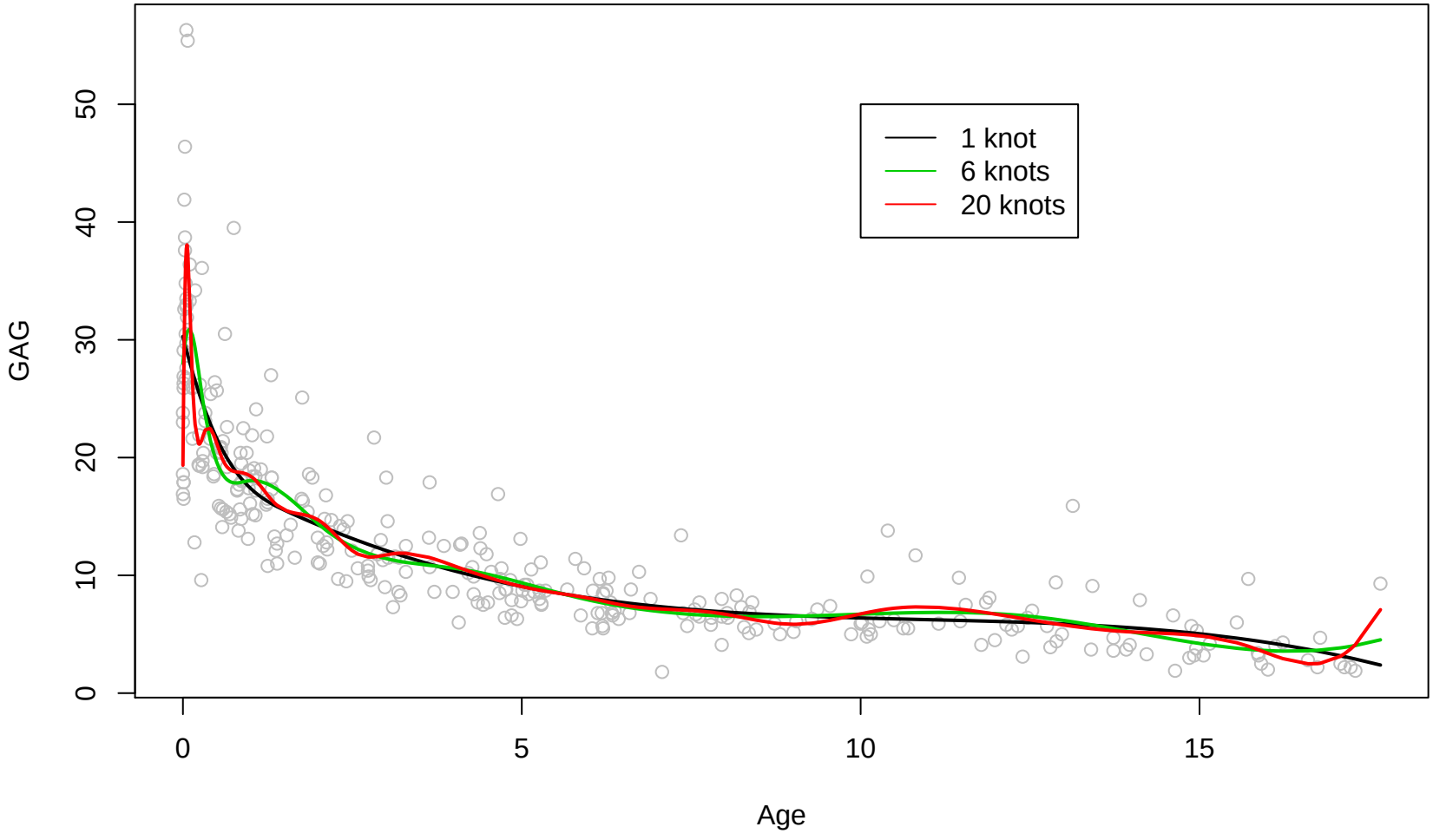

1 knot
2 knots
3 knots
f(x)
x
0
20
40
60
80
100
2
0
−2
−4
−6
−8
−10
4

Example

Venables and Ripley provide a data set, GAGurine, in the MASS library. It is described as follows:

Data were collected on the concentration of a chemical GAG in the urine of 314 children aged from zero to seventeen years. The aim of the study was to produce a chart to help a paediatrican to assess if a child's GAG concentration is "normal".


GAG
0
10
20
30
40
50
0
5
10
15
Age


GAG
0
10
20
30
40
50
0
5
10
15
Age
1 knot
6 knots
20 knots

Choosing the number and position of knots

- Knots are usually placed at quantiles of the data or at regularly spaced intervals.
- Choosing the number, rather than the placement, seems to be more crucial to the fit.
- Therefore choose a number of knots that represents the curvature you believe to be present in the data. This comes with experience.
- You may also want to place knots at points in the data where you expect significant changes in the relationship between the predictor and the outcome to occur.