




Inspire…Educate…Transform.

Supervised models

Logistic Regression, Time Series Forecasting

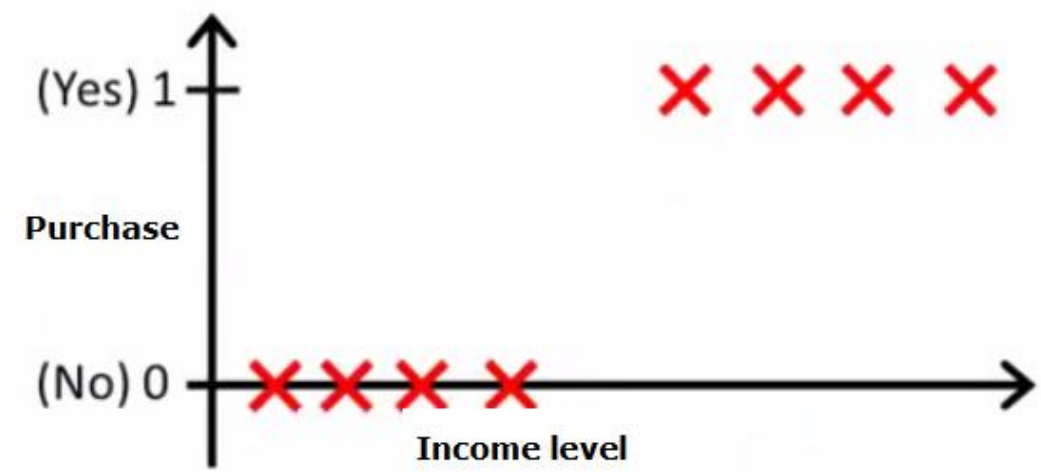
Dr. Sridhar Pappu

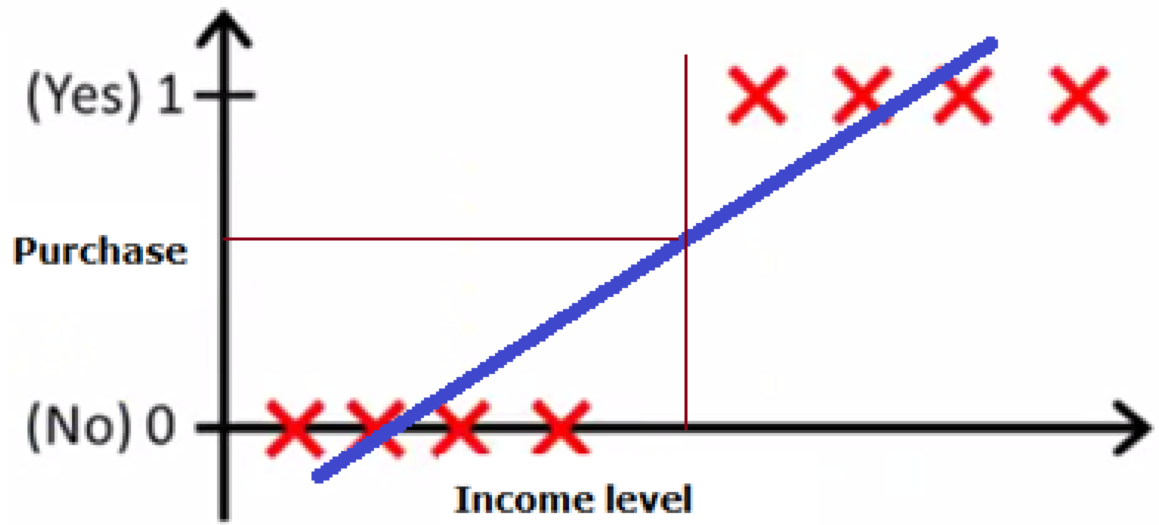
Executive VP – Academics, INSOFE

July 05, 2015

LOGISTIC REGRESSION

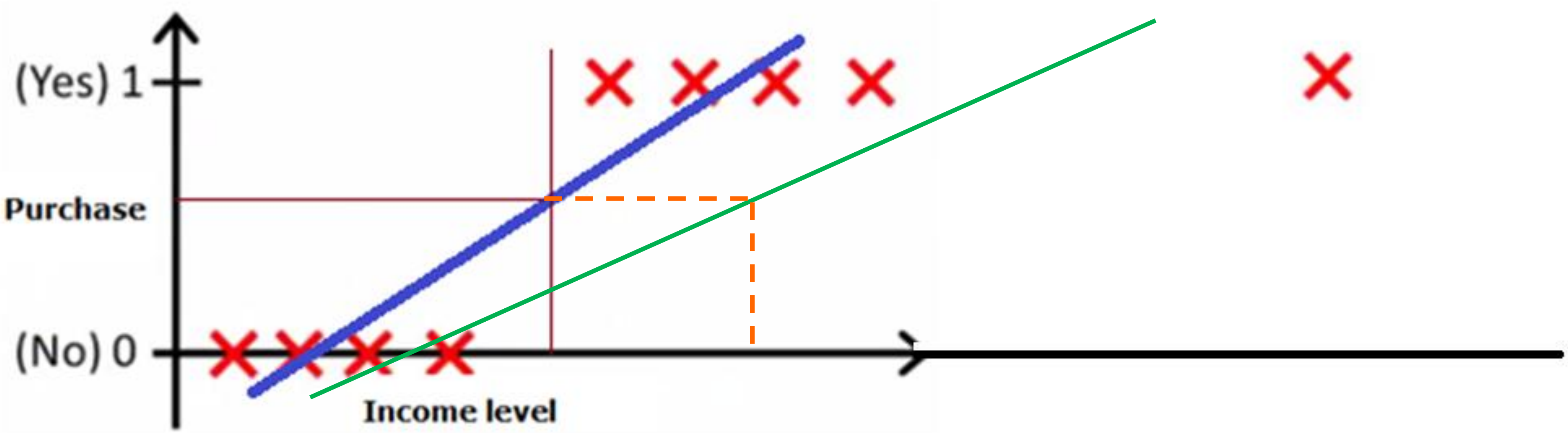
Classification tasks: Regression





It could fail

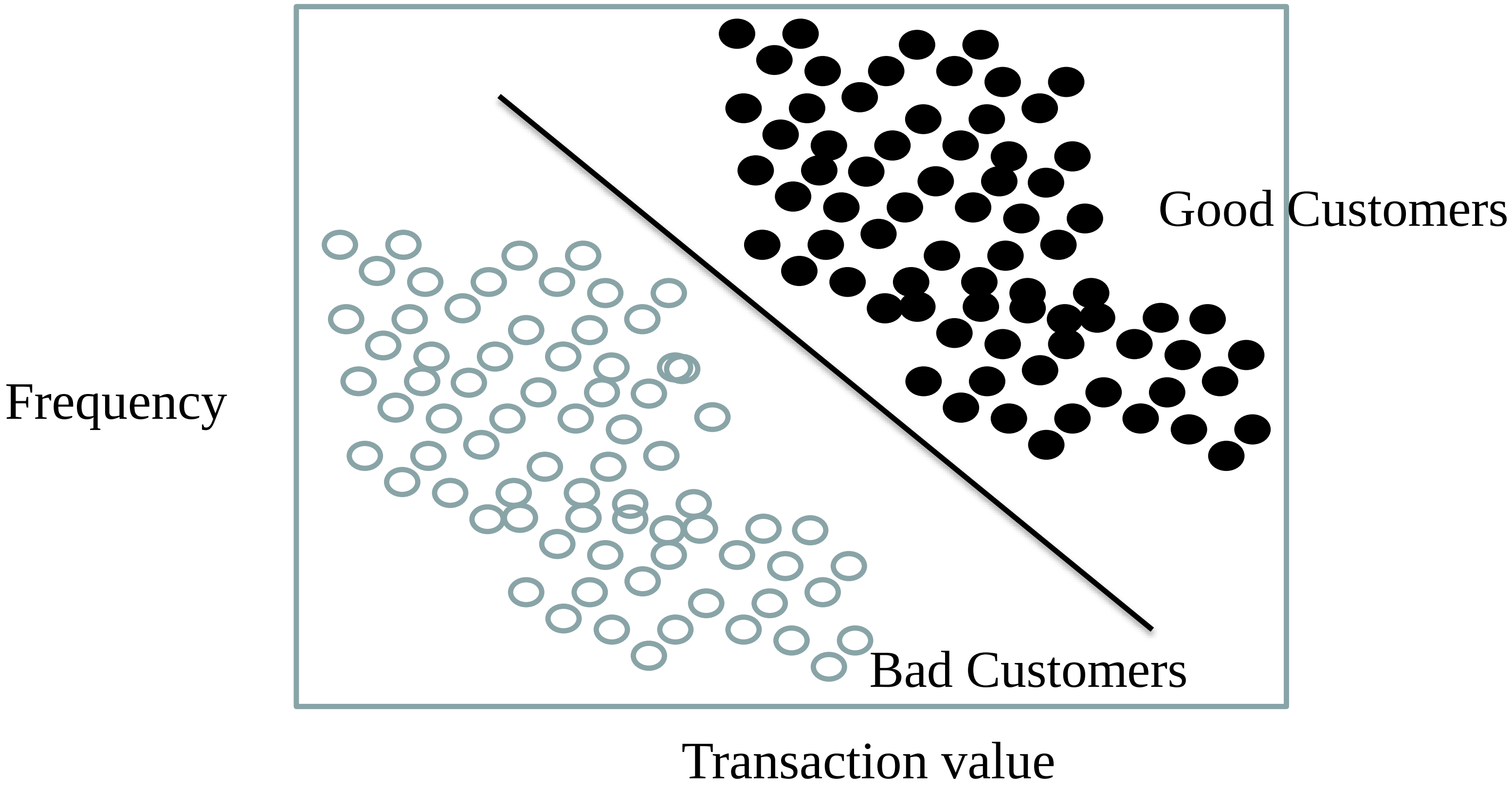
Least squares is not a good cost function



- In addition, linear regression hypothesis can be much larger than 1 or much smaller than zero and hence thresholding becomes difficult.

- Error terms do not follow normal distribution.
- Error terms are not independent.
- Error variances are heteroscedastic.

Logistic Regression

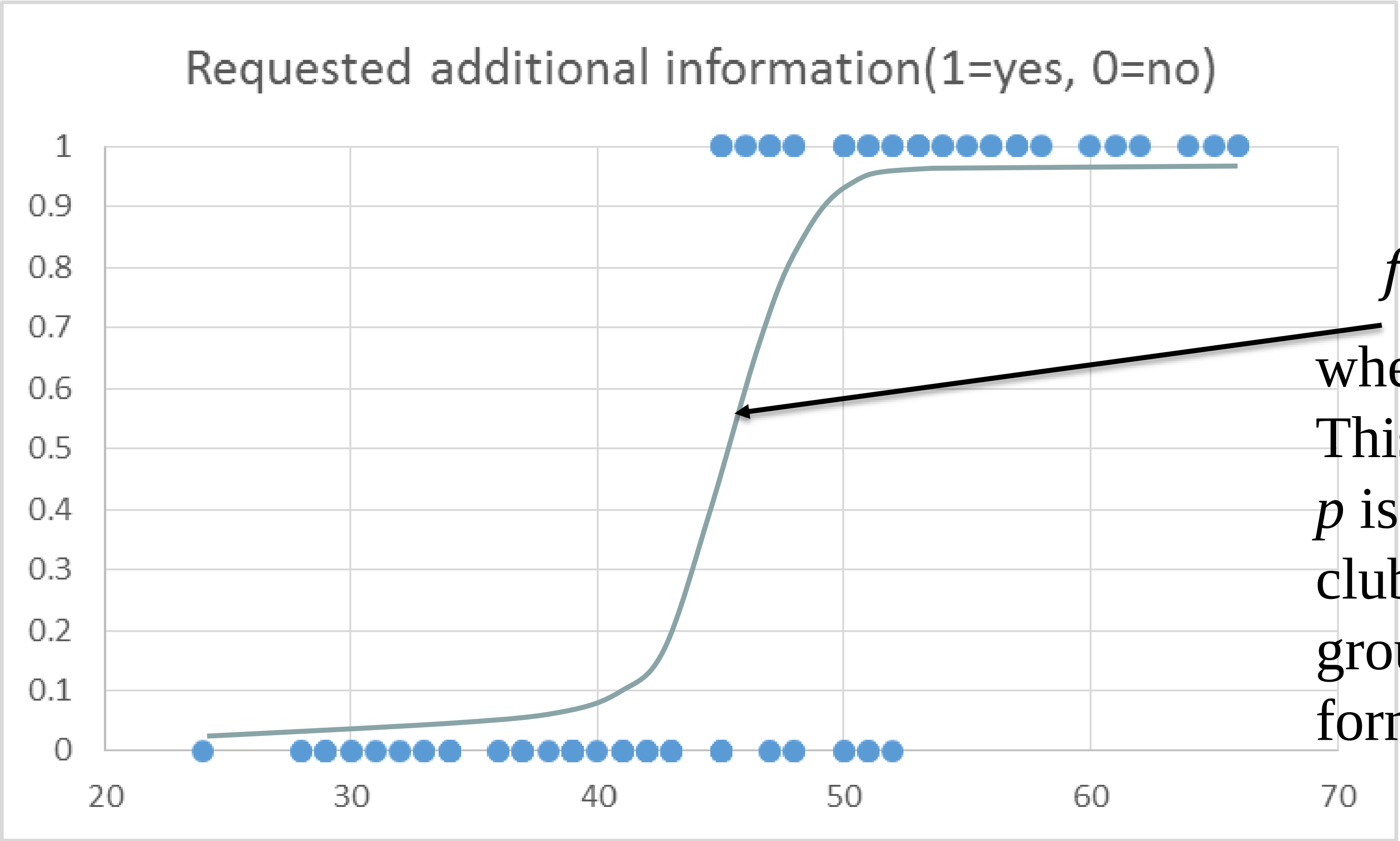


Example

An auto club mails a flier to its members offering to send more information regarding a supplemental health insurance plan if the member returns a brief enclosed form.

Can a model be built to predict if a member will return the form or not?

Example



$$f(x) = p = \frac{e^{\mu}}{1 + e^{\mu}}$$

where $\mu = \beta_0 + \beta_1 x_1$

This is a logistic model. p is the probability that a club member fits into group 1 (returns the form; success).

Logistic model

$$f(x) = p = \frac{e^{\beta_0 + \beta_1 x_1 + \cdots + \beta_k x_k}}{1 + e^{\beta_0 + \beta_1 x_1 + \cdots + \beta_k x_k}}$$

Odds Ratio is obtained by the probability of an event occurring divided by the probability that it will not occur.

Logistic model can be transformed into an odds ratio:

$$S = Odds\ ratio\ or\ Odds = \frac{p}{1-p}$$

Attention Check – Probability and Odds

If the probability of winning is 6/12, what are the odds of winning?

1:1 (Note, the probability of losing also is 6/12)

If the odds of winning are 13:2, what is the probability of winning?

13/15

If the odds of winning are 3:8, what is the probability of losing?

8/11

If the probability of losing is 6/8, what are the odds of winning?

2:6 or 1:3

Attention Check – Probability and Odds

Sign Up FREE BETS	bet365	Sky Bet	totesport	BoyleSports	Betfred	sportingbet	BetVictor	Paddy Power	StanJames	888sport	Ladbrokes	Coral	William Hill	winner	betfair	betway	BetBright	Titanbet	Unibet	bwin	32Red	betfair exchange	BETDAQ	Matchbook
	£200	£50	£25	£50	£30	£50	£25	£50	£10	£x3	£25	£20		£200	£30	£50	£30	£50	£20	£30	£10	£20	£20	£20
Novak Djokovic	11/10	6/5	11/10	5/4	11/10	6/5	11/10	6/5	6/5	6/5	5/4	11/10	6/5	6/5	6/5	6/5		6/5	6/5	6/5	6/5	5/4	23/20	13/10
Andy Murray	2	2	2	12/5	2	5/2	2	9/4	11/5	5/2	2	2	9/4	9/4	12/5	2		9/4	5/2	2	5/2	12/5	23/10	12/5
Roger Federer	11/2	9/2	13/2	6	13/2	6	13/2	6	13/2	6	13/2	13/2	6	6	13/2	6		6	6	11/2	6	32/5	33/5	33/5
Stan Wawrinka	11	12	10	12	10	12	12	10	11	11	10	12	10	11	12	11		11	11	11	11	64/5	12	59/5
Tomas Berdych	33	33	33	33	33	20	33	33	40	34	40	33	33	40	40	33		40	34	33	34	51	49	54
Milos Raonic	33	40	25	33	25	40	28	33	40	34	40	28	25	40	33	33		40	34	33	34	66	63	64
Grigor Dimitrov	33	40	33	50	33	40	40	33	50	44	50	50	50	50	40	33		50	44	33	44	61	59	59
Jo-Wilfried Tsonga	50	40	66	40	66	40	40	50	66	60	66	50	50	50	40	50		50	60	50	60	89	86	89
Nick Kyrgios	50	80	66	50	66	66	66	50	66	60	66	50	66	50	66	50		50	60	50	60	94	96	99
Marin Cilic	50	50	50	66	50	66	50	50	50	90	50	50	50	66	40	50		66	90	50	90	113	108	99
Dustin Brown	100	100	66	80	66	125	80		125	150	66	80	100	100	125	100		100	150		150	142	166	
Kevin Anderson	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]		[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]

Disclaimer: Gambling/Betting is injurious to financial health. Dr. Sridhar or INSOFE do not endorse this addiction, and this has been explained only for educational purposes.

Source: http://www.oddschecker.com/tennis/wimbledon/mens/winner
Last accessed: July 03, 2015

Logistic model

$$S = Odds\,ratio = \frac{p}{1-p}$$

$$S = \frac{\dfrac{e^{\beta_0 + \beta_1 x_1 + \cdots + \beta_k x_k}}{1 + e^{\beta_0 + \beta_1 x_1 + \cdots + \beta_k x_k}}}{1 - \dfrac{e^{\beta_0 + \beta_1 x_1 + \cdots + \beta_k x_k}}{1 + e^{\beta_0 + \beta_1 x_1 + \cdots + \beta_k x_k}}}$$

$$\therefore, S = e^{\beta_0 + \beta_1 x_1 + \cdots + \beta_k x_k}$$

$$\ln(S) = \ln\left(e^{\beta_0 + \beta_1 x_1 + \cdots + \beta_k x_k}\right)$$

$$= \beta_0 + \beta_1 x_1 + \cdots + \beta_k x_k$$

Logistic model

The log of the odds ratio is called logit, and the transformed model is linear in βs.

Interpreting the output

Binary Logistic Regression: Ask for Insurance Information versus Age							
Response Information							
			Value	Count			
Variable			1	36	(Event)		
Ask for Insurance Information			0	56			
			Total	92			
Logistic Regression Table							
					Odds	95%	CI
Predictor	Coef	SE Coef	Z	P	Ratio	Lower	Upper
Constant	-20.754	4.61715	-4.49	0.000			
Age	0.43368	0.096946	4.47	0.000	1.54	1.28	1.87
Log-Likelihood = -24.708							
Test that all slopes are zero: G = 73.739, DF = 1, P-Value = 0.000							

What is the logit equation?

$$\ln(S) = -20.754 + 0.43368Age$$

Determining Logistic Regression Model

Suppose we want a probability that a 50-year old club member will return the form.

$\ln(S) = -20.754 + 0.43368 * 50 = 0.93$

$S = e^{0.93} = 2.535$

The odds that a 50-year old returns the form are 2.535 to 1.

Determining Logistic Regression Model

$$\hat{p} = \frac{S}{S+1} = \frac{2.535}{2.535+1} = 0.7171$$

Using a probability of 0.50 as a cutoff between predicting a 0 or a 1, this member would be classified as a 1.

Testing the Overall Model – G Statistic

In regression analysis, *F* test was used. Here, *G* statistic is used.

G = 2{[log likelihood with variable] – [log likelihood without variable]}

Log likelihood without variable = $ln\left[\left(\frac{n_0}{N}\right)^{n_0}\left(\frac{n_1}{N}\right)^{n_1}\right]$,

where n_0 is the # of "0" observations, n_1 is the # of "1" observations and N is the total # of observations.

Testing the Overall Model – G Statistic

Log likelihood without variable = $ln\left[\left(\frac{n_0}{N}\right)^{n_0}\left(\frac{n_1}{N}\right)^{n_1}\right] = -61.578$

G = 2{[-24.708] – [-61.578]} = 73.74

p-value indicates overall significance of the model.

Likewise, the z-values and the associated p-values provide significance of individual predictor variables.

Testing the Overall Model - AIC

R outputs AIC (Akaike's Information Criterion) and you need to pick the model with the lowest AIC. Below is a sample output:

```
Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept)  -15.834      4.923  -3.216 0.001298 **
logconc        5.578      1.680   3.319 0.000902 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for binomial family taken to be 4.08)

    Null deviance: 83.631  on 8  degrees of freedom
Residual deviance: 29.346  on 7  degrees of freedom
AIC: 62.886
```

Testing the Overall Model - AIC

- $AIC = -2 * \ln(likelihood) + 2 * k$

  where *k* is the number of parameters in the model including the constant and the error.
- AIC provides a means for model selection.
- It does not test a model in the sense of null hypothesis and hence doesn't tell anything about the quality of the model. **It is only a relative measure between multiple models.**

Intuition

- Overly large coefficient magnitudes, overly large error bars on the coefficient estimates, and the wrong sign on a coefficient could be indications of correlated inputs.

Applications

- Predicting stock price movement (up/down)
- Predict whether a patient has diabetes or not
- Predict whether a customer will buy or not
- Predict the likelihood of loan default

TIME SERIES FORECASTING

Why time series

- Causal independent variables are
  - Unknown to us
  - Not available
  - Might not fit the data well
  - Difficult to forecast

Typical time series

$$\hat{y}_{t+1} = f(y_t, y_{t-1}, y_{t-2} \cdots) + f(x_1, x_2, x_3 \dots)$$

f can be linear or nonlinear

IMPORTANT CONCEPTS

Autocorrelation (ACF) and Partial ACF (PACF)

- ACF: n^{th} lag of ACF is the correlation between a day and n days before that.

- PACF: The same as ACF with all intermediate correlations removed. It is the k_{th} coefficient of the ordinary least squares regression.

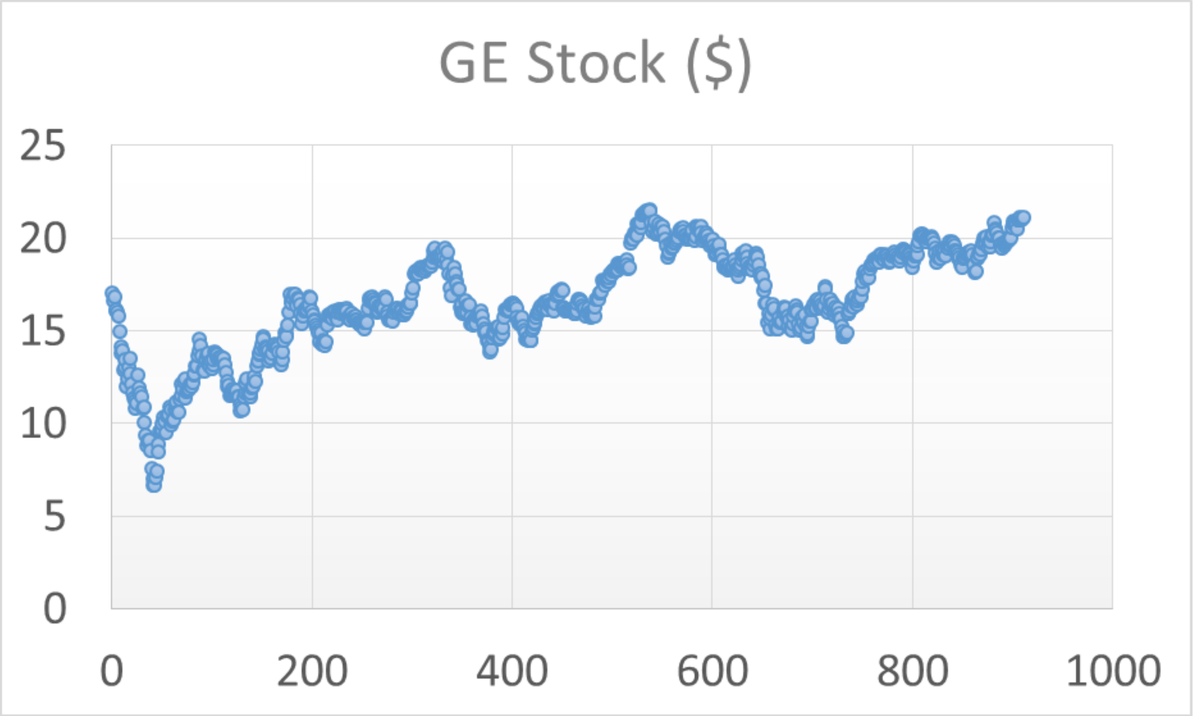
$$[y_t] = \beta_0 + \sum_{i=1}^{k} \beta_i [y_{t-i}] \text{ where}$$

$[y_t]$ is the input time series, k is the lag order and β_i is the i_{th} coefficient of the linear multiple regression.

