Time Series Forecasting

- Mohamed Imran

Typical Time Series

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

f can be linear or nonlinear

Important Concept

In time series, the order of observation is of primary importance. So is autocorrelation. They play a very important role in identifying the models and their characteristics

Autocorrelation (ACF) and Partial ACF (PACF)

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

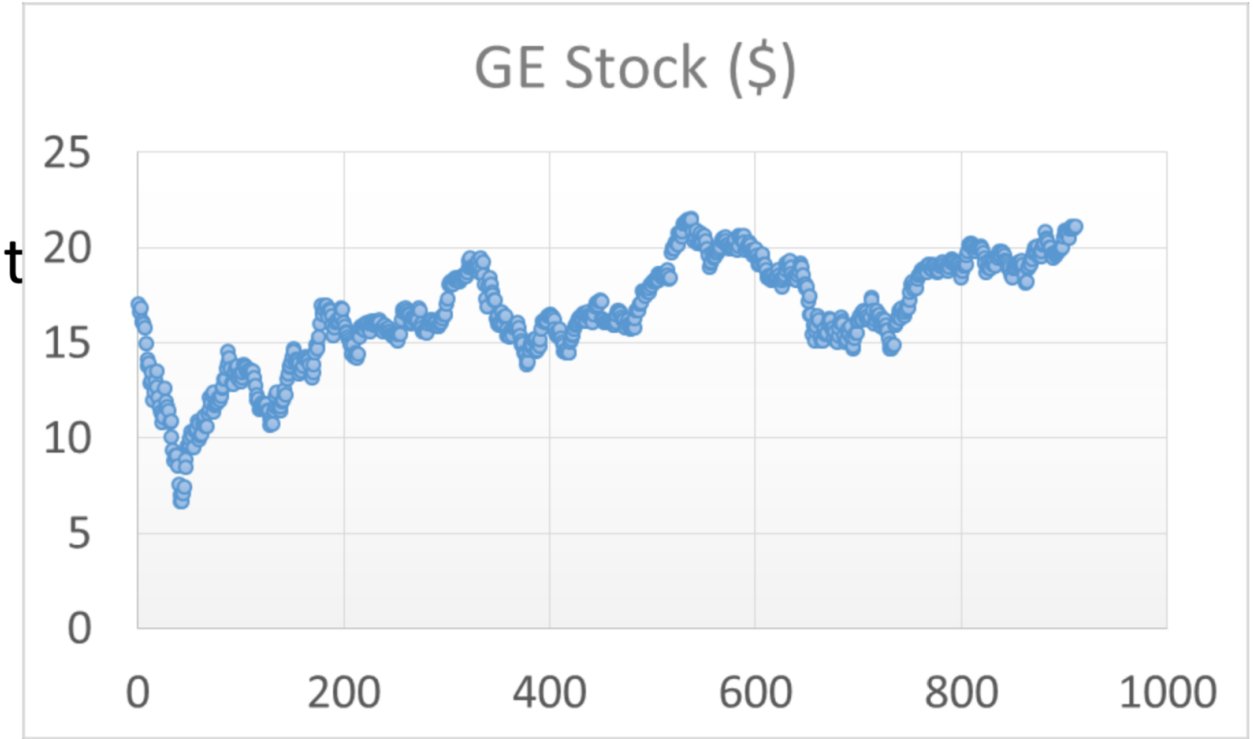
- PACF: The same as ACF with all intermediate correlations removed. It is the k_{th} coefficient of the ordinary least squares regression.