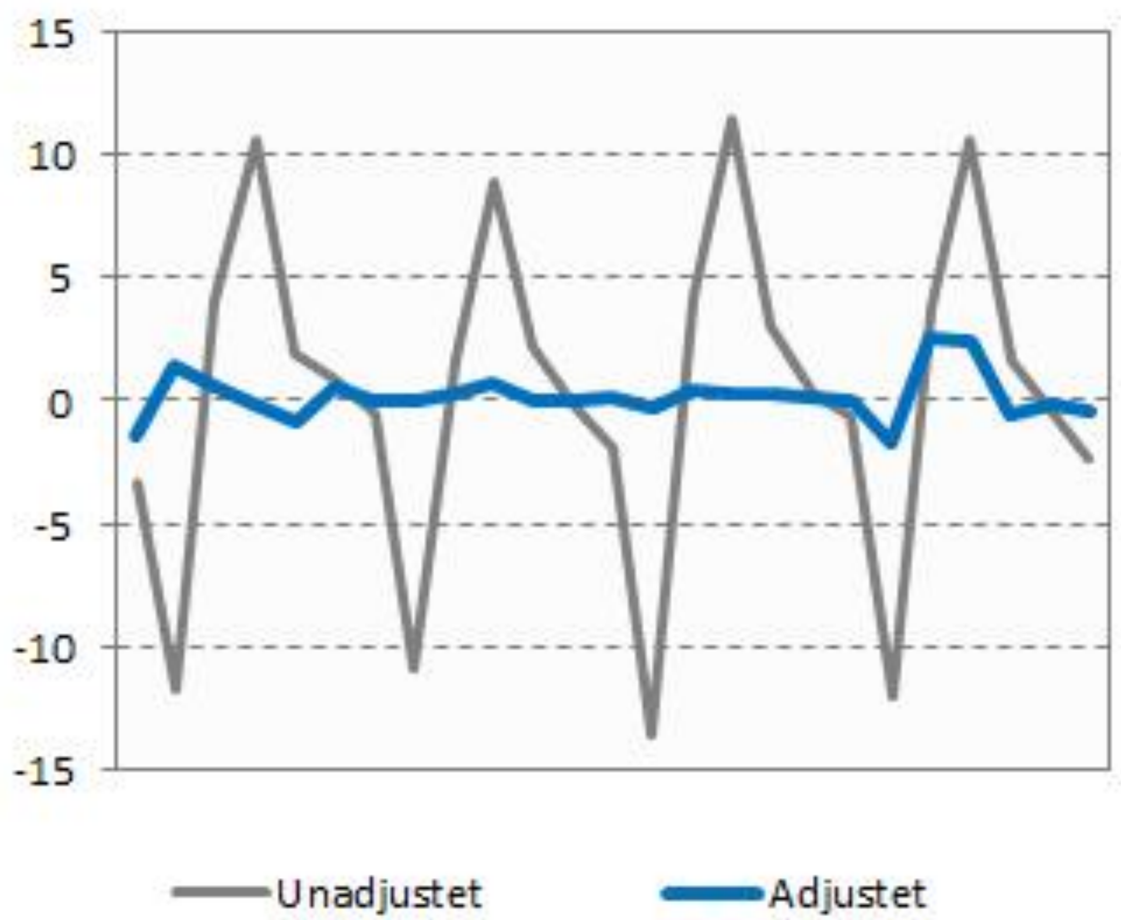
EXCEL ACTIVITY

Components of time series

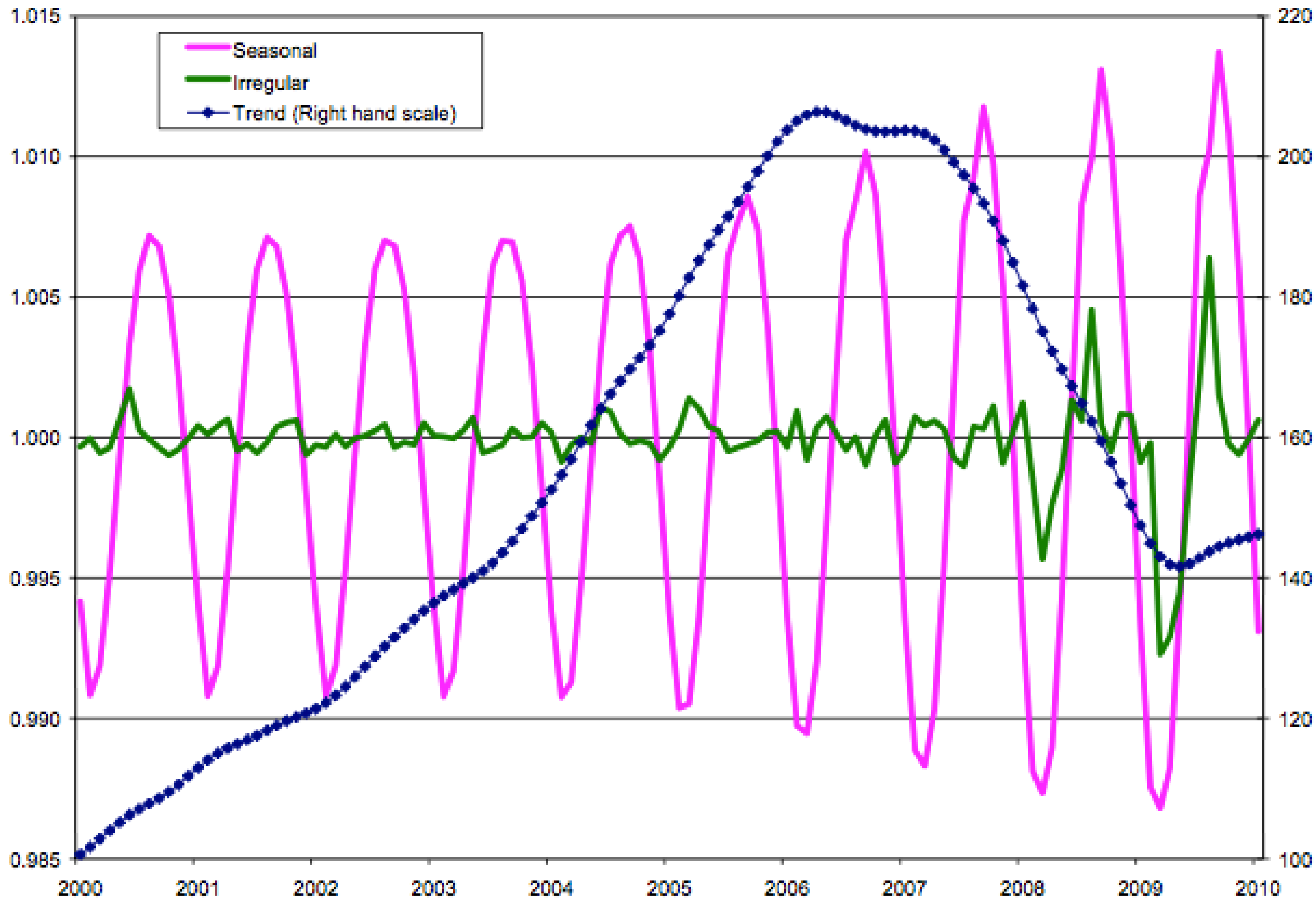
- Trend
- Seasonality/Cyclicality
- Random component



Seasonality



Trend, seasonality and randomness

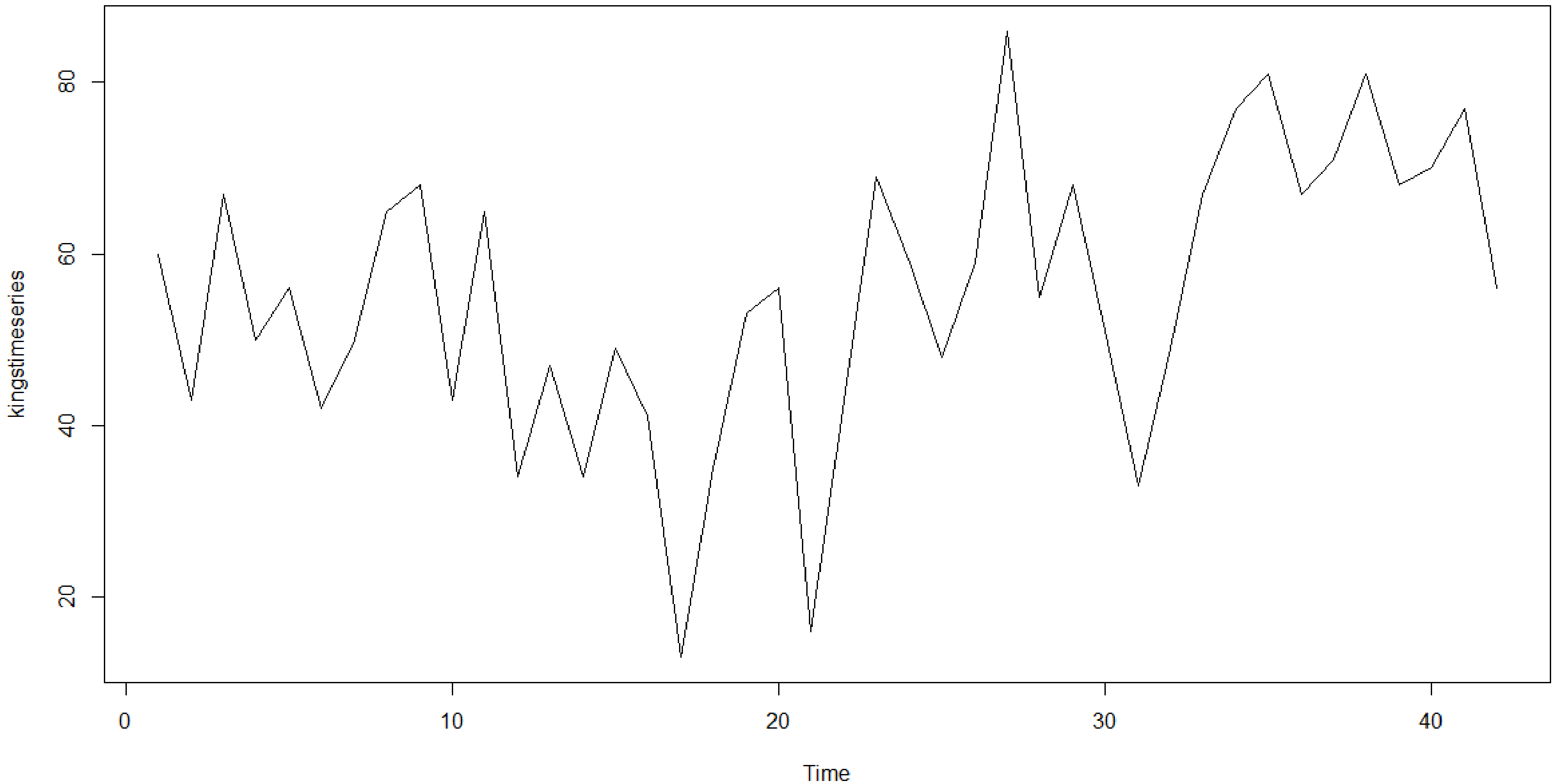


Trend, seasonality and randomness

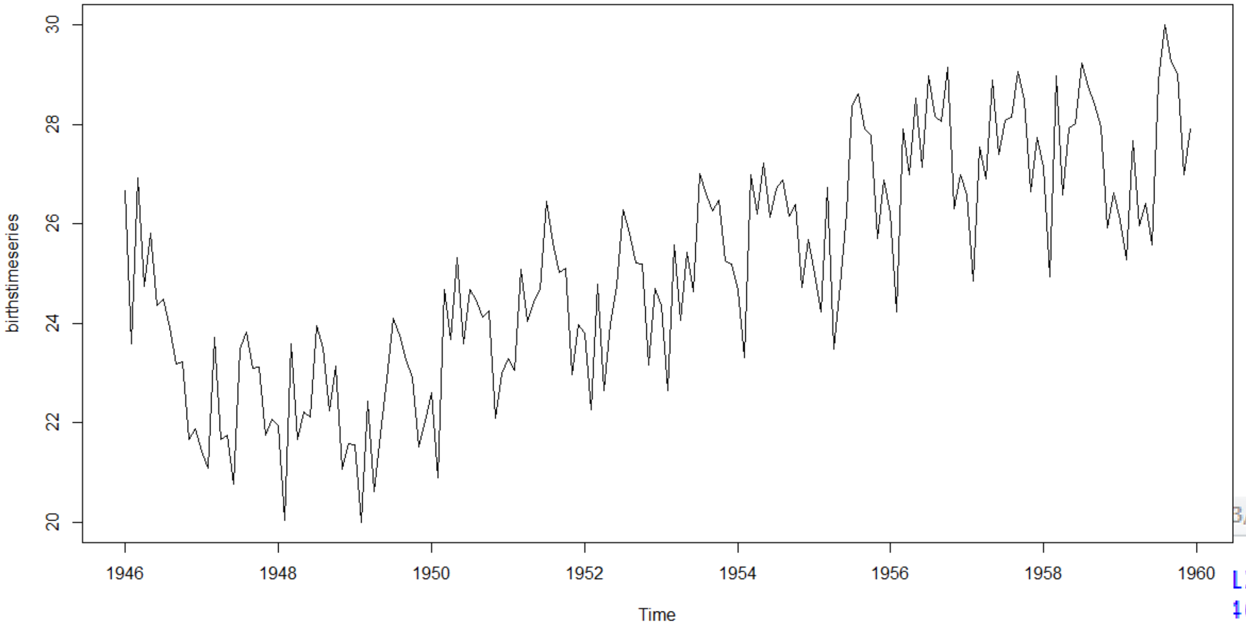
- They vary significantly

Trend, seasonality and randomness

```
> kingstimeseries <- ts(kings)
> kingstimeseries
Time Series:
Start = 1
End = 42
Frequency = 1
 [1] 60 43 67 50 56 42 50 65 68 43 65 34 47 34 49 41 13 35 53 56 16 43 69 59 48 59 86 55 68 51 33 49 67 77
[35] 81 67 71 81 68 70 77 56
```



Trend, seasonality and randomness



3/Batch 12/CSE 7202c/Day03_20150705/TimeSeries/

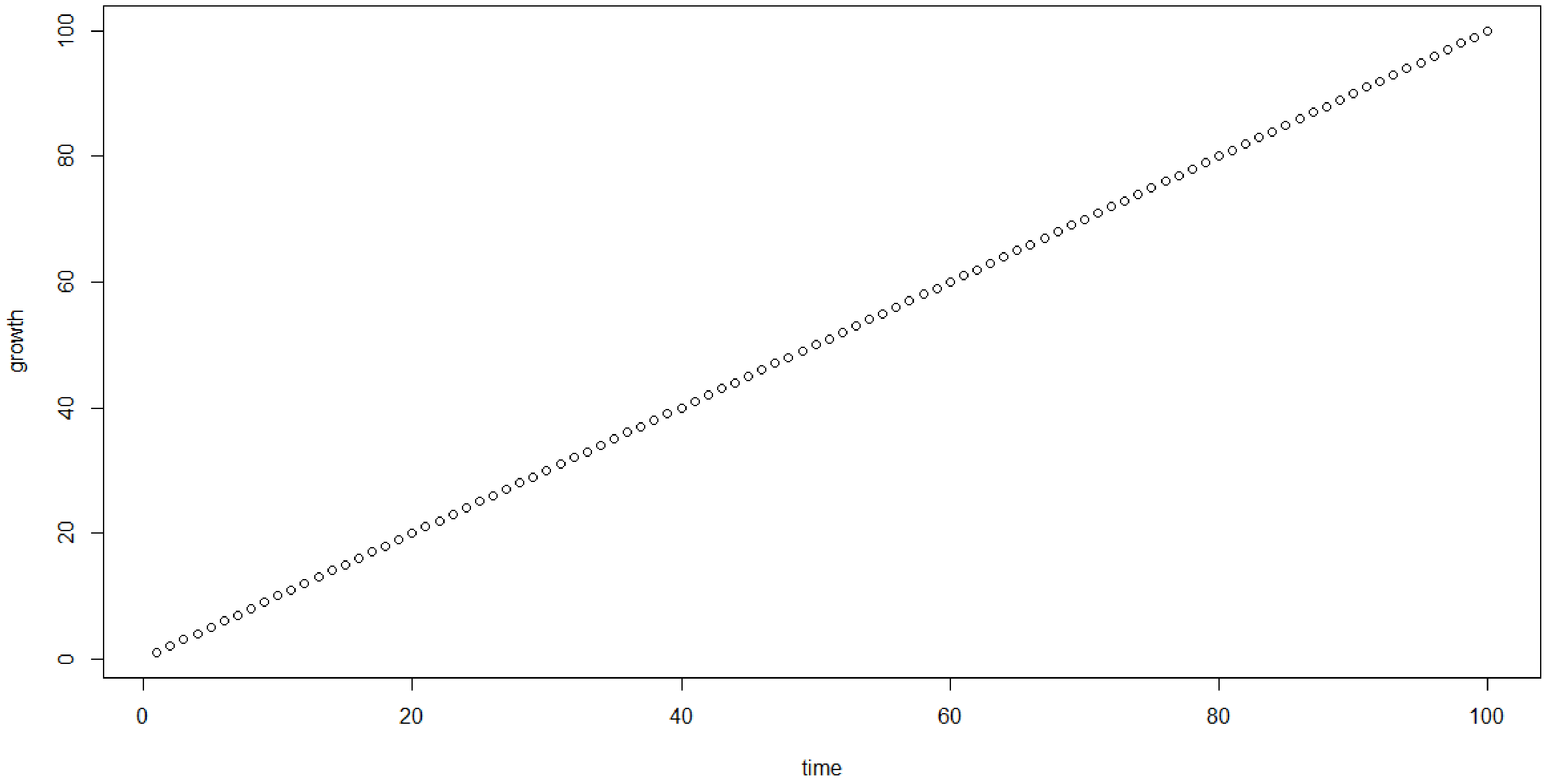
```
12,
46,1))
> birthstimeseries
```

	Jan	Feb	Mar	Apr	May	Jun	Jul	Aug	Sep	Oct	Nov	Dec
1946	26.663	23.598	26.931	24.740	25.806	24.364	24.477	23.901	23.175	23.227	21.672	21.870
1947	21.439	21.089	23.709	21.669	21.752	20.761	23.479	23.824	23.105	23.110	21.759	22.073
1948	21.937	20.035	23.590	21.672	22.222	22.123	23.950	23.504	22.238	23.142	21.059	21.573
1949	21.548	20.000	22.424	20.615	21.761	22.874	24.104	23.748	23.262	22.907	21.519	22.025
1950	22.604	20.894	24.677	23.673	25.320	23.583	24.671	24.454	24.122	24.252	22.084	22.991
1951	23.287	23.049	25.076	24.037	24.430	24.667	26.451	25.618	25.014	25.110	22.964	23.981
1952	23.798	22.270	24.775	22.646	23.988	24.737	26.276	25.816	25.210	25.199	23.162	24.707
1953	24.364	22.644	25.565	24.062	25.431	24.635	27.009	26.606	26.268	26.462	25.246	25.180
1954	24.657	23.304	26.982	26.199	27.210	26.122	26.706	26.878	26.152	26.379	24.712	25.688
1955	24.990	24.239	26.721	23.475	24.767	26.219	28.361	28.599	27.914	27.784	25.693	26.881
1956	26.217	24.218	27.914	26.975	28.527	27.139	28.982	28.169	28.056	29.136	26.291	26.987
1957	26.589	24.848	27.543	26.896	28.878	27.390	28.065	28.141	29.048	28.484	26.634	27.735
1958	27.132	24.924	28.963	26.589	27.931	28.009	29.229	28.759	28.405	27.945	25.912	26.619
1959	26.076	25.286	27.660	25.951	26.398	25.565	28.865	30.000	29.261	29.012	26.992	27.897

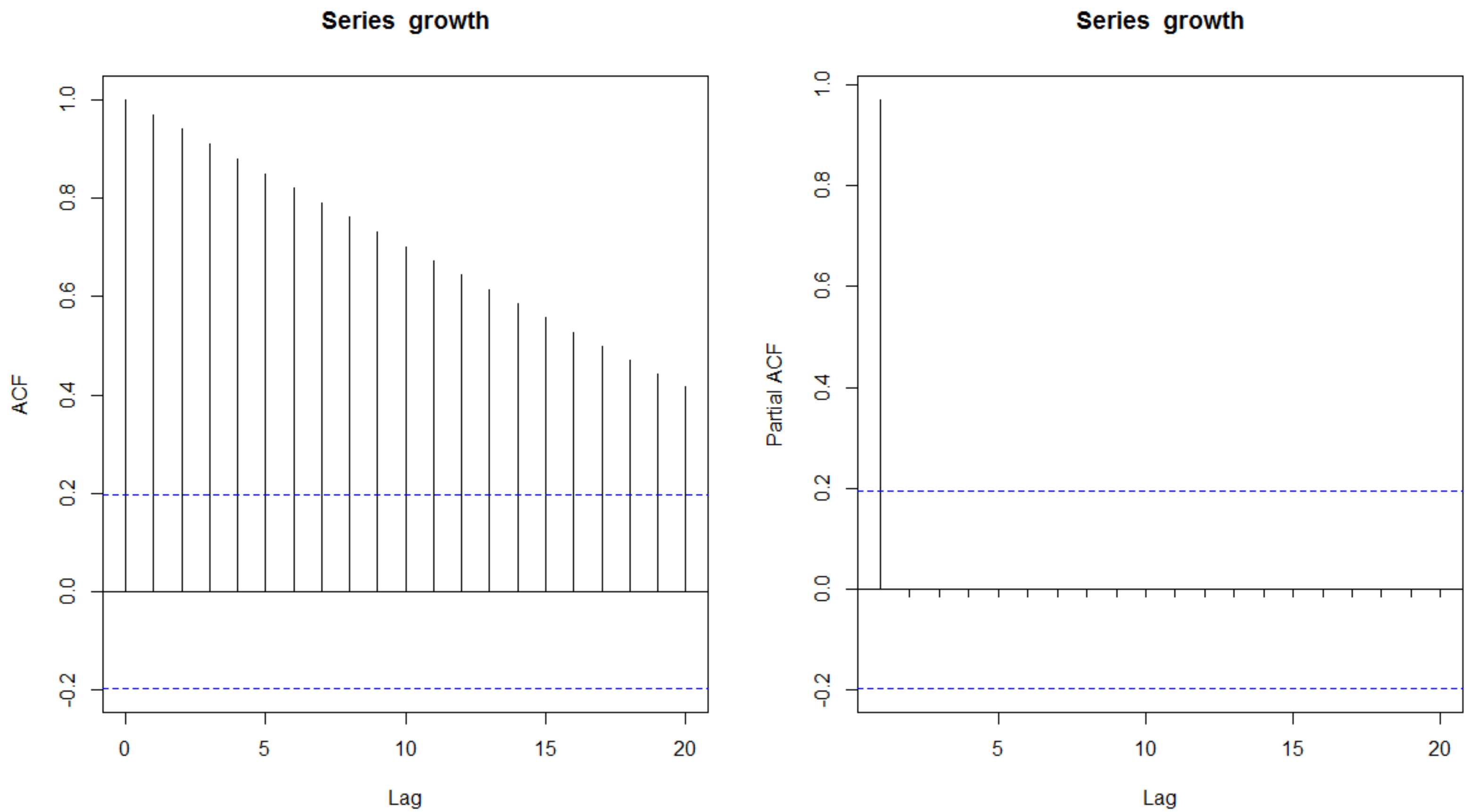
```
>
```

ACF and PACF – Idealized Trend, seasonality and randomness

ACF and PACF – Idealized Trend

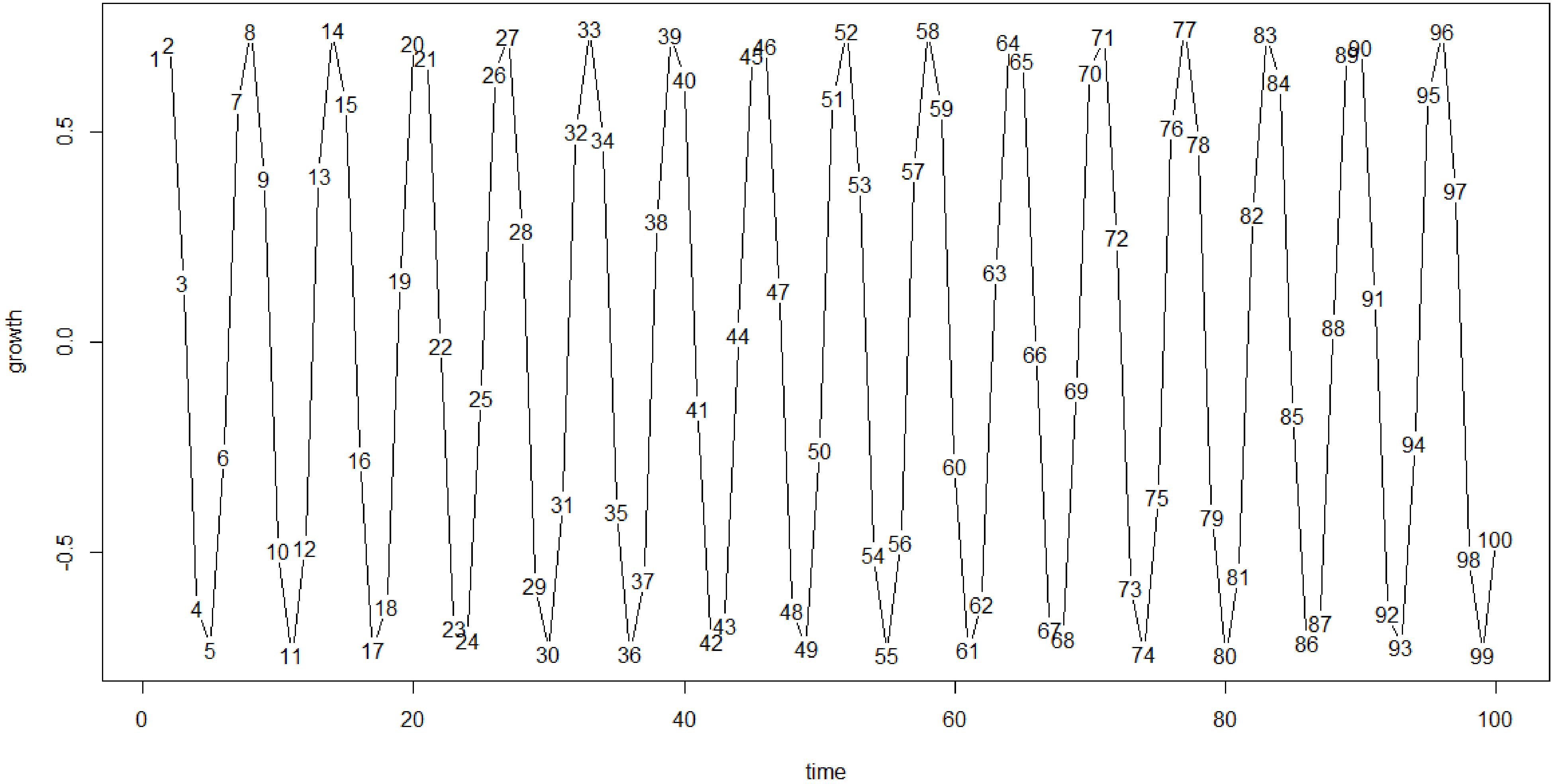


ACF and PACF – Idealized Trend

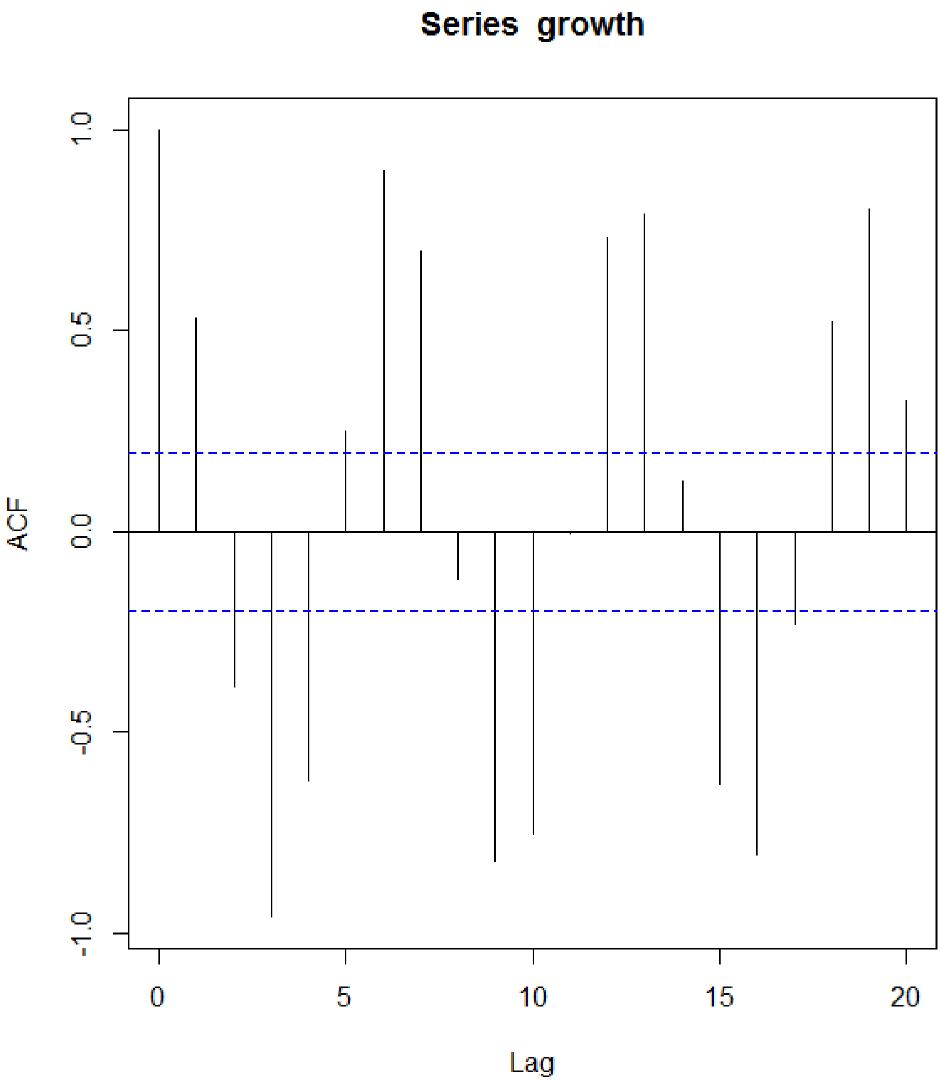


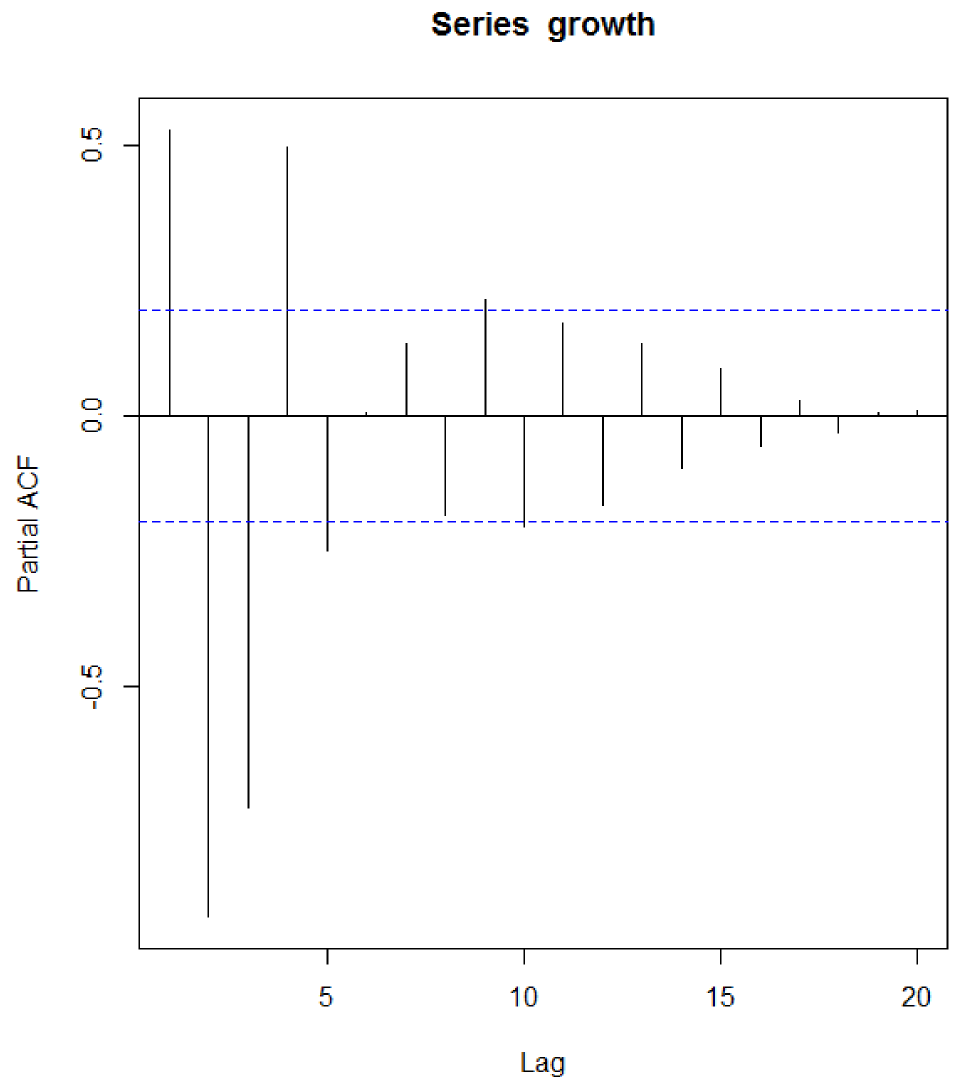
- ACF is a bar chart of correlation coefficients of the time series and its lags.
- PACF is a plot of the partial correlation coefficients of the time series and its lags.

ACF and PACF – Idealized Seasonality

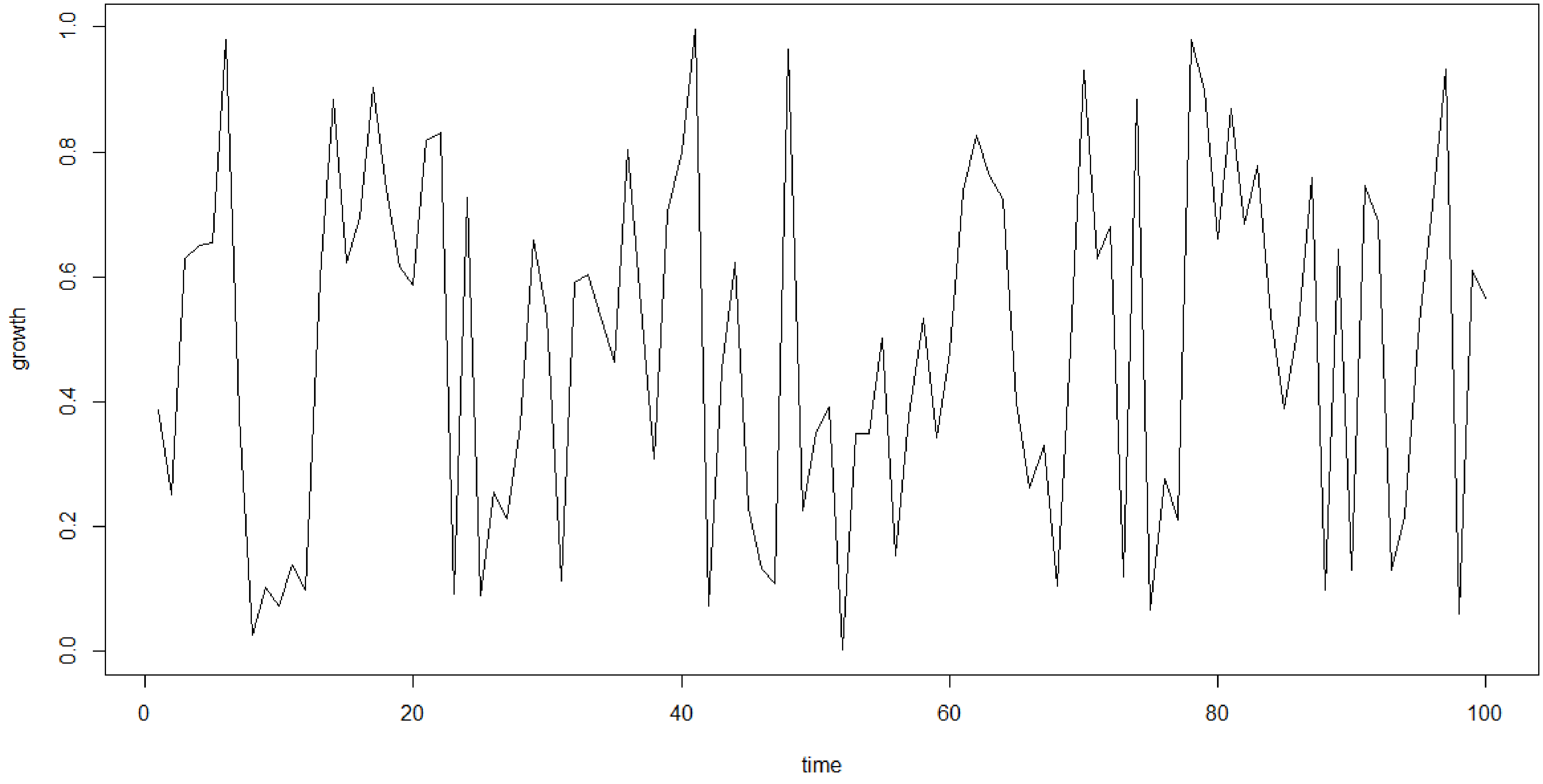


ACF and PACF – Idealized Seasonality

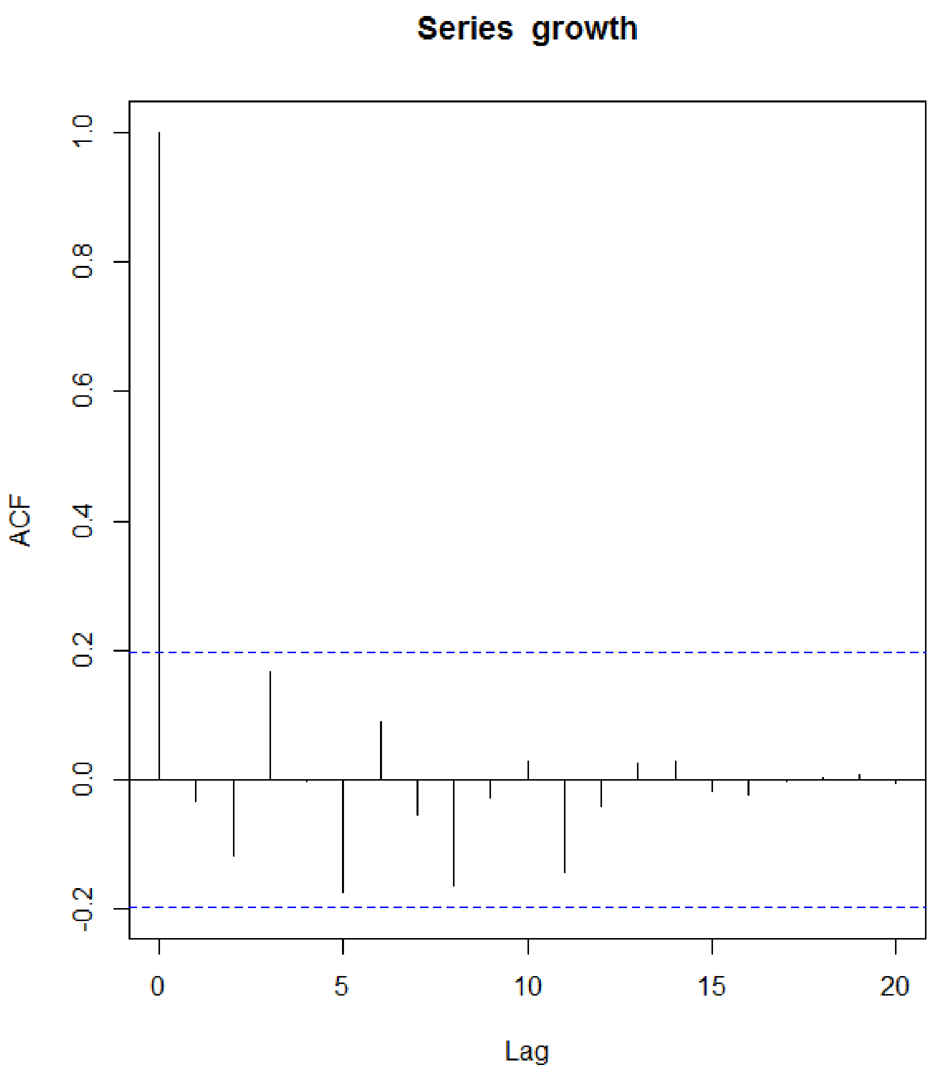


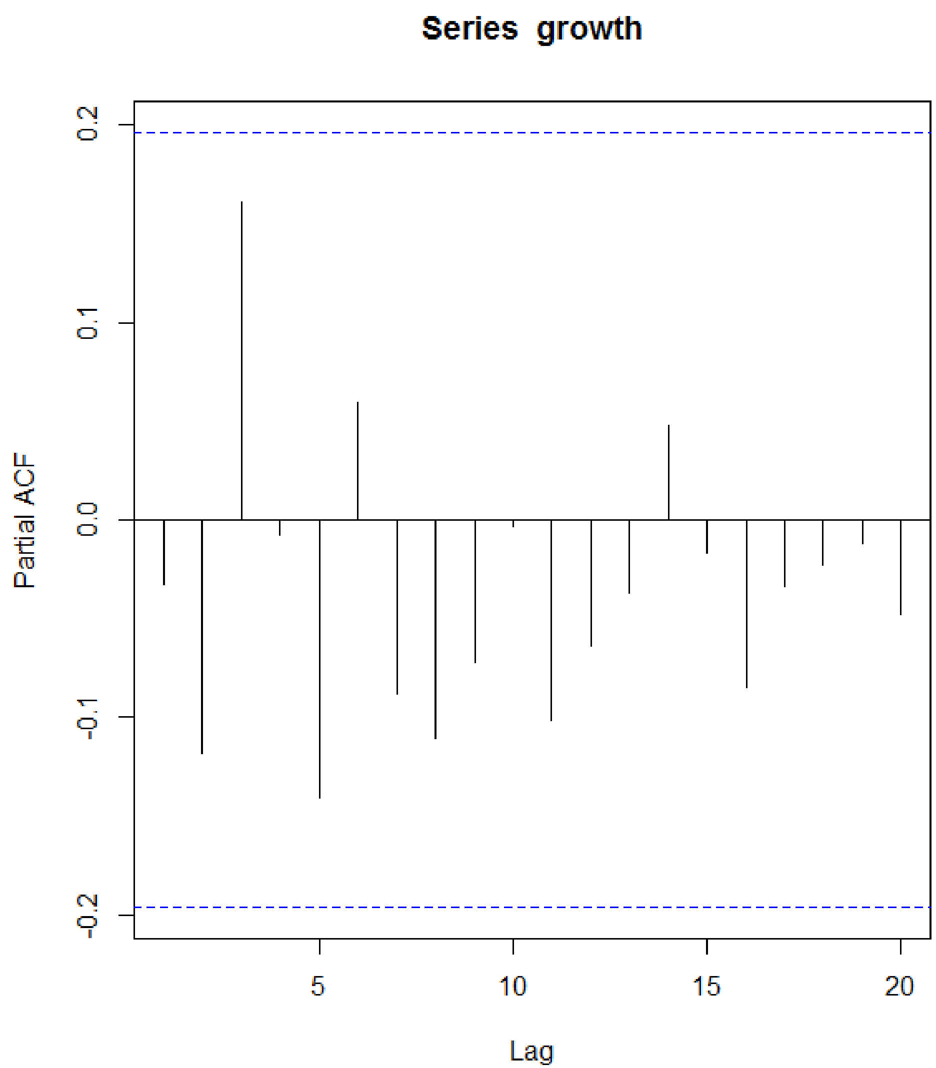


ACF and PACF – Idealized Randomness

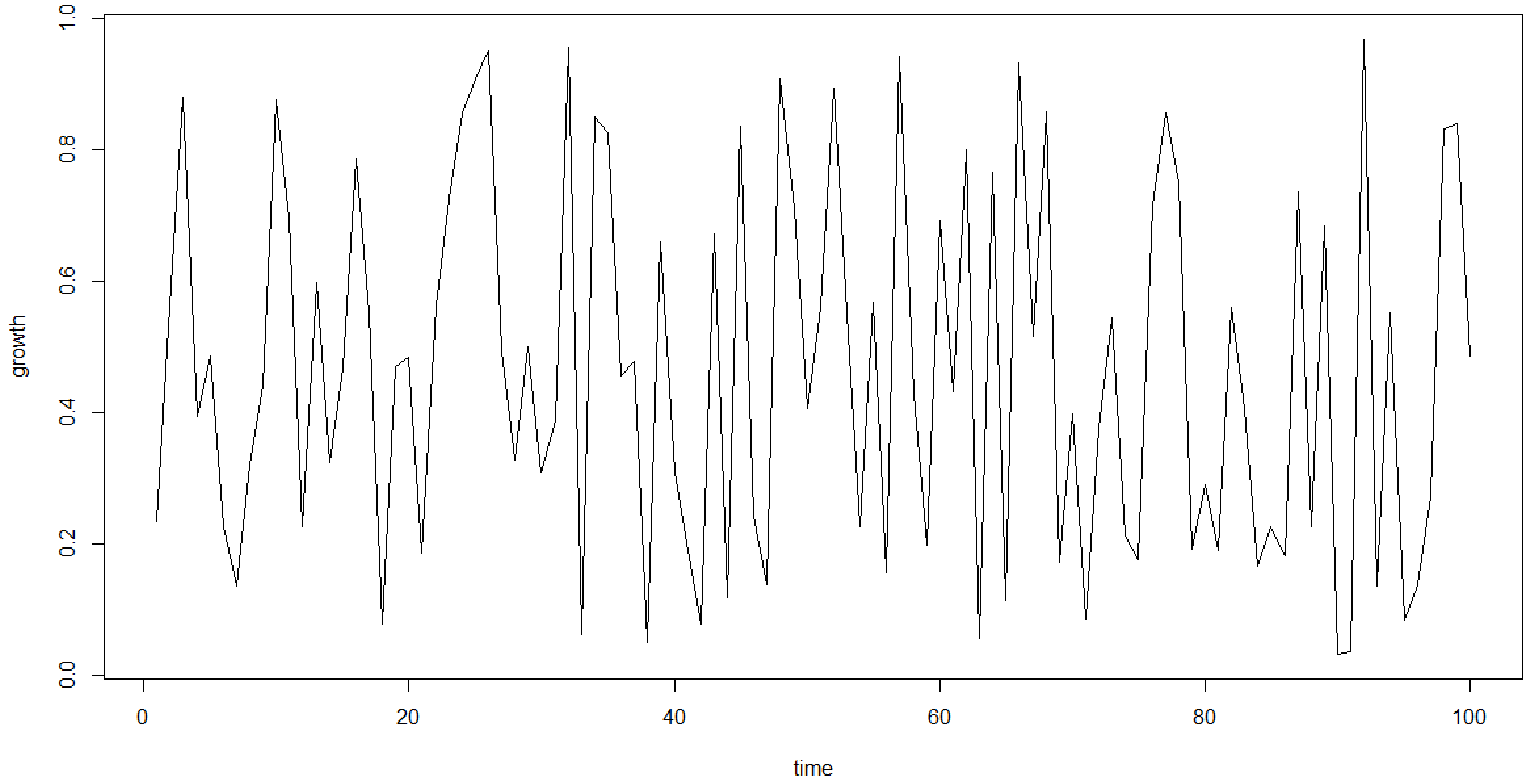


ACF and PACF – Idealized Randomness

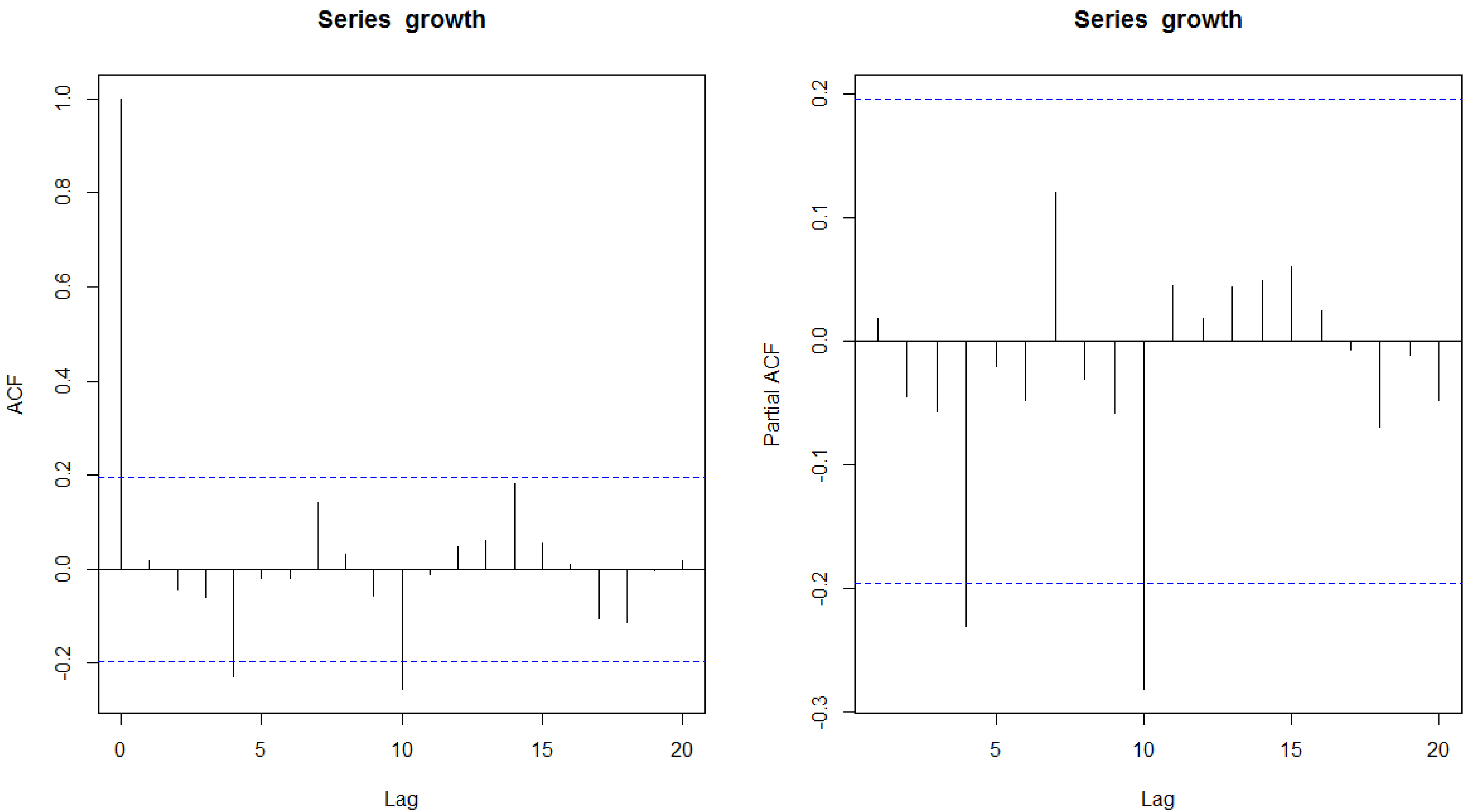




ACF and PACF – Idealized Randomness



ACF and PACF – Idealized Randomness

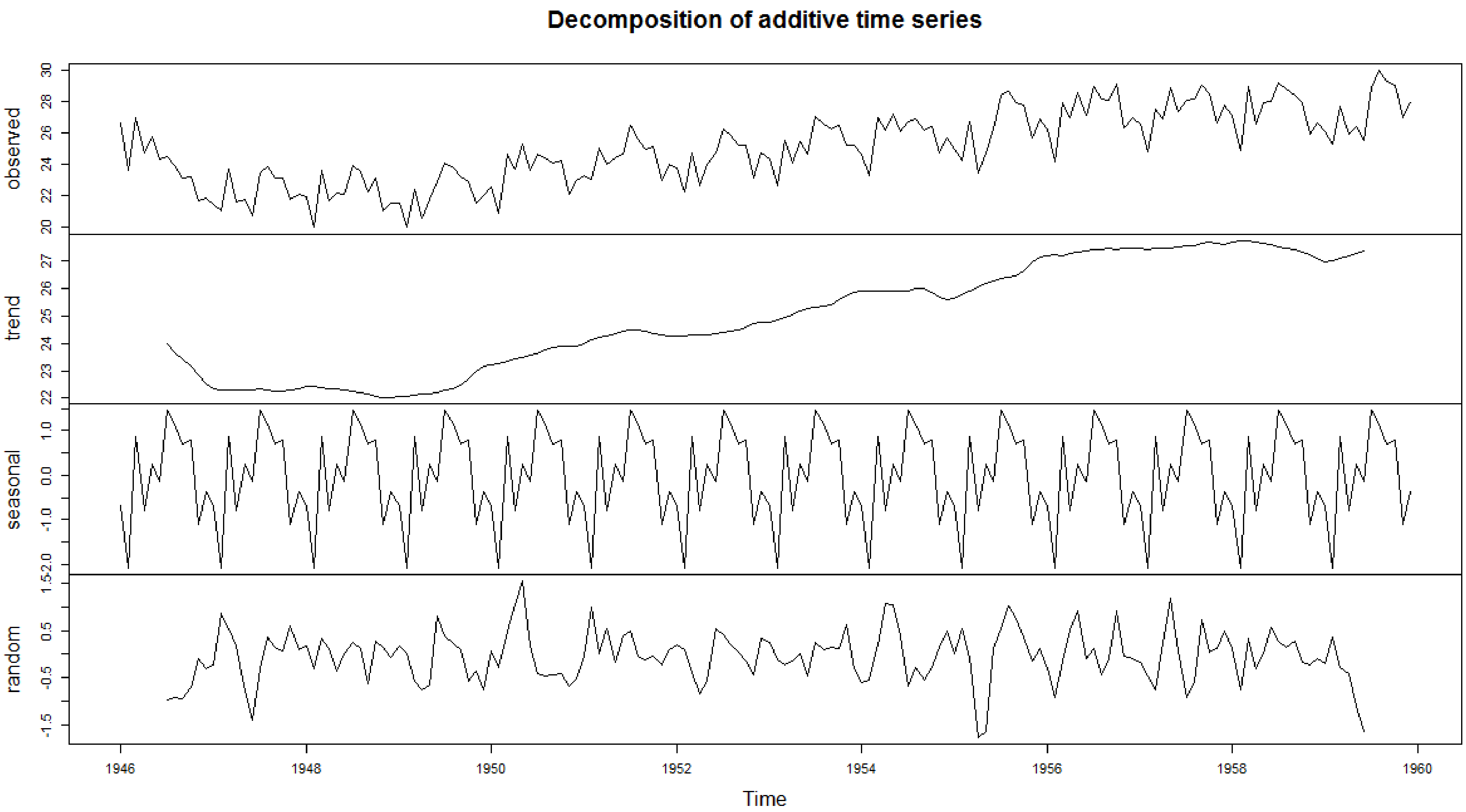


ACF and PACF – Idealized Trend, Seasonality and Randomness

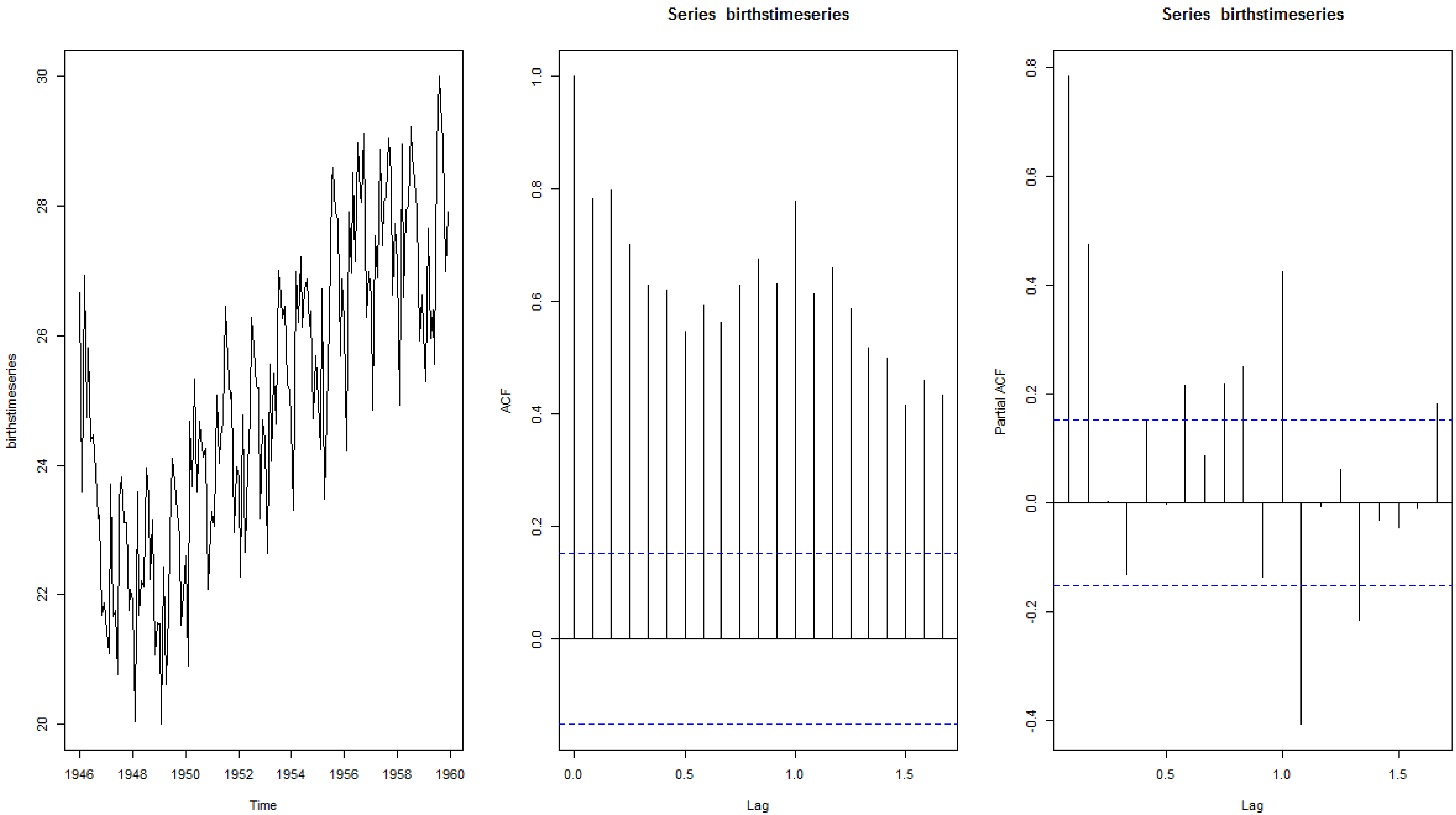
- Ideal Trend: Decreasing ACF and 1 or 2 lags of PACF
- Ideal Seasonality: Cyclicality in ACF and a few lags of PACF with some positive and some negative
- Ideal Random: A spike may or may not be present; even if present, magnitude will be small

ACF and PACF (Real-world): Decomposing Time Series into the 3 Components

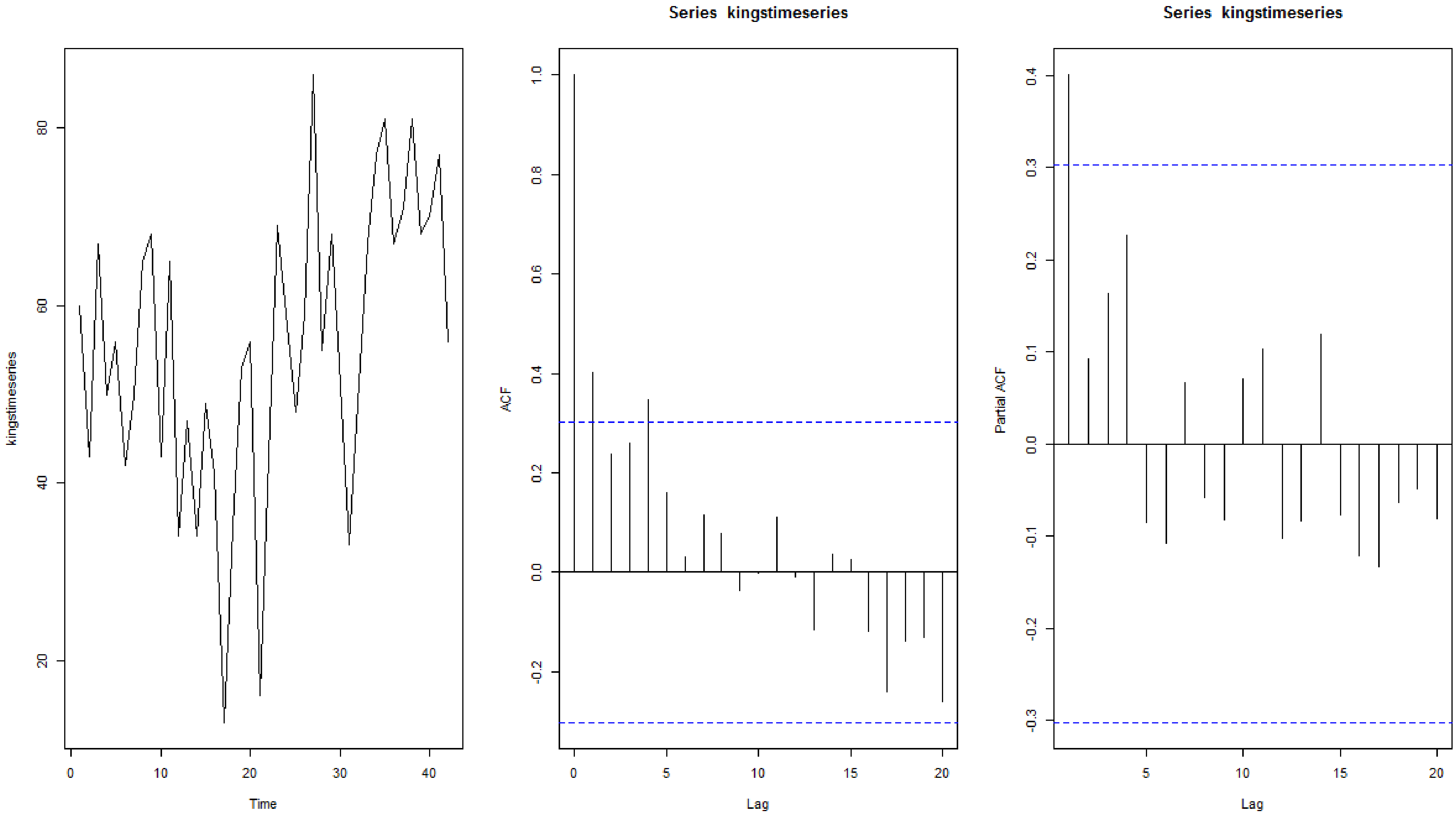
ACF and PACF (Real-world): Decomposing Time Series into the 3 Components - Births



ACF and PACF (Real-world): Decomposing Time Series into the 3 Components - Births



ACF and PACF (Real-world): Decomposing Time Series into the 3 Components – Kings' ages at death

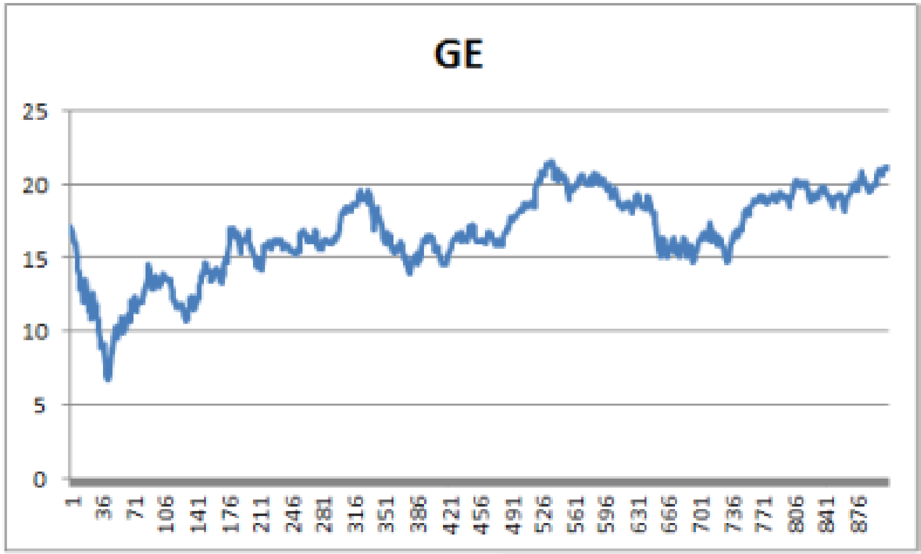


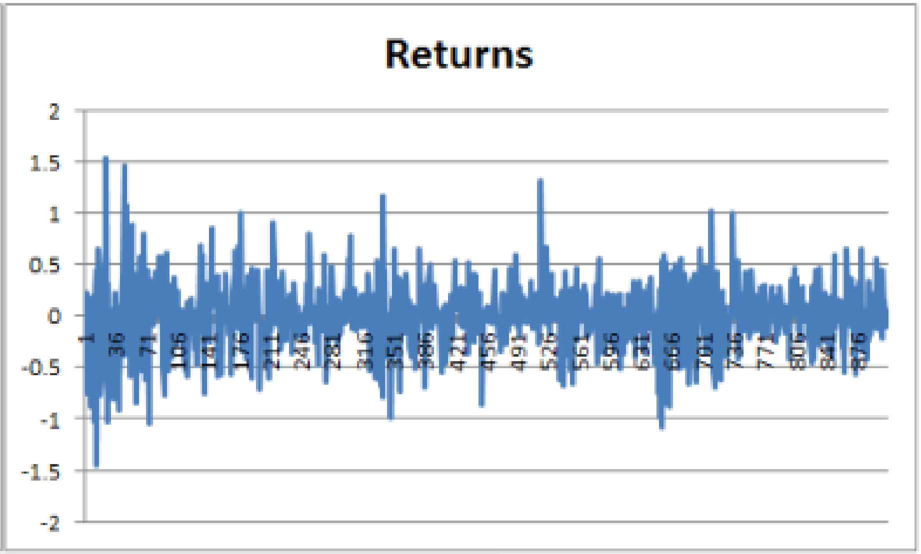
Stationary and Non-Stationary

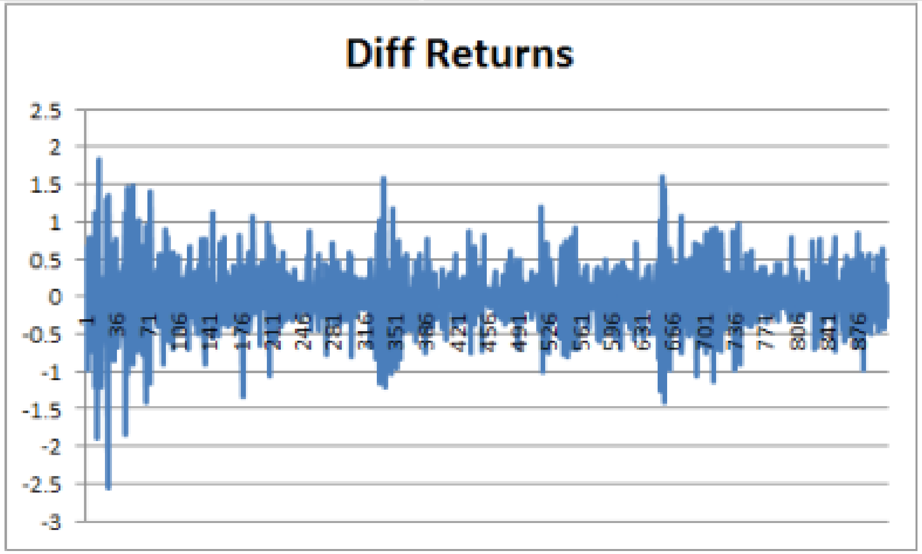
- Stationary data has a constant mean
- If the data is stationary, forecasting is easier!
- Differencing to convert non-stationary to stationary

EXCEL ACTIVITY

Removing trend from data







ACF and PACF of stationary and non-stationary

- Non-stationary series have an ACF that remains significant for half a dozen or more lags, rather than quickly declining to zero.

- You must difference such a series until it is stationary before you can identify the process.

A CRUDE WAY OF SOLVING TIME SERIES (CURVE FITTING)

Goodness of fit

- MSE (Mean square error)
- MAE (Mean absolute error)
- RMSE (Root mean square error)

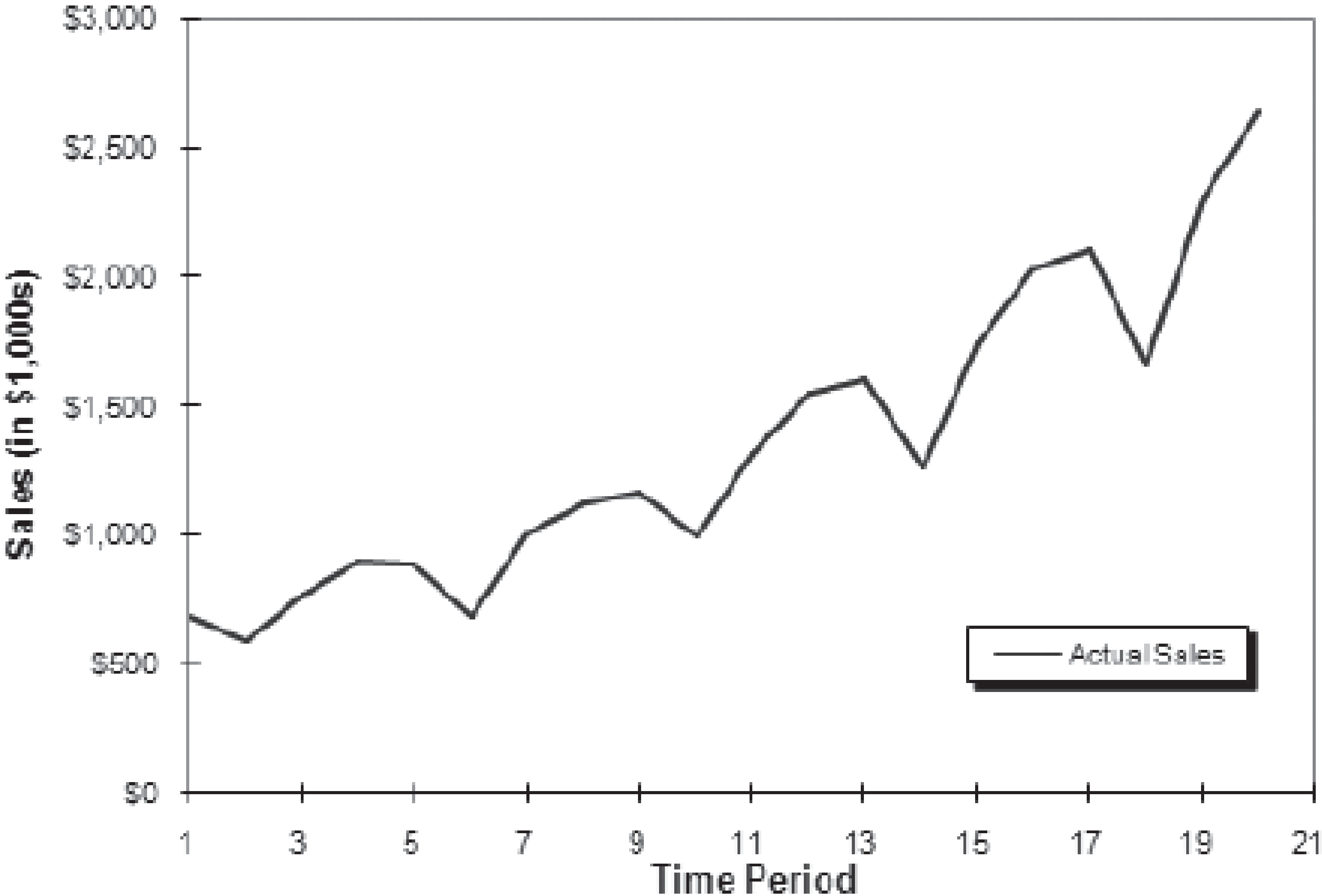
- MAPE (Mean absolute percent error)

- NMSE (Normalized mean square error)
- NMAE (Normalized mean absolute error)
- NMAPE (Normalized mean absolute percent error)

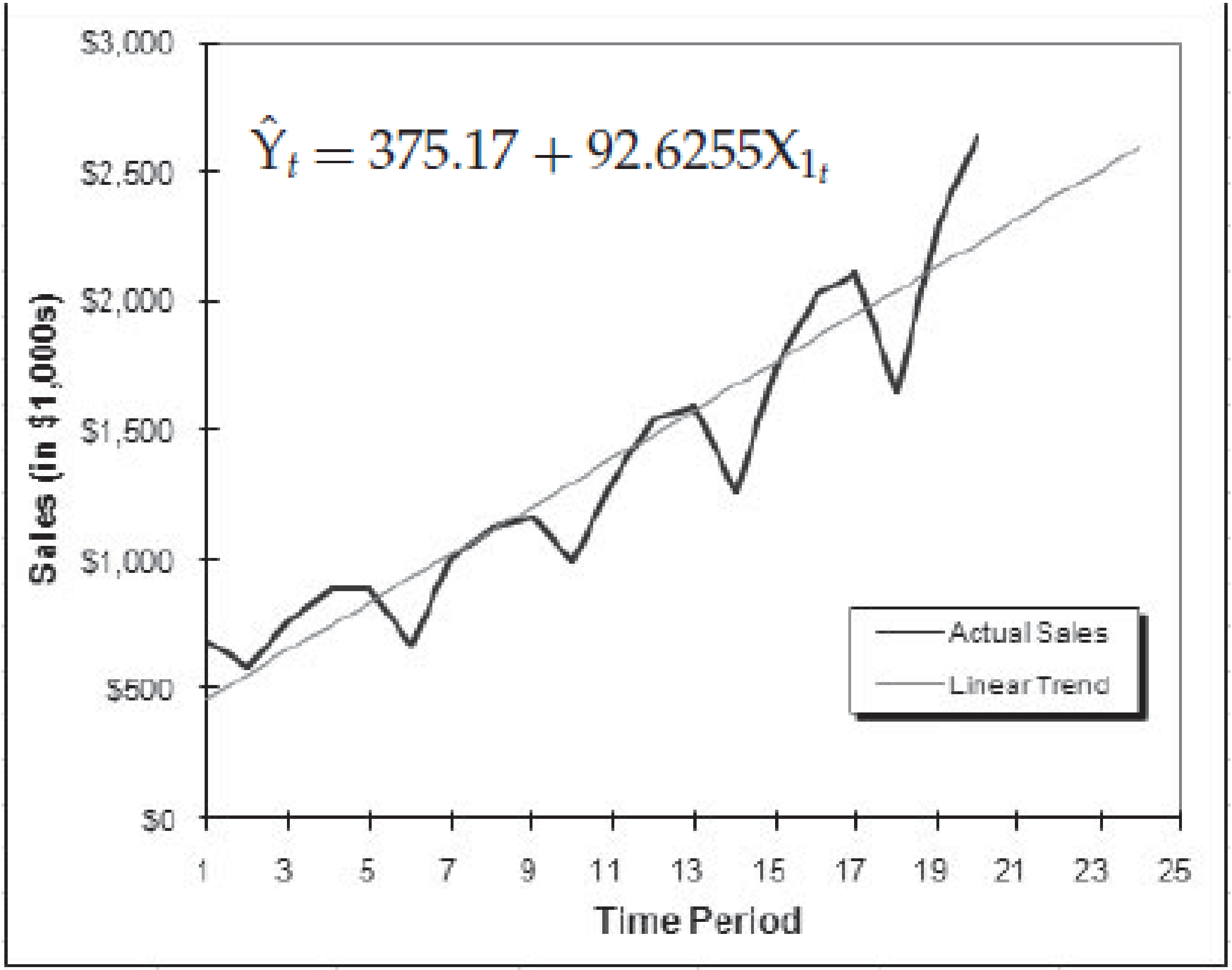
Regression on time

- Use when trend is the most pronounced

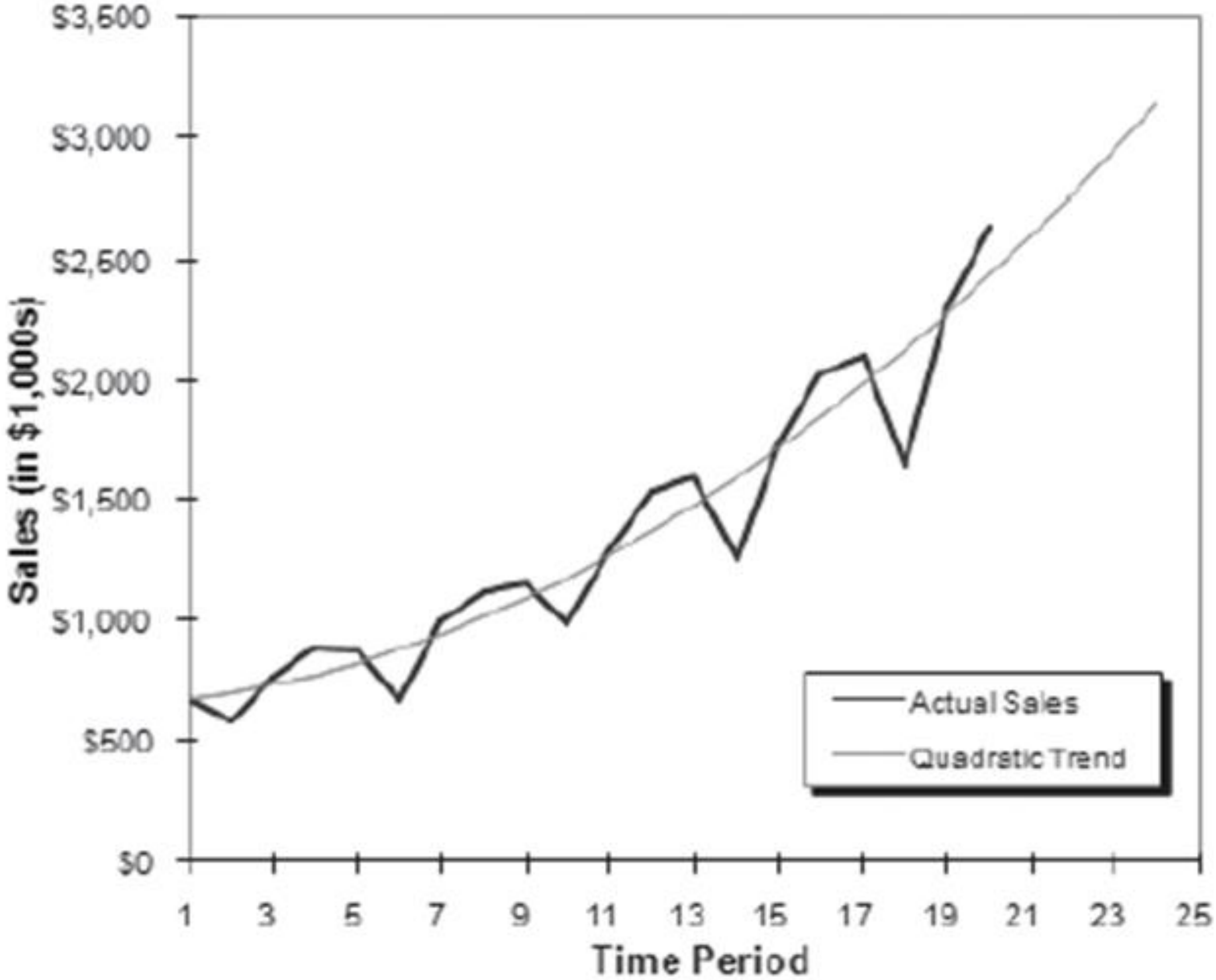
- ACF decays exponentially and PACF has very few spikes


$3,000
$2,500
$2,000
$1,500
$1,000
$500
$0
Sales (in $1,000s)
1
3
5
7
9
11
13
15
17
19
21
Time Period
Actual Sales

Regression analysis



Quadratic trend



Seasonal regression models

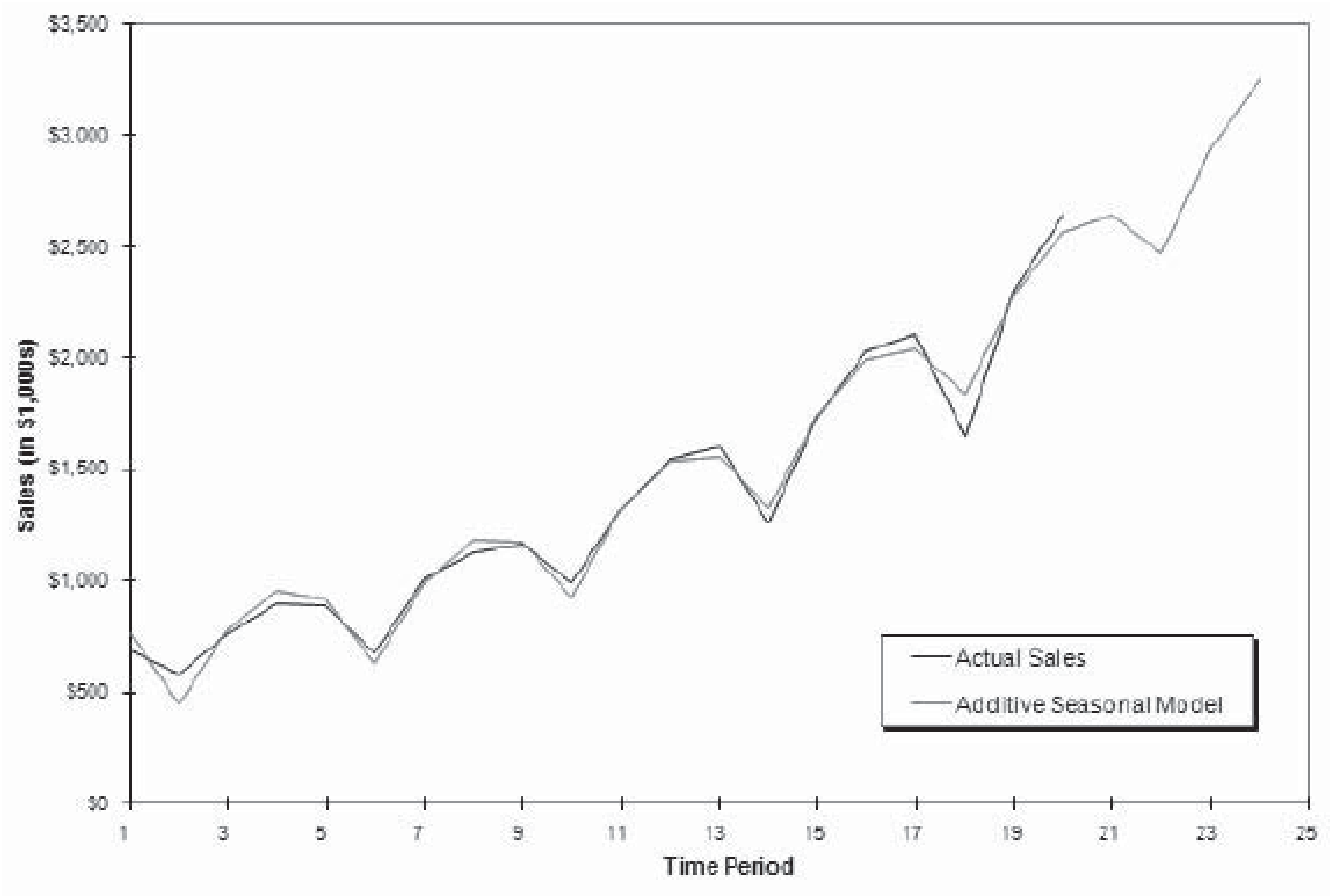
	Value of		
Quarter	X_{3_t}	X_{4_t}	X_{5_t}
1	1	0	0
2	0	1	0
3	0	0	1
4	0	0	0

$$Y_t = \beta_0 + \beta_1 X_{1_t} + \beta_2 X_{2_t} + \beta_3 X_{3_t} + \beta_4 X_{4_t} + \beta_5 X_{5_t} + \varepsilon_t$$

where, $X_{1t} = t$ and $X_{2t} = t^2$.

CSE 7202c

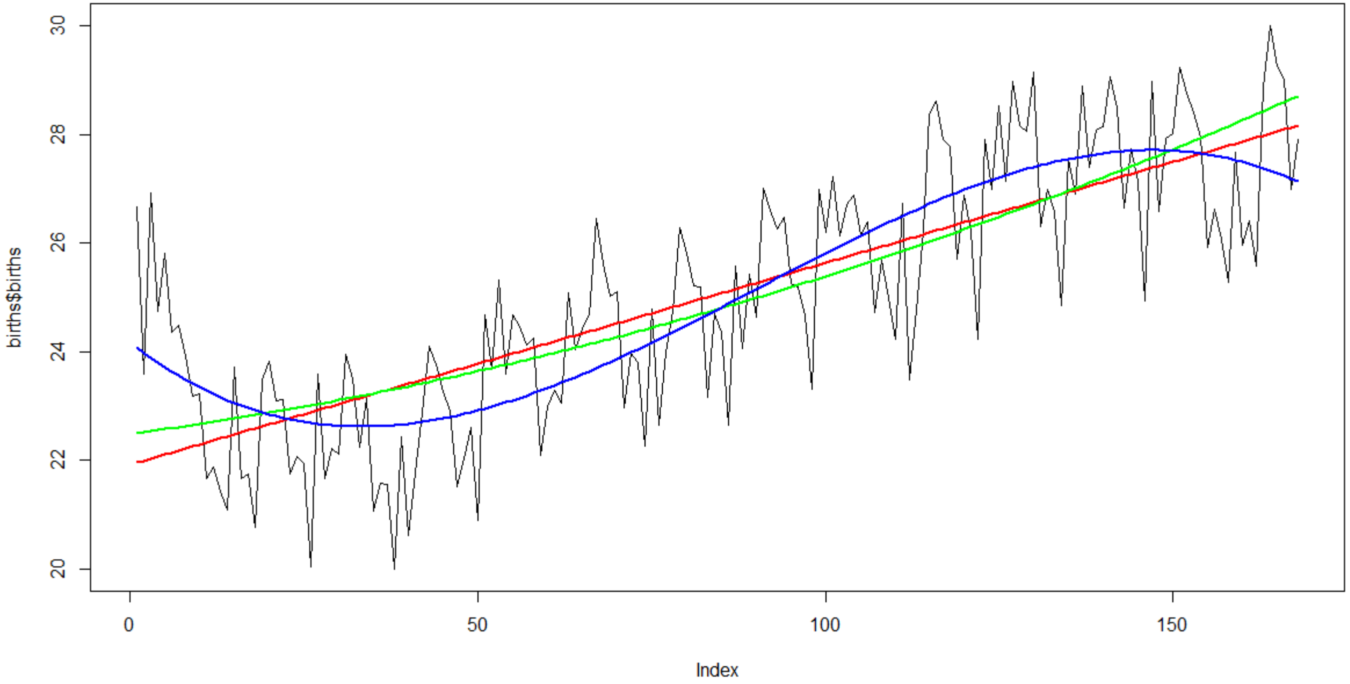
Seasonal regression models



CSE 7202c

Seasonal regression models

Seasonal regression models - Births

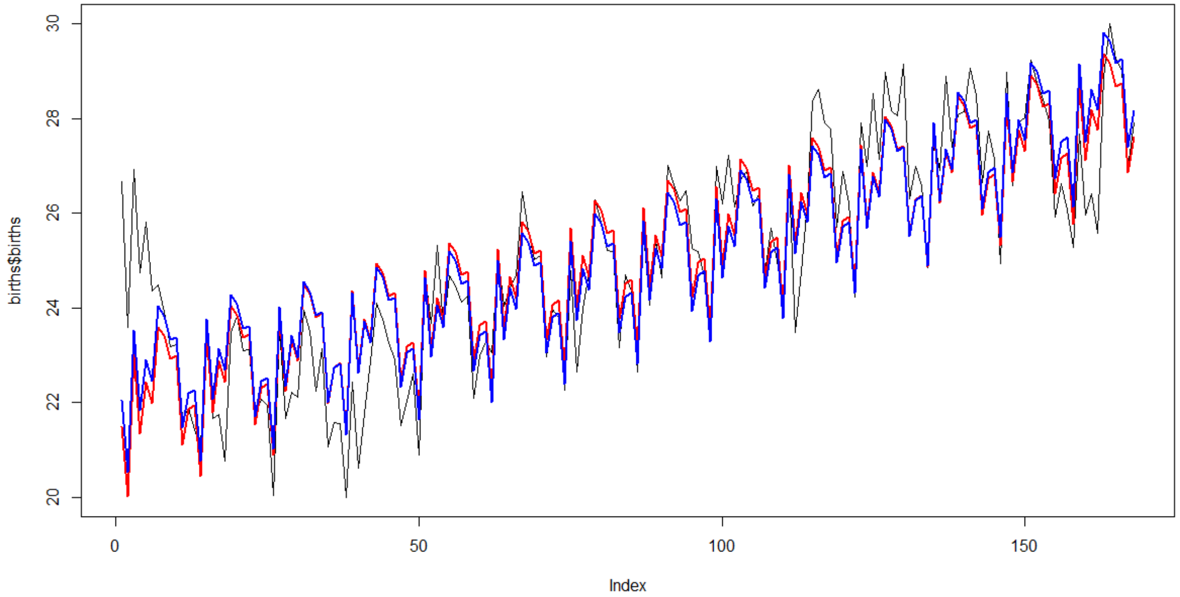


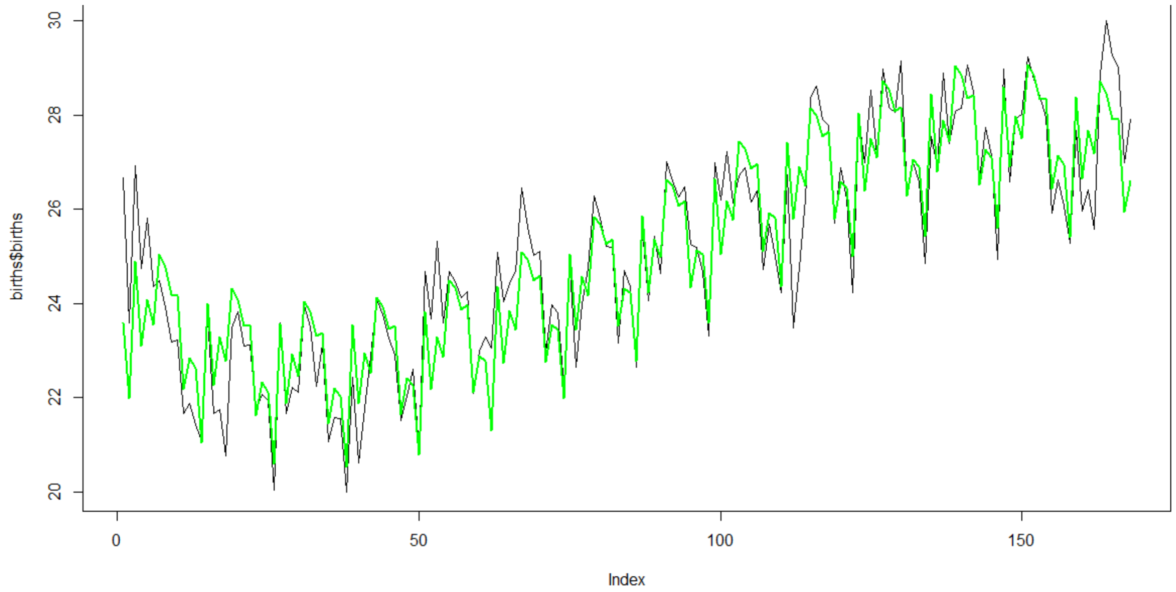
Data Editor

File Edit Help

	births	time	var3	var4	var5	var6
1	26.663	1				
2	23.598	2				
3	26.931	3				
4	24.74	4				
5	25.806	5				
6	24.364	6				
7	24.477	7				
8	23.901	8				
9	23.175	9				
10	23.227	10				
11	21.672	11				
12	21.87	12				
13	21.439	13				
14	21.089	14				
15	23.709	15				
16	21.669	16				
17	21.752	17				
18	20.761	18				
19	23.479	19				

Seasonal regression models - Births





Data Editor

File Edit Help

	births	time	seasonal	var4	var5
1	26.663	1	1		
2	23.598	2	2		
3	26.931	3	3		
4	24.74	4	4		
5	25.806	5	5		
6	24.364	6	6		
7	24.477	7	7		
8	23.901	8	8		
9	23.175	9	9		
10	23.227	10	10		
11	21.672	11	11		
12	21.87	12	12		
13	21.439	13	1		
14	21.089	14	2		
15	23.709	15	3		
16	21.669	16	4		
17	21.752	17	5		
18	20.761	18	6		
19	23.479	19	7		

Another crude way of incorporating seasonality

- Take the trend prediction and actual prediction

- Depending on additive or multiplicative model compute the deviation and map it as seasonality effect for each prediction

- Take averages of the seasonality value. Use this to make future predictions

Case

Year	Quarter	Time variable (this is created)	Revenues
2008	I	1	10.2
	II	2	12.4
	III	3	14.8
	IV	4	15
2009	I	5	11.2
	II	6	14.3
	III	7	18.4
	IV	8	18

```
Call:
lm(formula = y ~ x)

Residuals:
    Min      1Q  Median      3Q     Max
-3.5595 -0.9384  0.4405  1.3265  1.9286

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  10.0393     1.5531   6.464  0.00065 ***
x             0.9440     0.3076   3.069  0.02196 *
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 1.993 on 6 degrees of freedom
Multiple R-squared: 0.6109,     Adjusted R-squared: 0.5461
F-statistic: 9.422 on 1 and 6 DF,  p-value: 0.02196
```

Seasonality: Multiplicative

Time	Observed values TSI (assuming no impact of cyclicality)	Predicted values (per the regression) T	SI = TSI/T
1	10.2	10.983	0.929
2	12.4	11.927	1.040
3	14.8	12.871	1.150
4	15	13.815	1.086
5	11.2	14.759	0.759
6	14.3	15.703	0.911
7	18.4	16.647	1.105
8	18	17.591	1.023

Quarterly seasonality

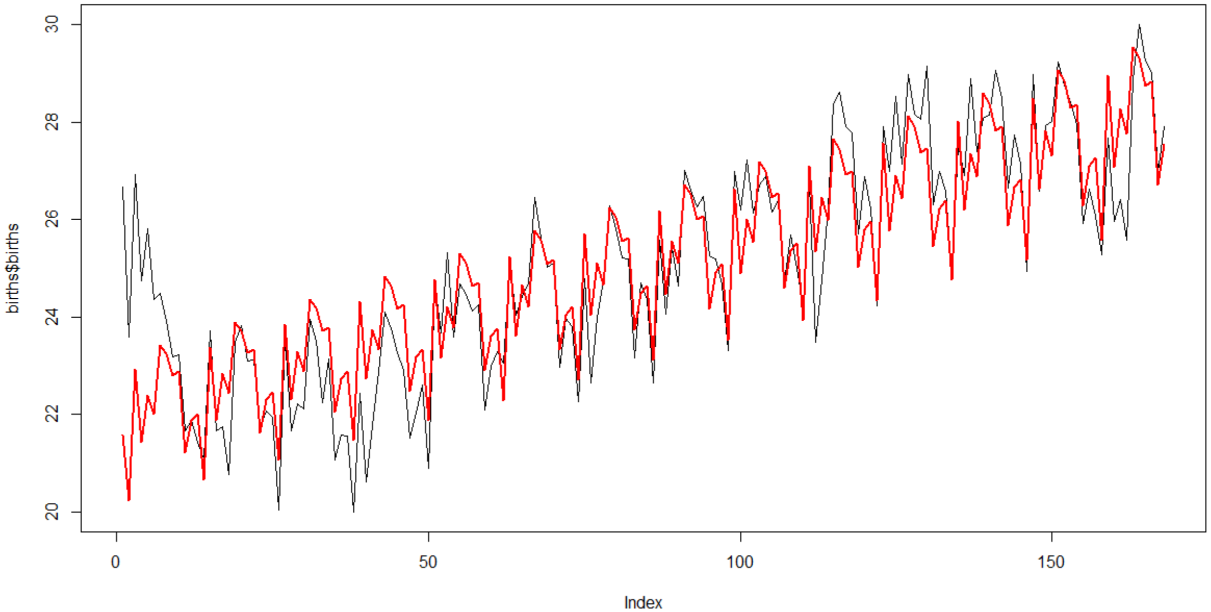
Time	Average seasonality factor
Q1	0.844
Q2	0.975
Q3	1.127
Q4	1.054

Computations

- Trend $Y_9 = 10.039 + 0.944(9) = 18.535$

- Corrected for seasonality and randomness: $18.535 * 0.844 = 15.643$

Seasonality: Multiplicative



Data Editor

File Edit Help

	births	time	seasonal	mae	var5
1	26.663	1	1	1.214304	
2	23.598	2	2	1.072901	
3	26.931	3	3	1.222374	
4	24.74	4	4	1.121036	
5	25.806	5	5	1.167374	
6	24.364	6	6	1.100294	
7	24.477	7	7	1.103546	
8	23.901	8	8	1.075775	
9	23.175	9	9	1.041357	
10	23.227	10	10	1.041954	
11	21.672	11	11	0.9705802	
12	21.87	12	12	0.9778208	
13	21.439	13	1	0.9569611	
14	21.089	14	2	0.93978	
15	23.709	15	3	1.054788	
16	21.669	16	4	0.9624398	
17	21.752	17	5	0.9645349	
18	20.761	18	6	0.9190777	
19	23.479	19	7	1.037695	

Issues with regressing on time

- It is too much of a curve fit For a statistician to sleep well!
- If there is no trend or if seasonality and fluctuations are more important than trend, then the coefficients behave weirdly

TIME SERIES: MORE ROBUST ANALYSES

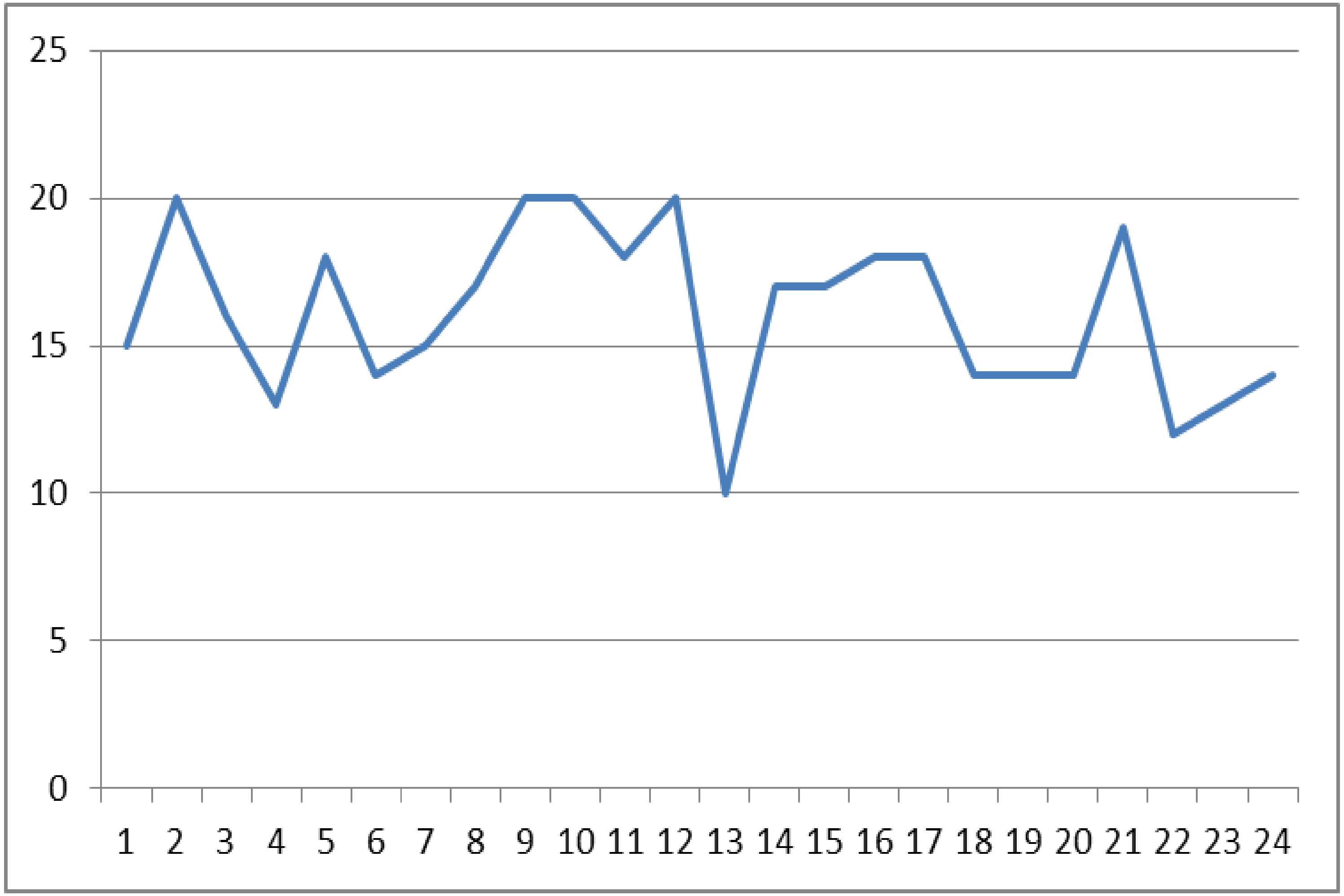
Identifying techniques for different processes

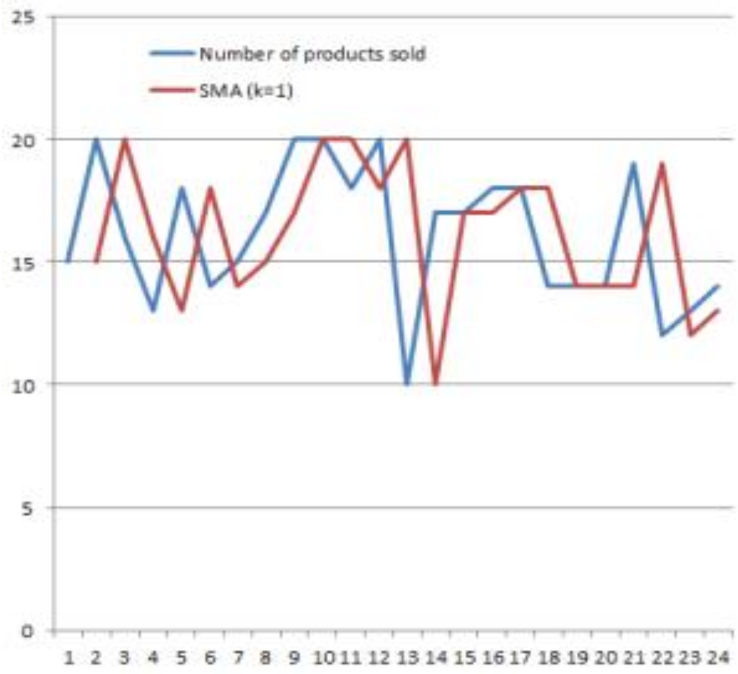
- We use different techniques for different processes
  - Random stationary
  - Seasonal
  - Trend
- First we need to identify them

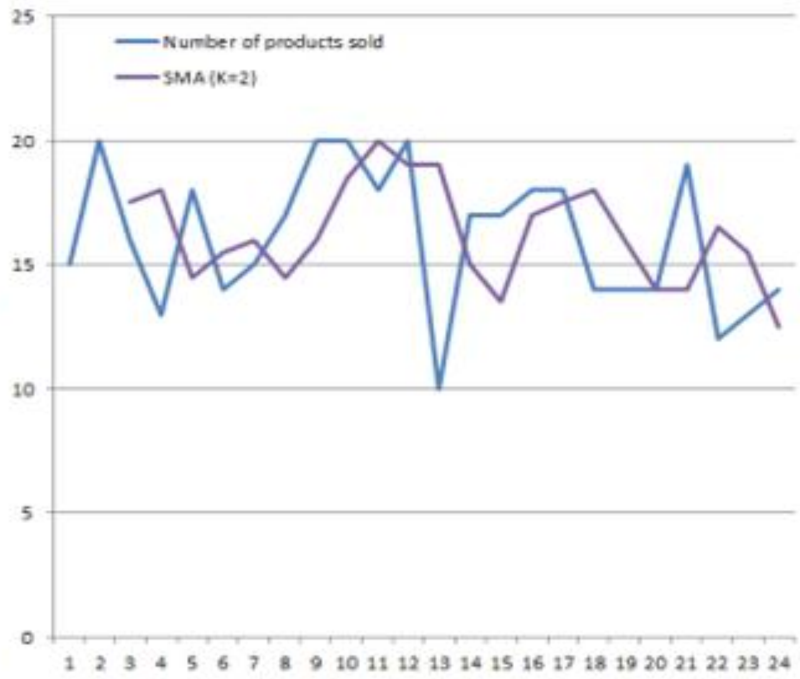
Stationary model: Case 1 – Simple Moving Averages

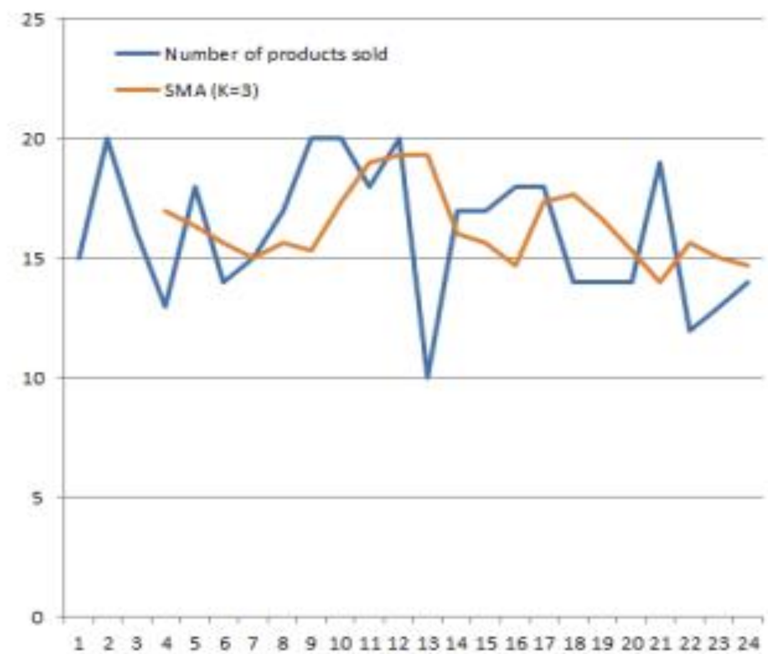
Number of products sold	SMA (k=1)	Error	SMA (K=2)	Error	SMA (K=3)	Error
15						
20	15	5				
16	20	4	17.5	1.5		
13	16	3	18	5	17	4
18	13	5	14.5	3.5	16.333333	1.666667
14	18	4	15.5	1.5	15.666667	1.666667
15	14	1	16	1	15	0
17	15	2	14.5	2.5	15.666667	1.333333
20	17	3	16	4	15.333333	4.666667
20	20	0	18.5	1.5	17.333333	2.666667
18	20	2	20	2	19	1
20	18	2	19	1	19.333333	0.666667
10	20	10	19	9	19.333333	9.333333
17	10	7	15	2	16	1
17	17	0	13.5	3.5	15.666667	1.333333
18	17	1	17	1	14.666667	3.333333
18	18	0	17.5	0.5	17.333333	0.666667
14	18	4	18	4	17.666667	3.666667
14	14	0	16	2	16.666667	2.666667
14	14	0	14	0	15.333333	1.333333
19	14	5	14	5	14	5
12	19	7	16.5	4.5	15.666667	3.666667
13	12	1	15.5	2.5	15	2
14	13	1	12.5	1.5	14.666667	0.666667
		2.913043		**2.681818**		**2.492063**

Stationary model: Moving Averages









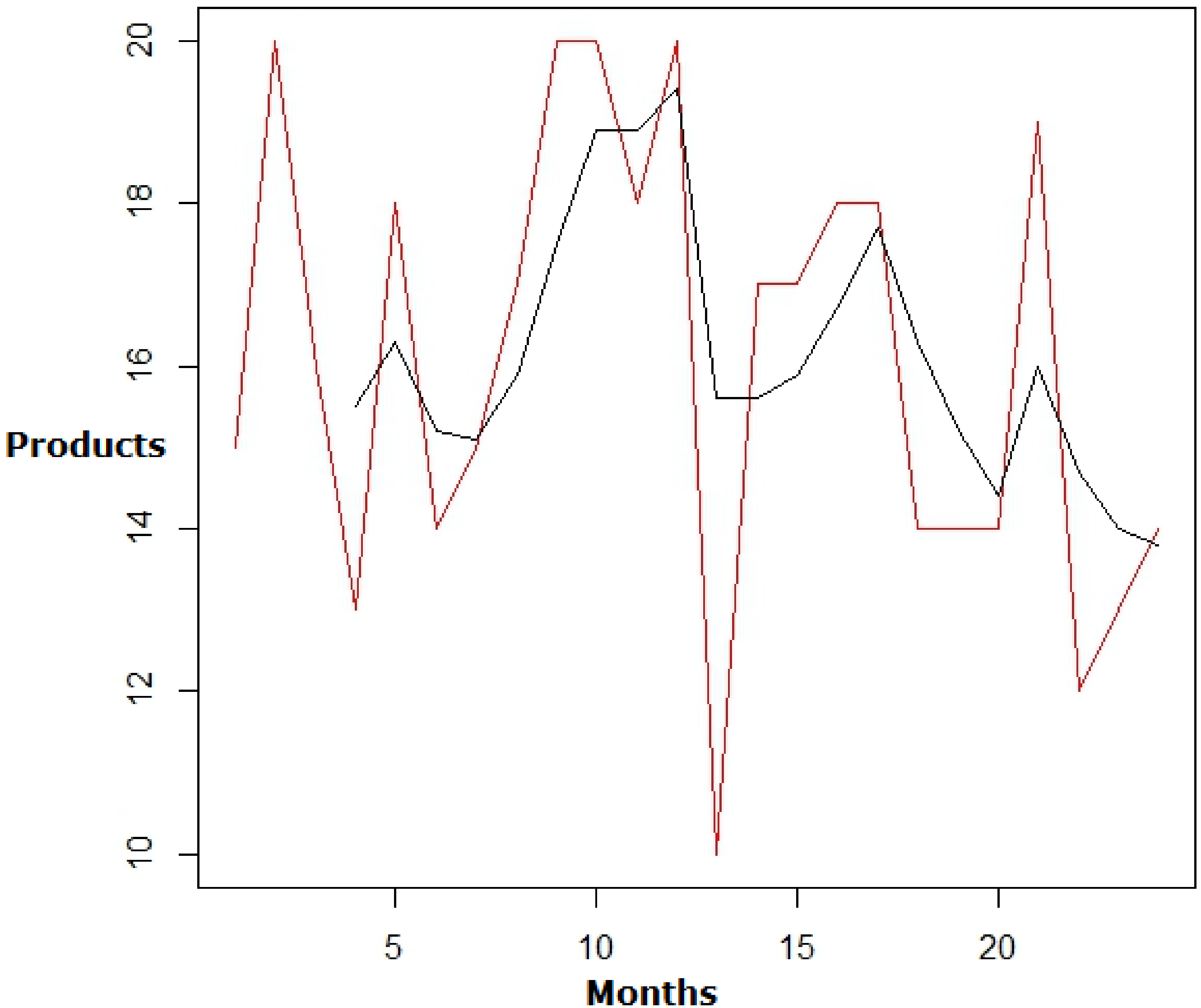
Only decision point is K

Stationary model: Weighted moving average

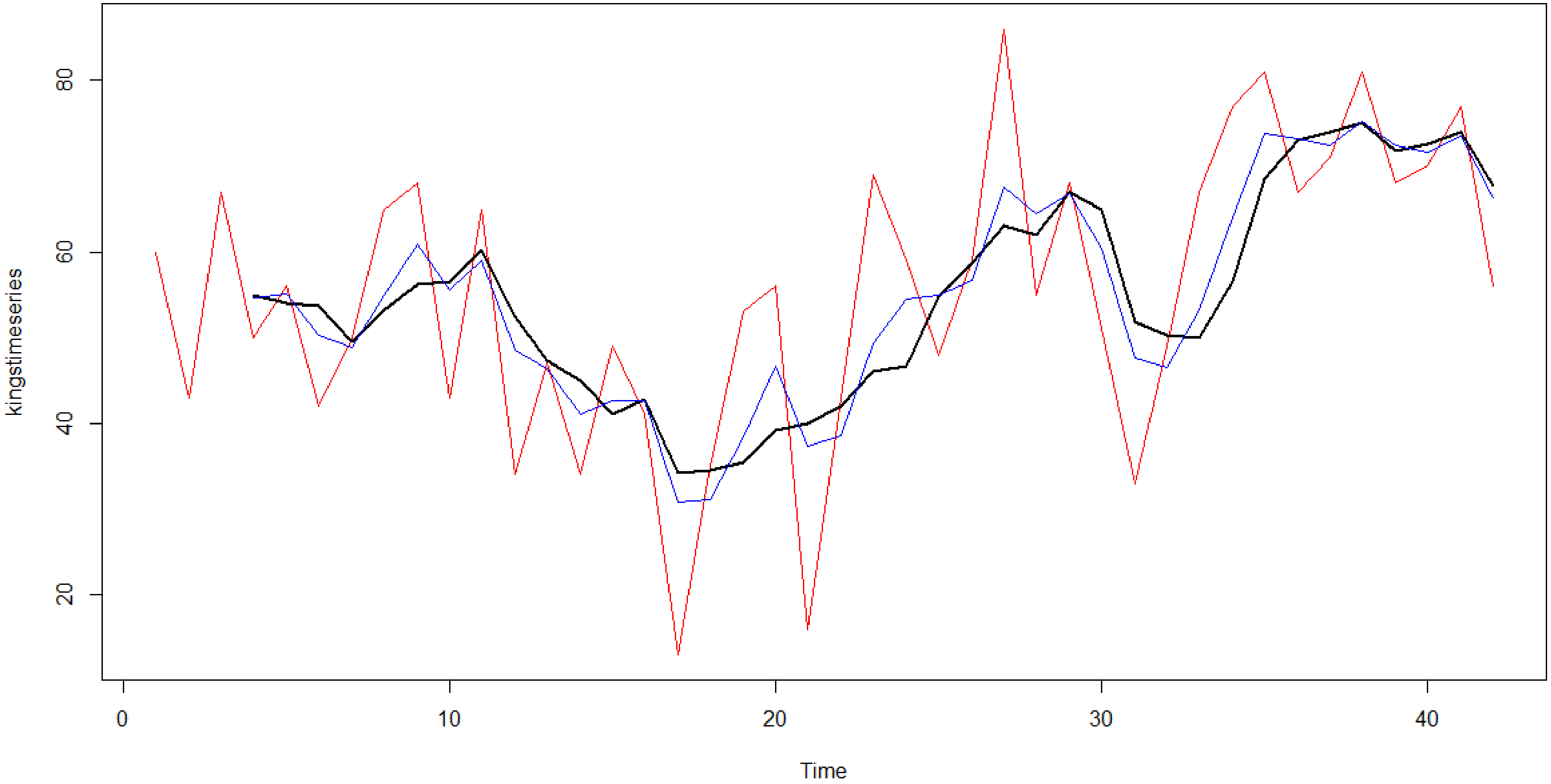
$$\hat{Y}_{t+1} = w_1 Y_t + w_2 Y_{t-1} + \cdots + w_k Y_{t-k+1}$$

- Typically we choose a time period of moving average and weights are chosen such that the error is minimized

WMA



SMA and WMA – Kings' ages at death



errorSMA	9.60897435897436
errorWMA	7.91794871794872

Stationary model: Exponential smoothing

$$\hat{Y}_{t+1} = \hat{Y}_t + \alpha(Y_t - \hat{Y}_t)$$

Above equation indicates that the predicted value for time period t+1 ($\widehat{Y_{t+1}}$) is equal to the predicted value for the previous period ($\widehat{Y_t}$) plus an adjustment for the error made in predicting the previous period's value ($\alpha(Y_t$ - $\widehat{Y_{t+1}}$)).

The parameter α can assume any value between 0 and 1 ($0 \leq \alpha \leq 1$).

Exponential smoothing in other ways

$$\widehat{Y_{t+1}} = \widehat{Y_t} + \alpha(Y_t - \widehat{Y_t})$$

$$= \alpha Y_t + (1-\alpha)\,\widehat{Y_t}$$

$$\widehat{Y_{t+1}} = Y_t - (1-\alpha)\left(Y_t - \widehat{Y_t}\right)$$

$$\hat{Y}_{t+1} = \alpha Y_t + \alpha(1-\alpha)Y_{t-1} + \alpha(1-\alpha)^2 Y_{t-2} + \cdots + \alpha(1-\alpha)^n Y_{t-n} + \cdots$$

EXCEL ACTIVITY – Effect of α

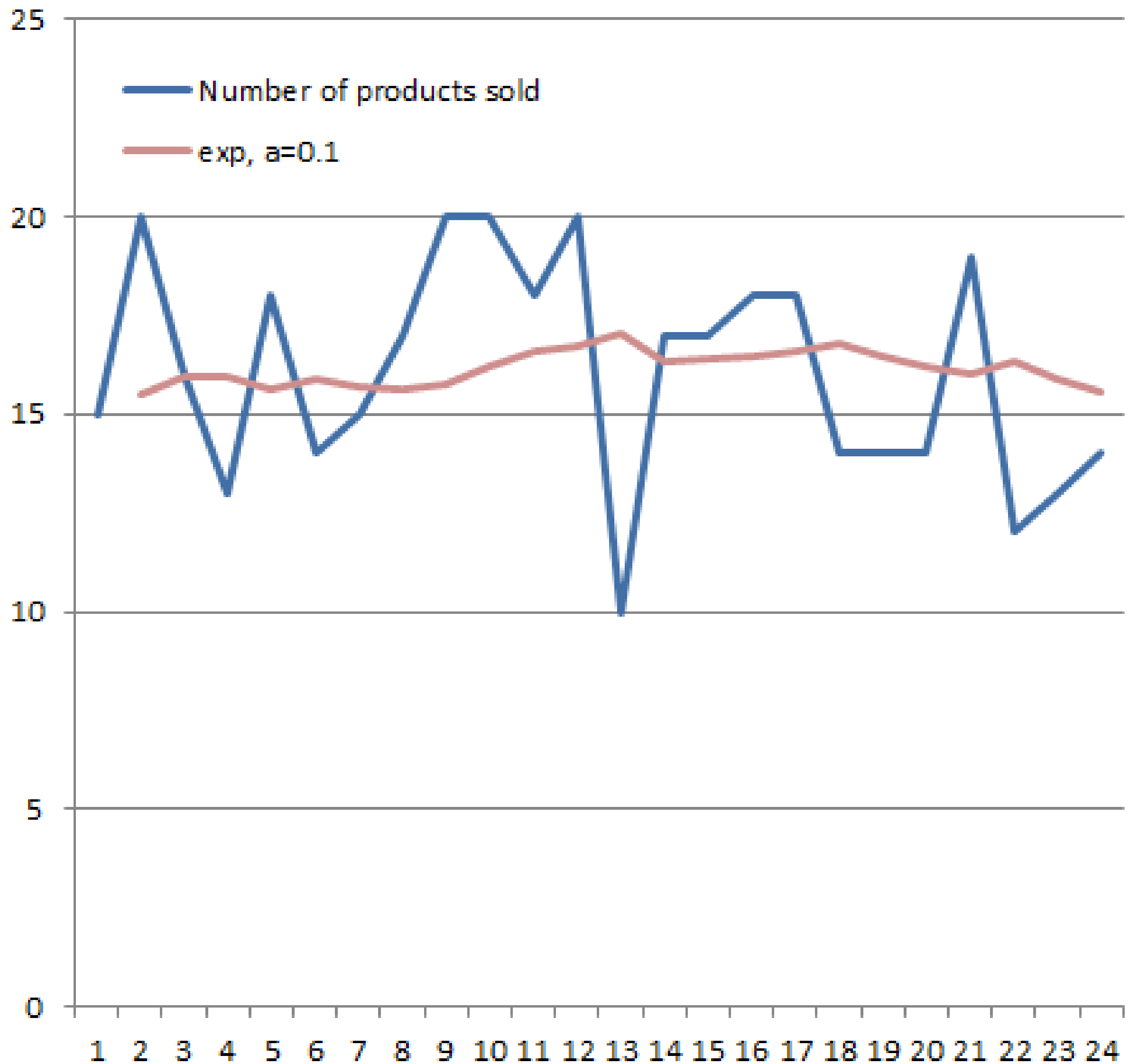
Various ways of understanding exponential smoothing

- Forecast
  - Interpolation between previous *forecast* and previous *observation*
  - Previous *forecast* plus fraction of previous error
  - Previous *observation* minus fraction 1- of previous error
  - *Exponentially weighted (i.e. discounted) moving average*

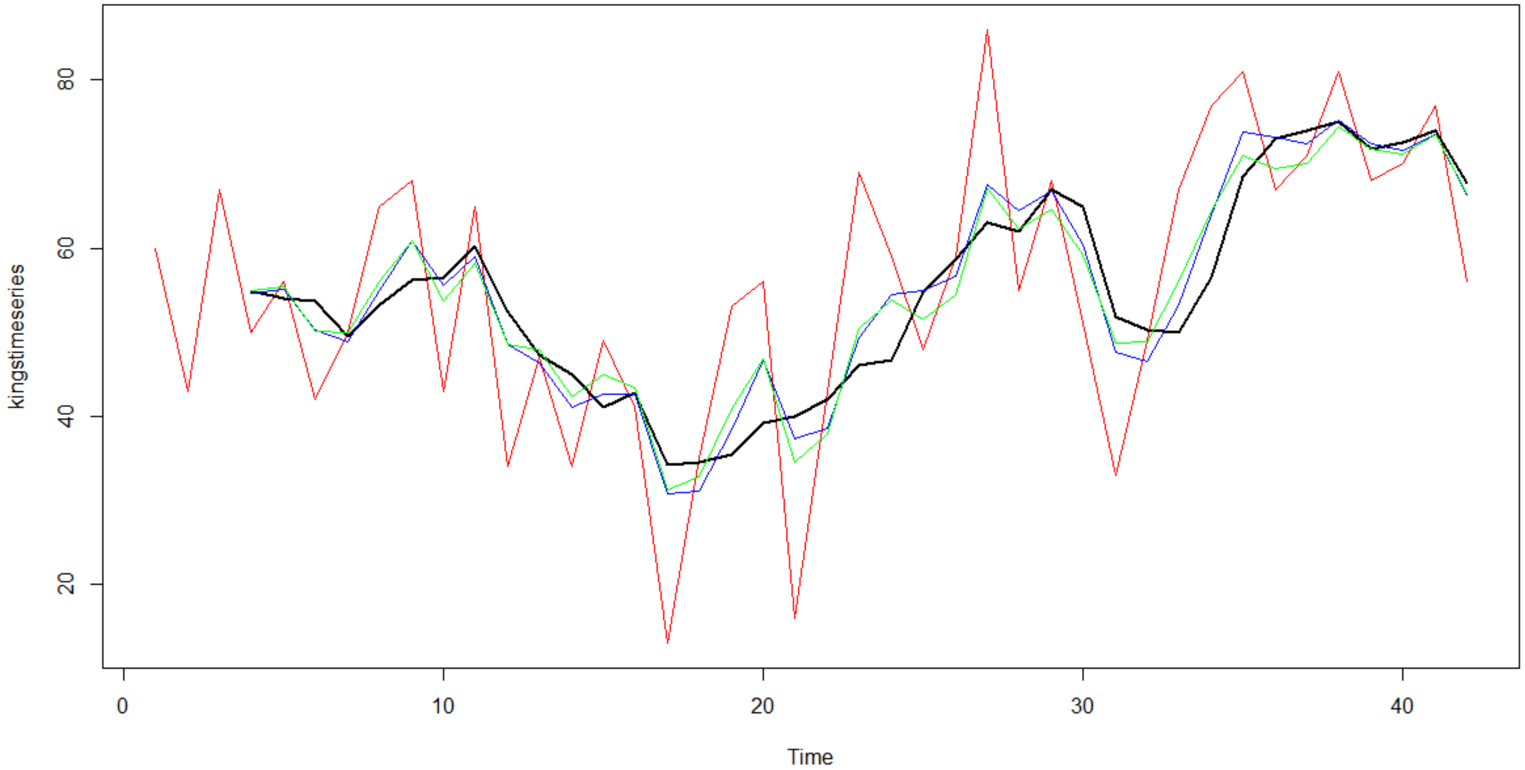
Exponential smoothing

- Y at t+1 $$\widehat{Y_{t+1}} = \widehat{Y_t} + \alpha(Y_t - \widehat{Y_t})$$
- Y at t+2
- All future predictions are same! This is in accordance with **stationary** assumption.

EMA

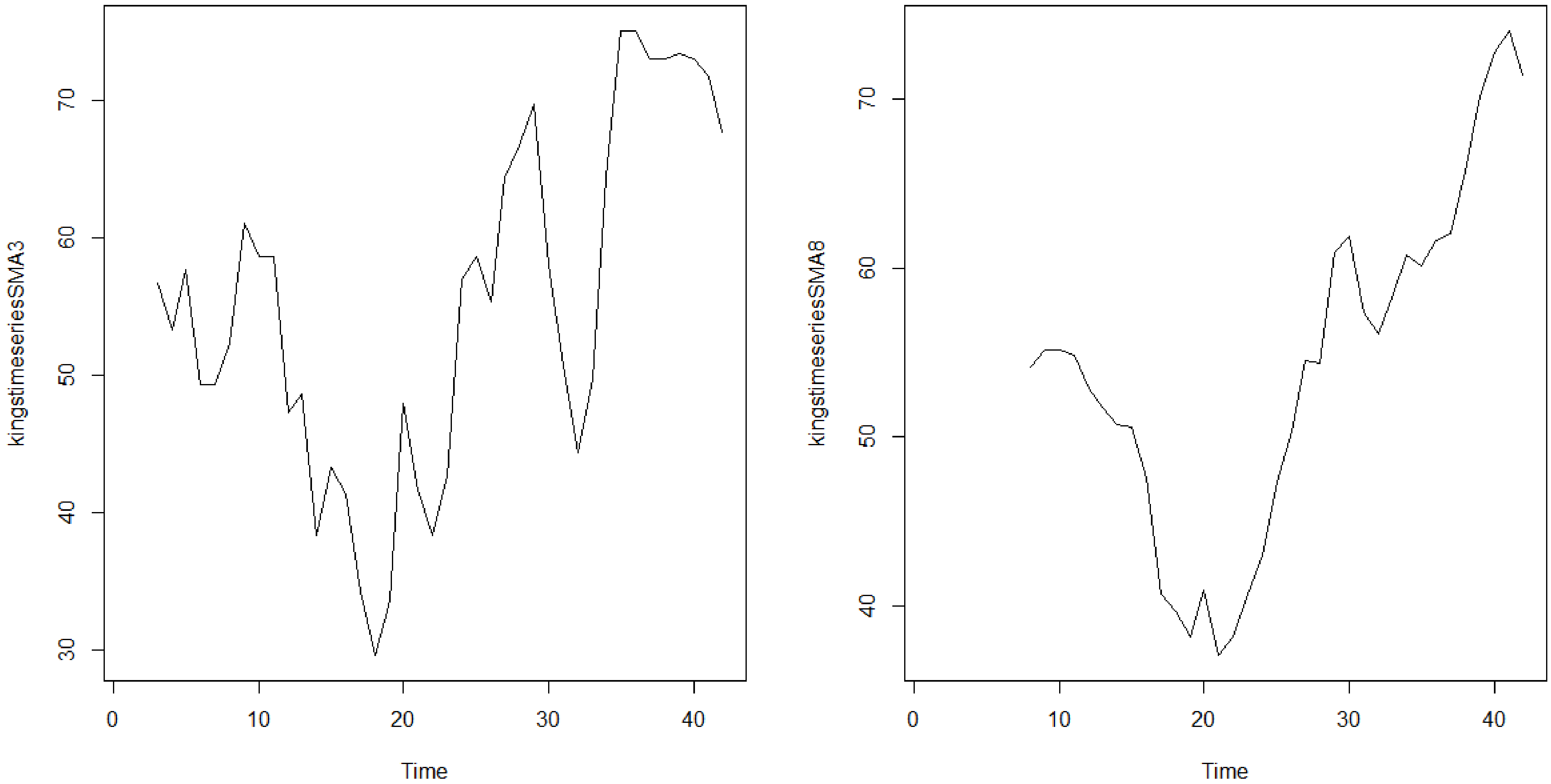

25
20
15
10
5
0
Number of products sold
exp, a=0.1
1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19 20 21 22 23 24

SMA, WMA and EMA – Kings' ages at death



errorEMA	7.45191533669607
errorSMA	9.60897435897436
errorWMA	7.91794871794872

Effect of *k* – Kings’ ages at death



CSE 7202c

ADDING TREND AND SEASONALITY TO MOVING AVERAGE PROCESSES

Holt-Winters method with Additive Seasonality

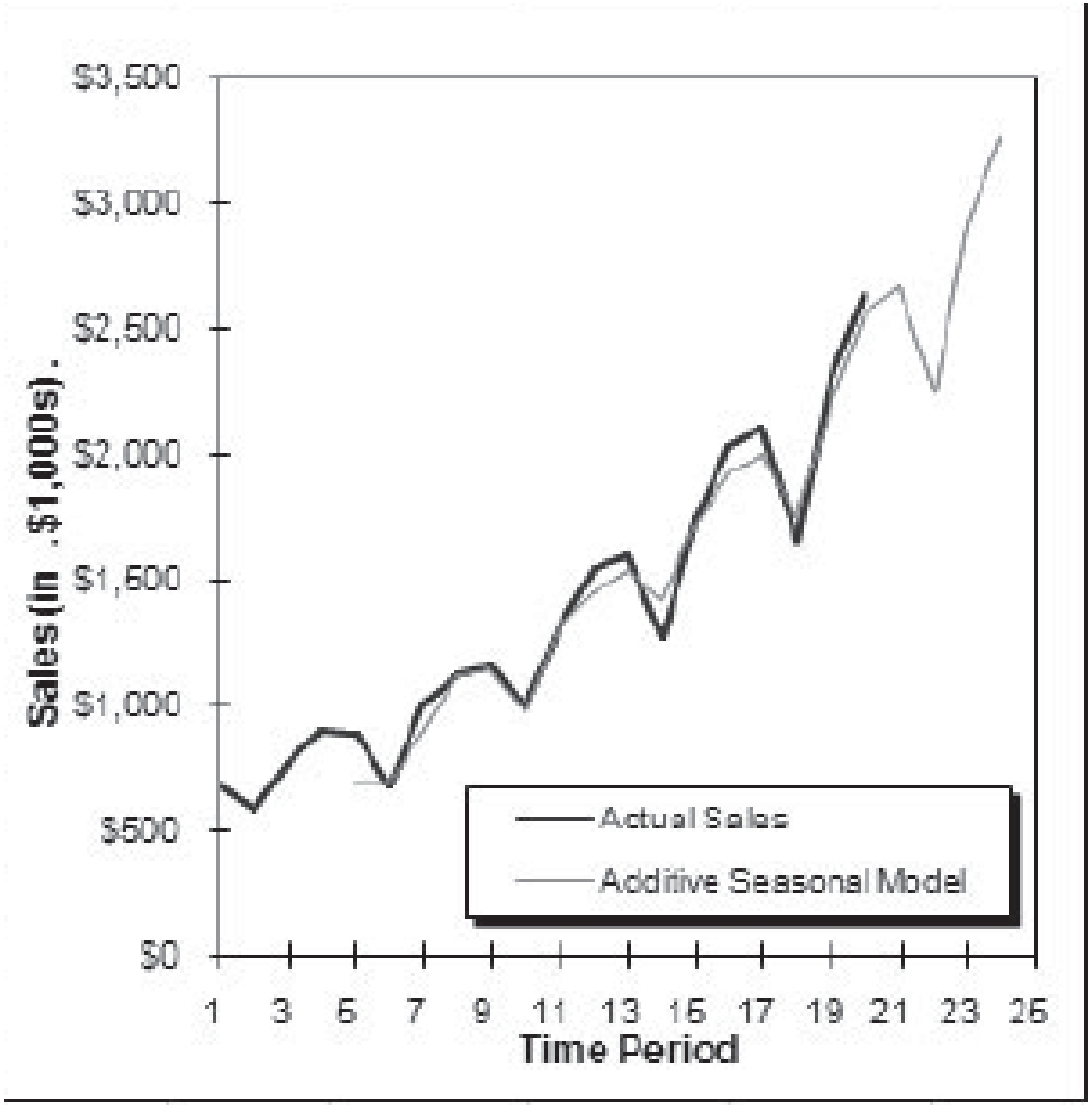
$$\hat{Y}_{t+n} = E_t + nT_t + S_{t+n-p}$$

Random Trend Seasonality

$$E_t = \alpha(Y_t - S_{t-p}) + (1-\alpha)(E_{t-1} + T_{t-1})$$

$$T_t = \beta(E_t - E_{t-1}) + (1-\beta)T_{t-1}$$

$$S_t = \gamma(Y_t - E_t) + (1-\gamma)S_{t-p}$$


$3,500
$3,000
$2,500
$2,000
$1,500
$1,000
$500
$0
Sales (in $1,000s)
1 3 5 7 9 11 13 15 17 19 21 23 25
Time Period
Actual Sales
Additive Seasonal Model

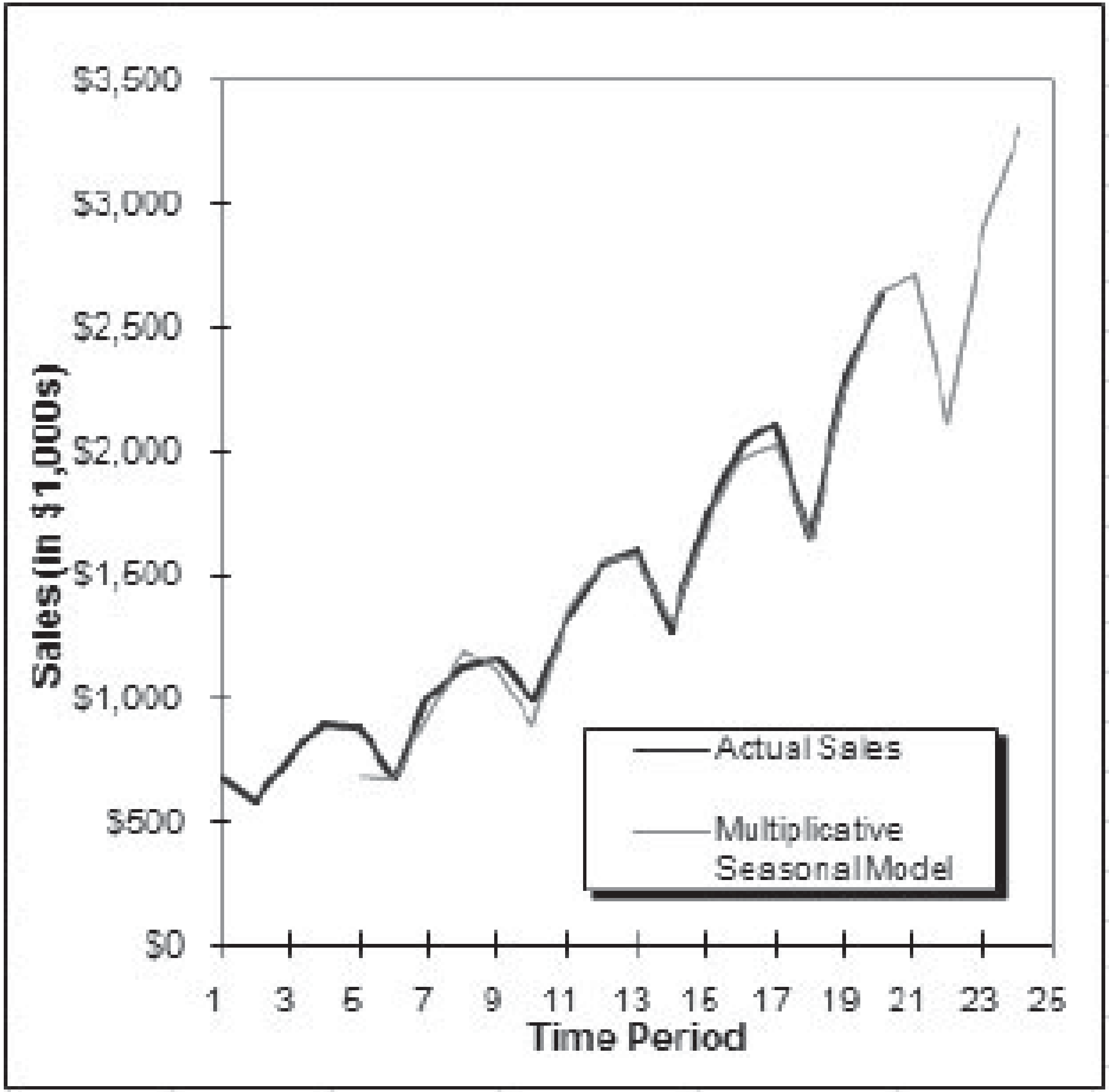
Holt-Winters method with Multiplicative Seasonality

$$\hat{Y}_{t+n} = (E_t + nT_t)S_{t+n-p}$$

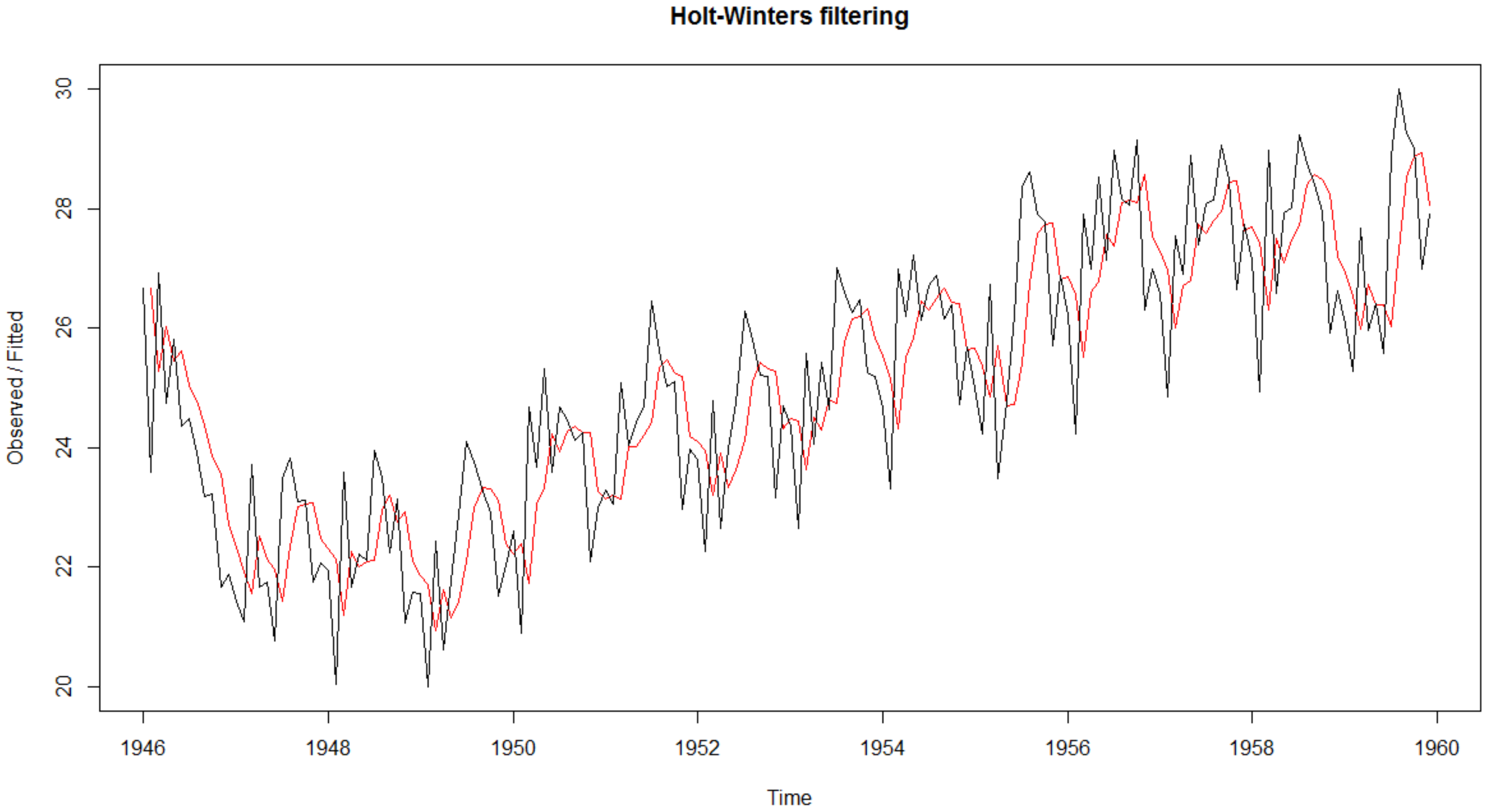
$$E_t = \alpha \frac{Y_t}{S_{t-p}} + (1-\alpha)(E_{t-1} + T_{t-1})$$

$$T_t = \beta(E_t - E_{t-1}) + (1-\beta)T_{t-1}$$

$$S_t = \gamma \frac{Y_t}{E_t} + (1-\gamma)S_{t-p}$$

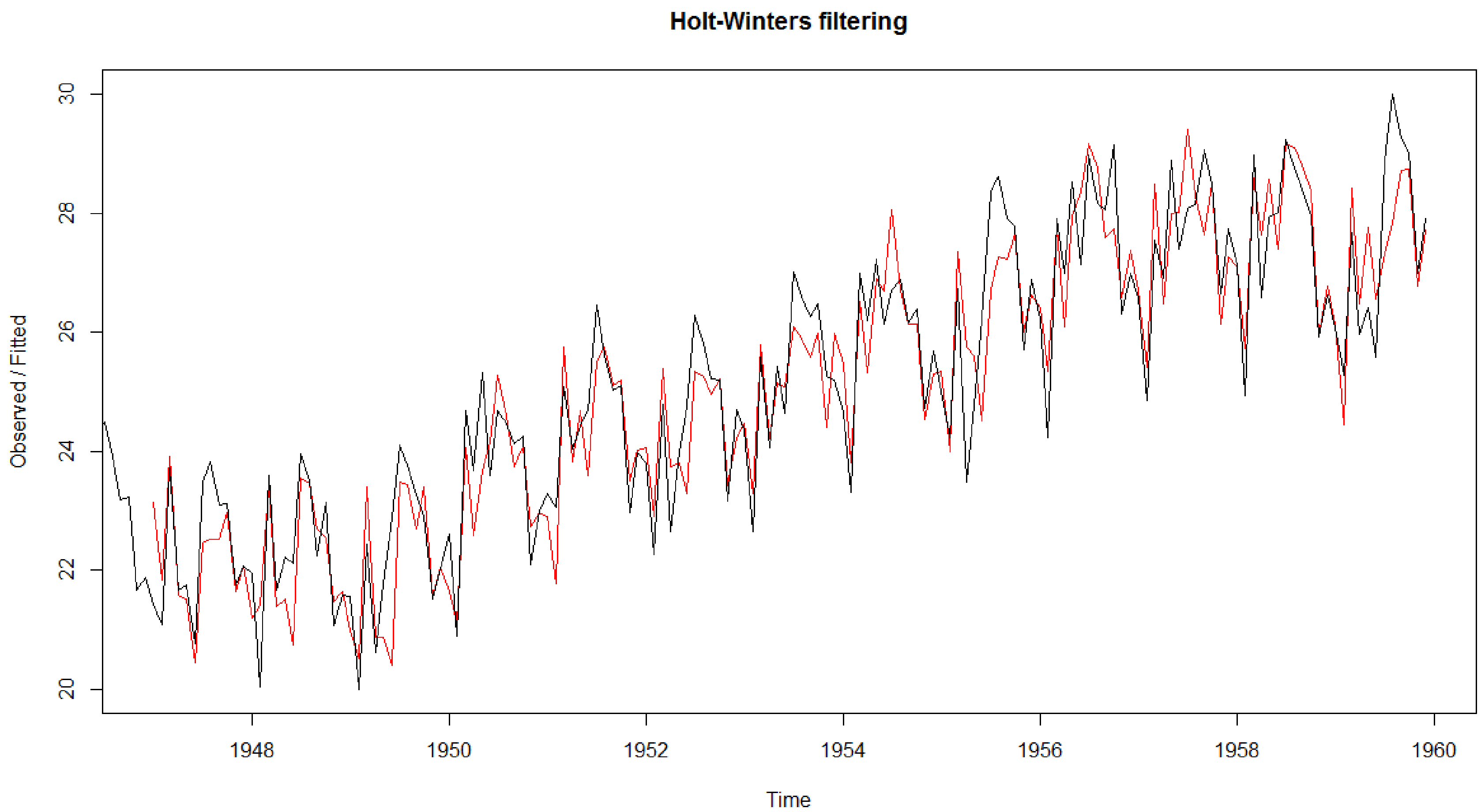

$3,500
$3,000
$2,500
$2,000
$1,500
$1,000
$500
$0
Sales (in $1,000s)
1 3 5 7 9 11 13 15 17 19 21 23 25
Time Period
Actual Sales
Multiplicative Seasonal Model

Holt-Winters method: Only Randomness




```
> birthsforecast$SSE
[1] 281.8759
```

Holt-Winters method: All Components

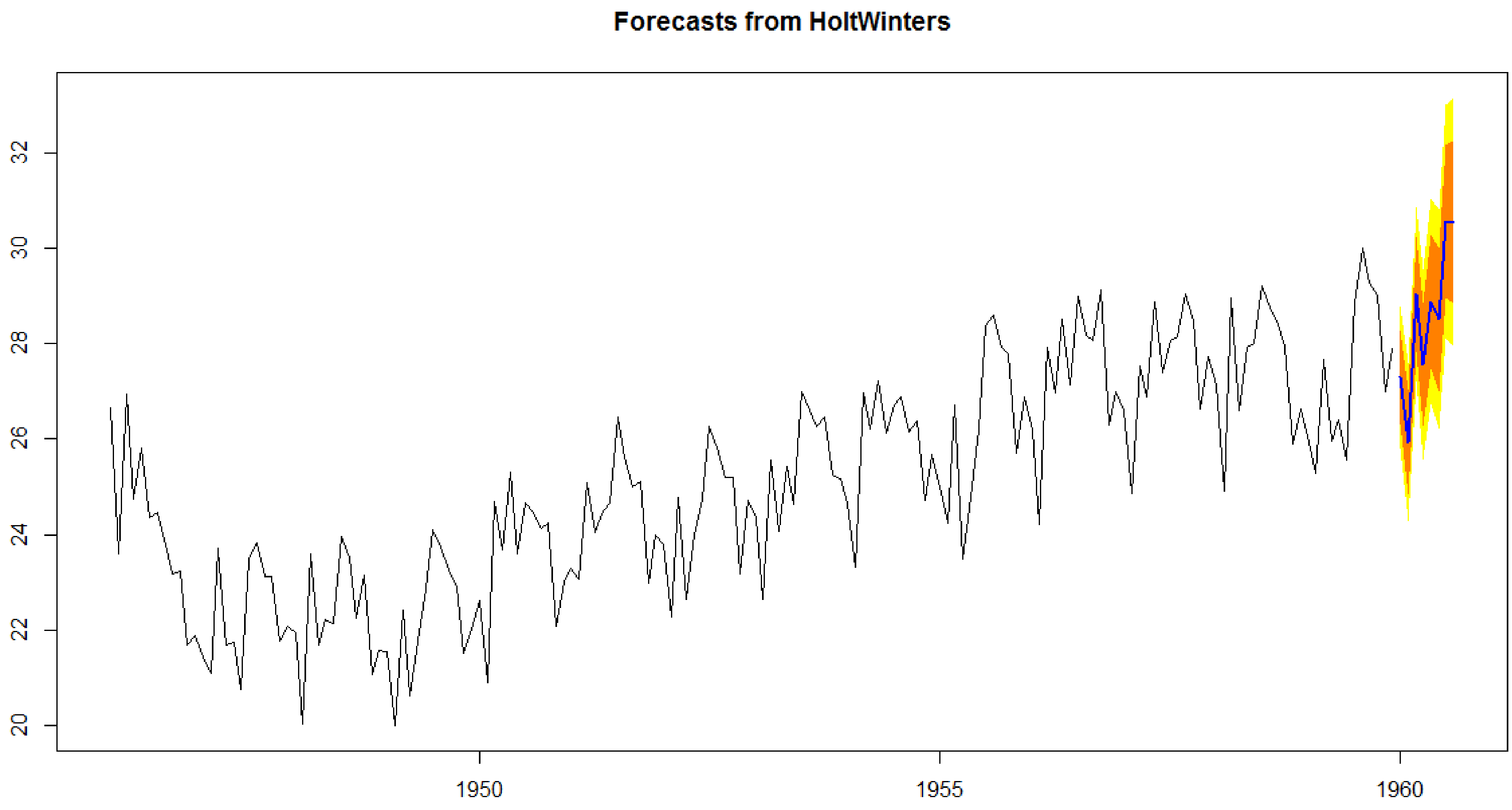



```
> birthsforecast$SSE
[1] 90.94058
```

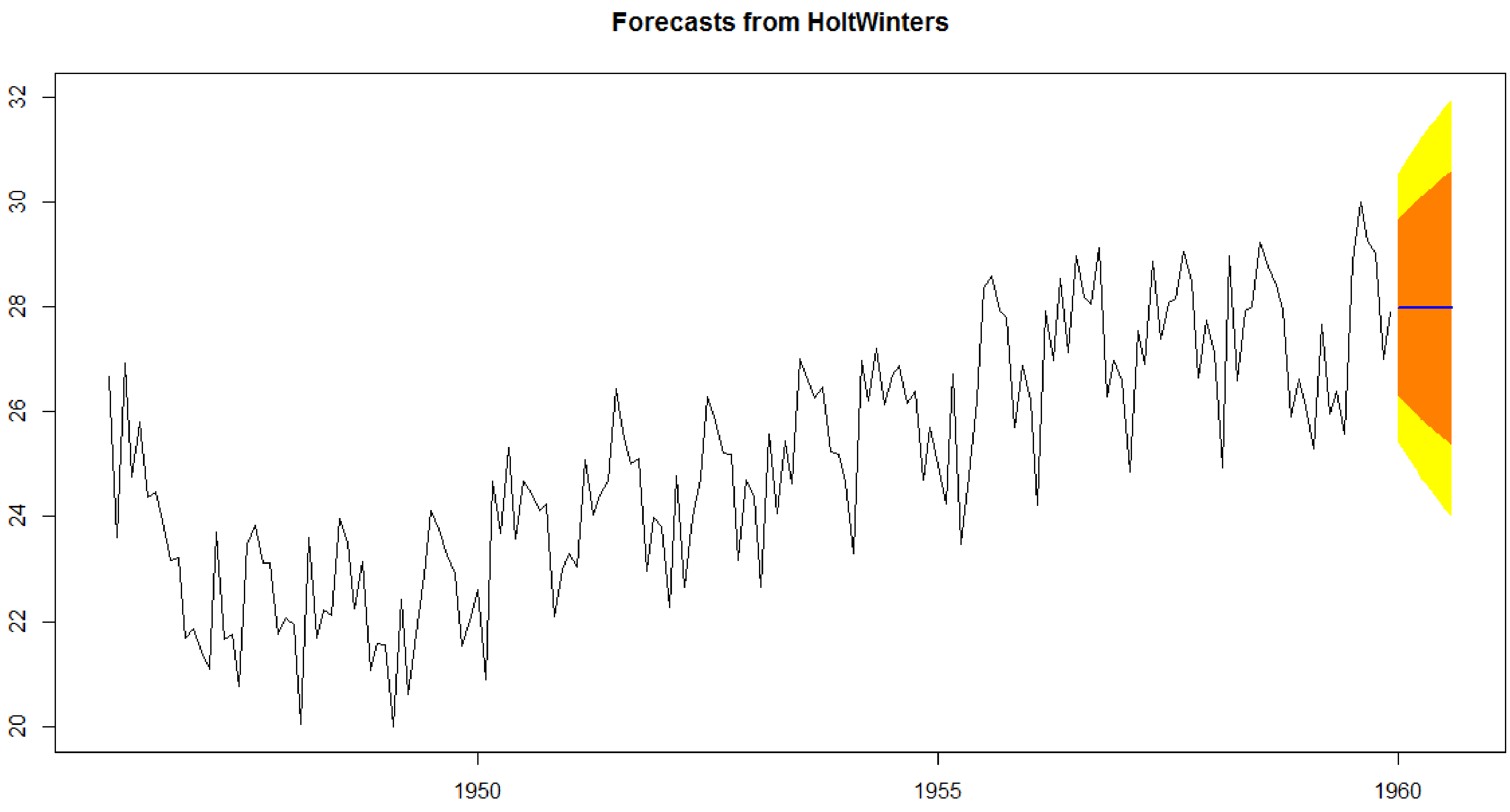
Holt-Winters method: All Components

```
Holt-Winters exponential smoothing with trend and additive seasonal component.

Call:
HoltWinters(x = birthstimeseries)

Smoothing parameters:
 alpha: 0.4823655
 beta : 0.02988495
 gamma: 0.563186

Coefficients:
           [,1]
a    28.04366357
b     0.04199921
s1   -0.78546221
s2   -2.19944507
s3    0.87813012
s4   -0.65164728
s5    0.63427267
s6    0.21182821
s7    2.23177191
s8    2.17167733
s9    1.52077678
s10   1.16900861
s11  -0.97500043
s12  -0.18636055
> birthsforecast$fitted
             xhat    level         trend       season
Jan 1947 23.13579 23.81055 -0.1567618007 -0.51798958
```

Holt-Winters method: Forecasting



Holt-Winters method: Forecasting with no trend and seasonality



Box–Jenkins methodology

- Model identification and model selection.
- Parameter estimation.
- Model checking

- http://www.ncss.com/wp-content/themes/ncss/pdf/Procedures/NCSS/The_Box-Jenkins_Method.pdf

Model selection

- Check ACF, PACF
- Identify important lag periods
- Create a data frame (table) with these past lag values as independent variables and value to be predicted as dependent variable
- Perform autoregression (AR models)
- To incorporate randomness, use MA

SHAPE	INDICATED MODEL
Exponential, decaying to zero	Autoregressive model. Use the partial autocorrelation plot to identify the order of the autoregressive model.
Alternating positive and negative, decaying to zero	Autoregressive model. Use the partial autocorrelation plot to help identify the order.
One or more spikes, rest are essentially zero	Moving average model, order identified by where plot becomes zero.
Decay, starting after a few lags	Mixed autoregressive and moving average model.
All zero or close to zero	Data is essentially random.
High values at fixed intervals	Include seasonal autoregressive term.
No decay to zero	Series is not stationary.

In practice

- There are techniques that automate model selection

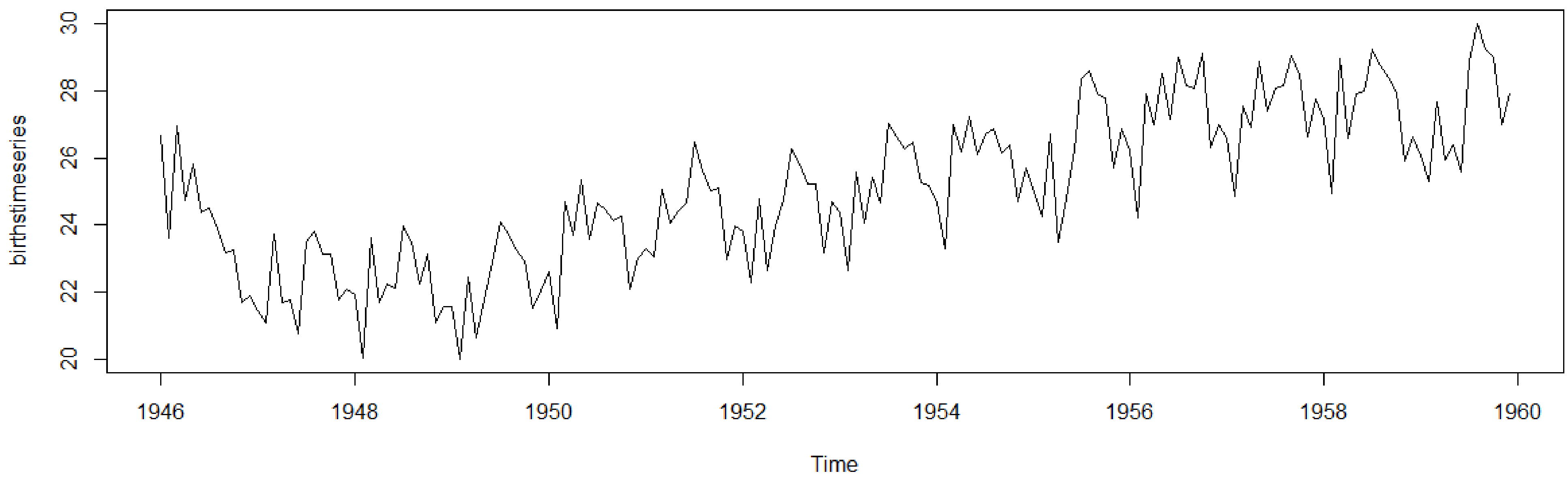
ARIMA(p,d,q) model

- *p* is the number of autoregressive terms (a linear regression of the current value of the series against one or more prior values of the series)
  - Maximum lag beyond which PACF is 0
- *d* is the number of non-seasonal differences, (*d* is the the order of the differencing used to make the time series stationary)

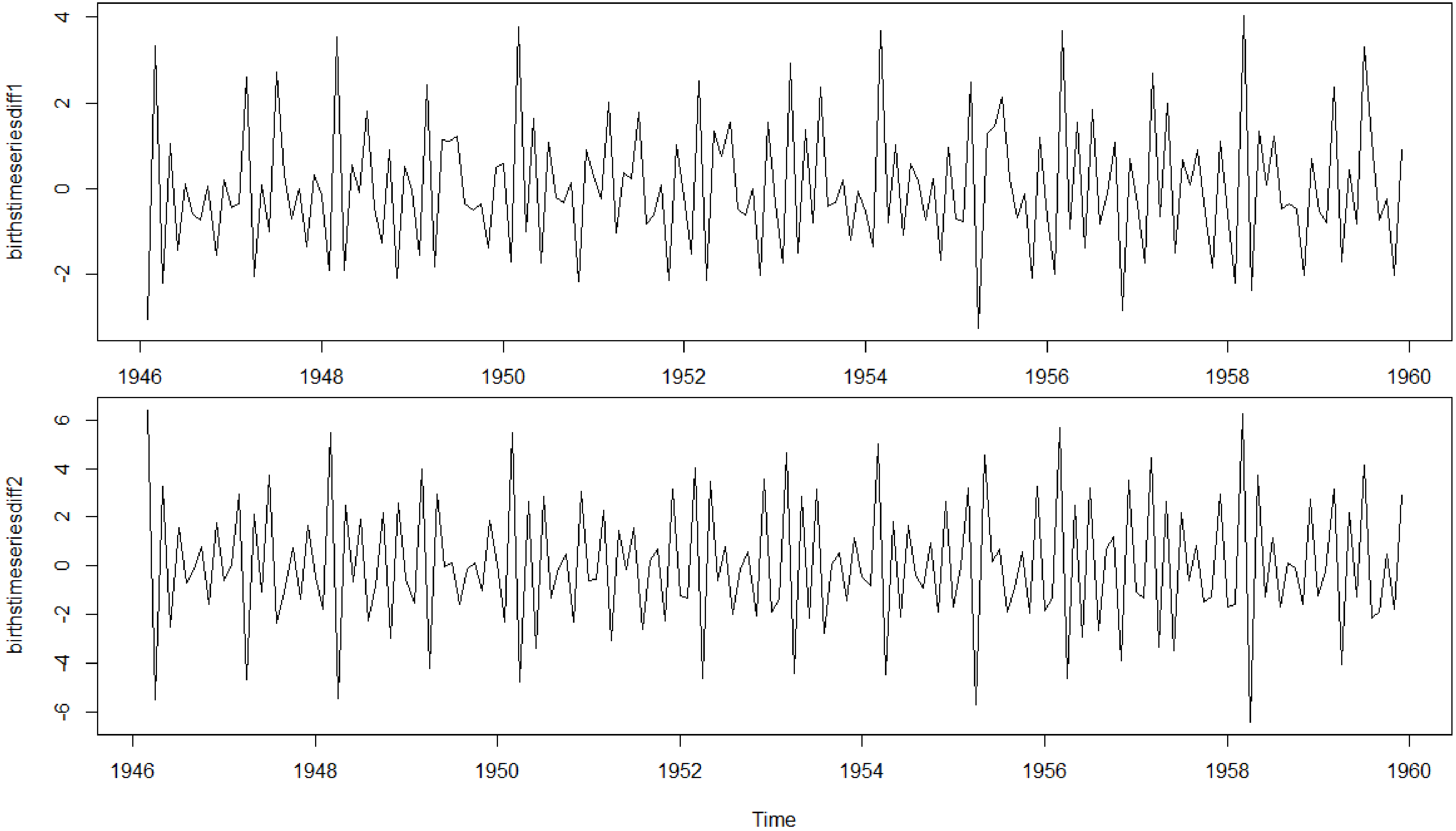
- q is the order of the moving average model (a linear regression of the current value of the series against the white noise or random shocks/spikes of one or more prior values of the series)
  - Maximum lag beyond which the ACF is 0

- Non-seasonal ARIMA models are denoted ARIMA(p,d,q)
- Seasonal ARIMA models are denoted ARIMA(p,d,q)(P,D,Q)$_m$, where m refers to the number of periods in each season and (P,D,Q) refer to the autoregressive, differencing and moving average terms of the seasonal part of the ARIMA model.

ARIMA(p,d,q)


birthstimeseries
20
22
24
26
28
30
1946
1948
1950
1952
1954
1956
1958
1960
Time

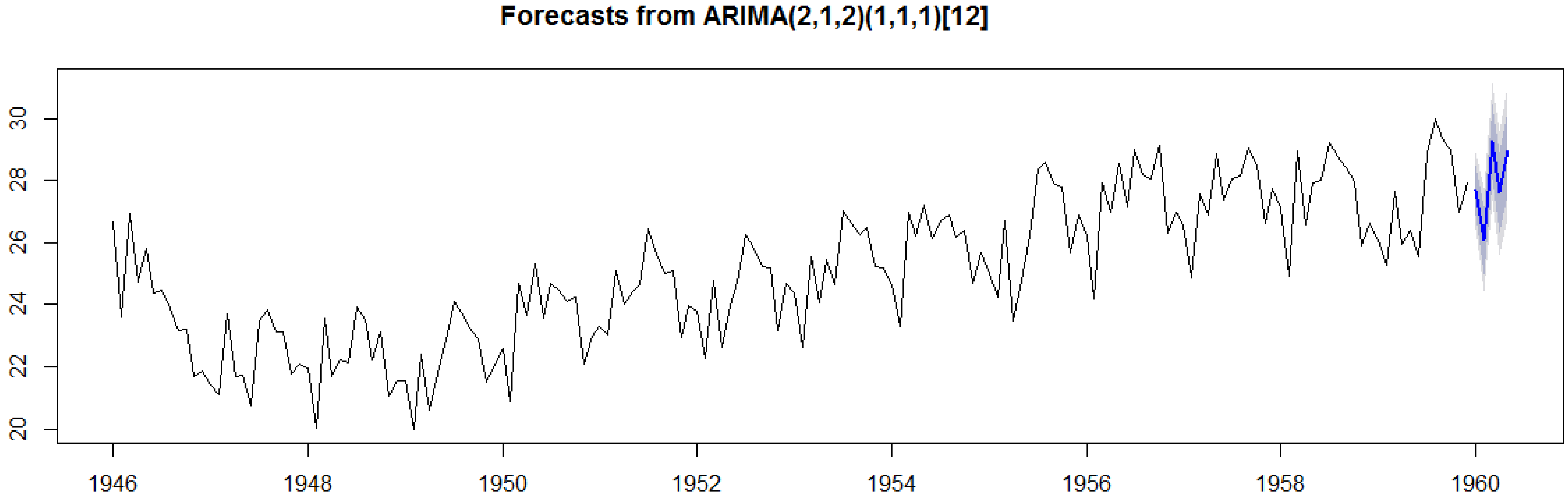
ARIMA(p,1,q) and ARIMA(p,2,q)



Seasonal ARIMA model

```
Series: birthstimeseries
ARIMA(2,1,2)(1,1,1)[12]

Coefficients:
         ar1      ar2      ma1     ma2     sar1     sma1
      0.6539  -0.4540  -0.7255  0.2532  -0.2427  -0.8451
s.e.  0.3004   0.2429   0.3228  0.2879   0.0985   0.0995

sigma^2 estimated as 0.3918:  log likelihood=-157.45
AIC=328.91   AICc=329.67   BIC=350.21
>
```

Seasonal ARIMA model - Forecast



TS has more cells, fridges

■ 6% of all salaried households in rural TS pay taxes, in AP it's 3%

DC CORRESPONDENT
HYDERABAD, JULY 3

As per the Socio Economic and Caste Census 2011 released on Friday, there are about 83.06 lakh households in Telangana State to Andhra Pradesh's 1.22 crore households.

The first installment of the SECC report that was released on Friday covers the rural parts of the country.

Rural Telangana has about 57.06 lakh households while AP has 92.97 lakh households. Despite this, rural TS has a higher number of salaried households than AP. Also, more salaried households in rural TS pay income or professional taxes than AP.

About 6 per cent of all salaried households in rural TS pay taxes while in AP it's half the number.

Also, more people in Telangana own mobiles phones and refrigerators. More than 83 per cent of the rural population in TS own mobile phones.

The land ownership patterns in the two states have also thrown up interesting facts. Unirrigated land in AP constitutes about 24 per cent while in TS it is 31 per cent.

Proportion of agricultural lands, which receive assured irrigation water in the two states, is the same at 36 per cent.

The quantum of land though is higher in AP. AP, however, has beaten TS in terms of education as a higher proportion of the TS rural population is still illiterate.

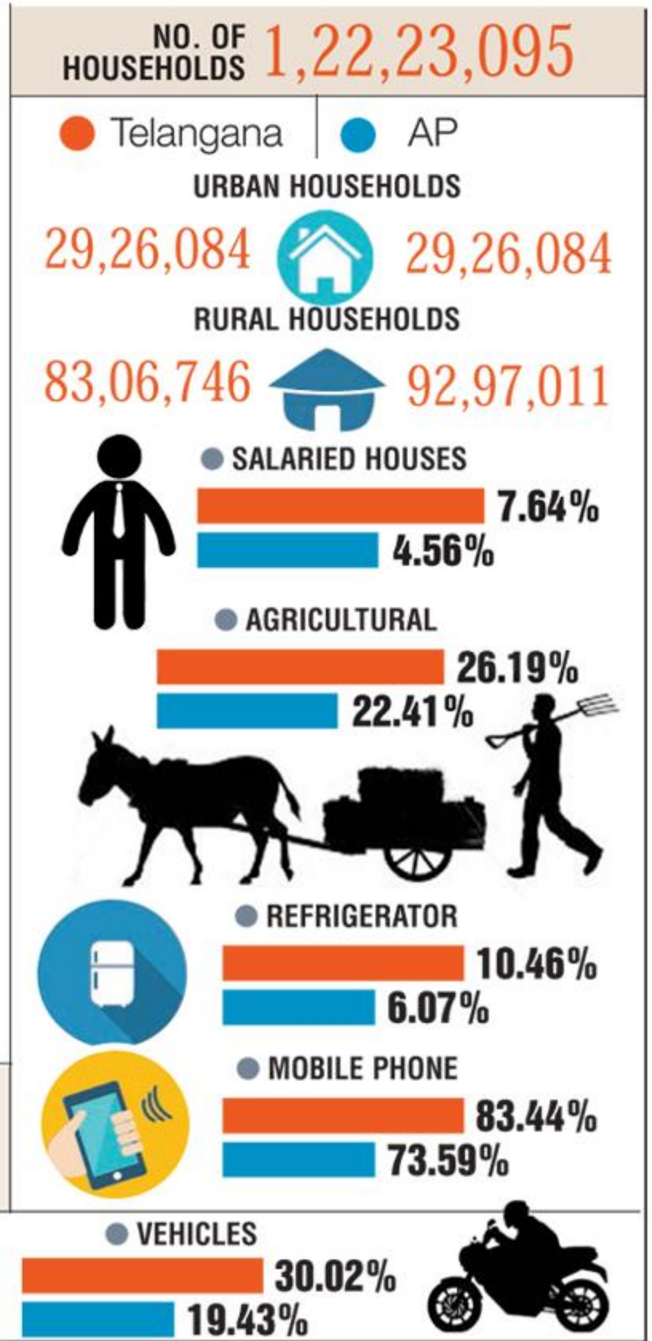
Majority of the rural populations in the two states derive their income from manual casual labour.

A whopping 59 per cent of AP's rural population comprises casual labourers while nearly half of the rural TS population comprises casual labourers.

The SECC 2011 data shows that about 26 per cent of the TS rural population derives its income from cultivation. The corresponding figure in AP is lesser at 22 per cent.

> A whopping 59 per cent of AP's rural population comprises casual labourers while nearly half TS population comprises casual labourers
>
> — SOCIA-ECONOMIC CASTE CENSUS

Telangana	AP
Kuchcha Households: 12.38%	Kuchcha Households: 9.12%
Pucca Households: 87.52%	Pucca Households: 90.65%
Houses with 1 room: 13.50%	Houses with 1 room: 26.12%



CSE 7202c

Vote to decide if Greece gets a last-ditch financial rescue

Slim lead for 'Yes' in Greece referendum

Athens, July 3: Supporters of Greece's bailout terms have taken a wafer-thin opinion poll lead over the 'No' vote backed by the leftist government, 48 hours before a referendum that may determine the country's future in the euro zone.

The poll by the respected ALCO institute, published in the *Ethnos* on Friday, put the 'Yes' camp on 44.8 per cent against 43.4 per cent for the 'No' vote. But the lead was well within the pollster's 3.1 percentage point margin of error, with 11.8 per cent saying they are still undecided.

Given a volatile public mood and a string of recent election results that ran counter to opinion poll predictions, the result is in effect completely open.

With banks shuttered all week, cash withdrawals

PROBLEMS IN FOCUS

IN ORDER TO WALK OUT OF THE HOUR OF ECONOMIC CRISIS, GREECE HAS TO ADDRESS SOME OF THESE PROBLEMS

DEFUSING PENSION BOMB

Olivier Passet, director of Greece's economy analysis, Xerfi, said resolving Greece's pension "time bomb... is priority number one."

International School of Engineering

Plot 63/A, 1st Floor, Road # 13, Film Nagar, Jubilee Hills, Hyderabad - 500 033

For Individuals: +91-9502334561/63 or 040-65743991

For Corporates: +91-9618483483

Web: http://www.insofe.edu.in

Facebook: https://www.facebook.com/insofe

Twitter: https://twitter.com/Insofeedu

YouTube: http://www.youtube.com/InsofeVideos

SlideShare: http://www.slideshare.net/INSOFE

LinkedIn: http://www.linkedin.com/company/international-school-of-engineering