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

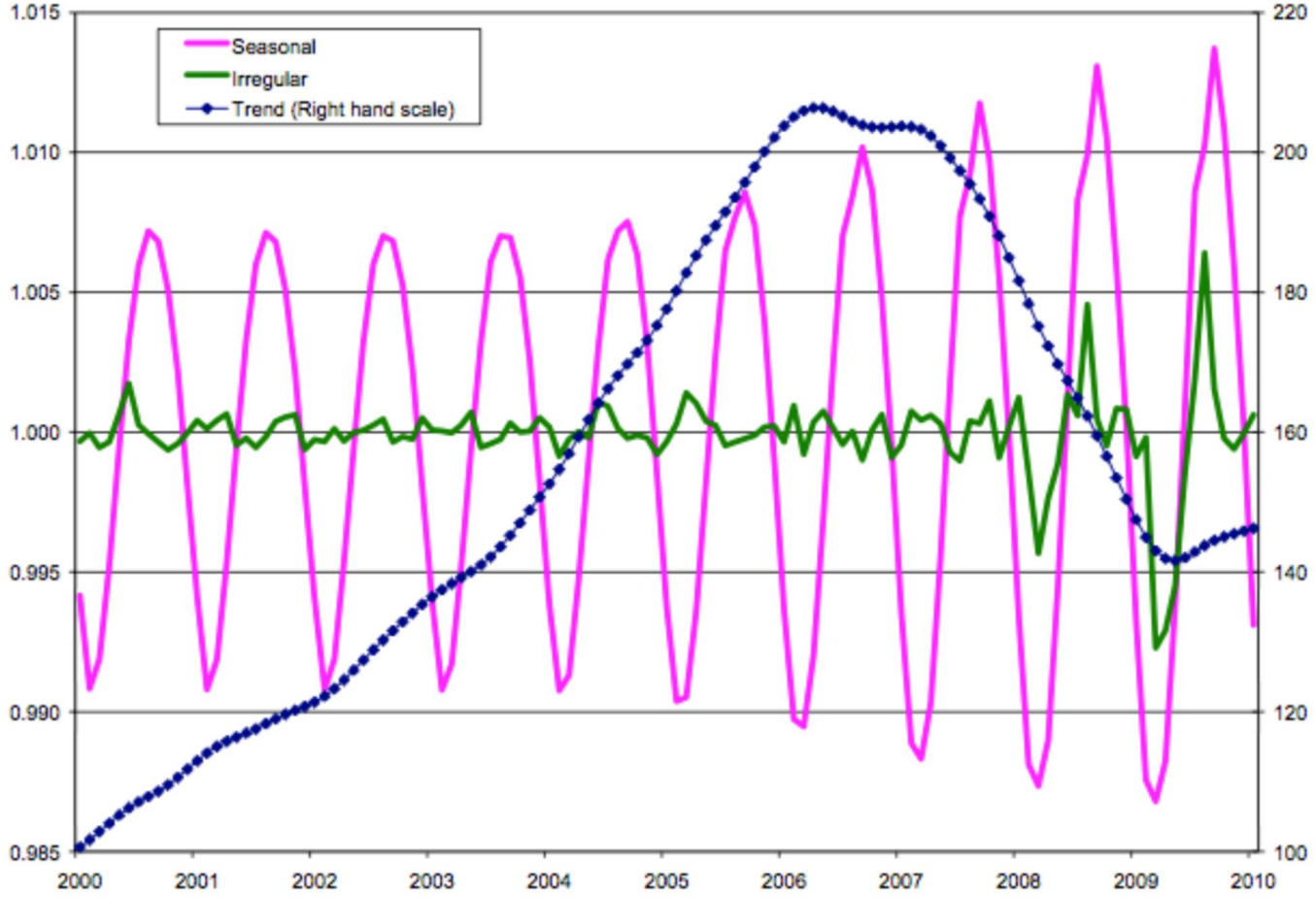
$[y_t]$ is the input time series, *k* is the lag order and β_i is the i_{th} coefficient of the linear multiple regression.

Components of Time Series

- Trend
- Seasonality
- Random component



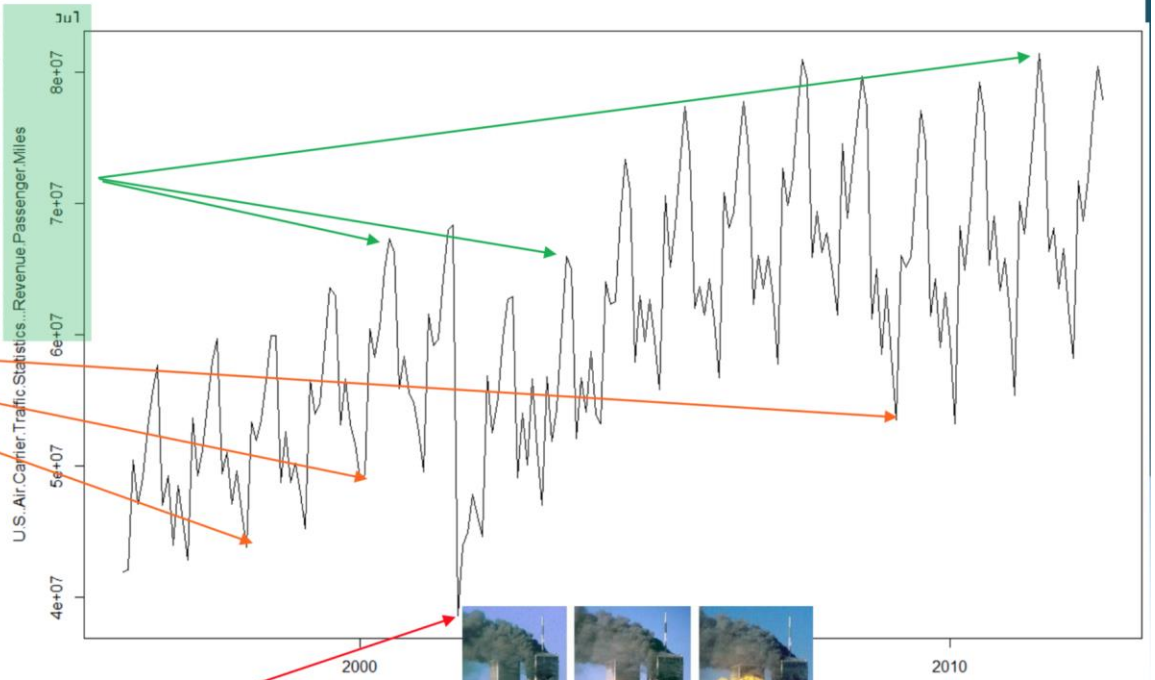
Trend, Seasonality and Randomness



US Air Carrier Traffic – Revenue Passenger Miles ('000)

RPM

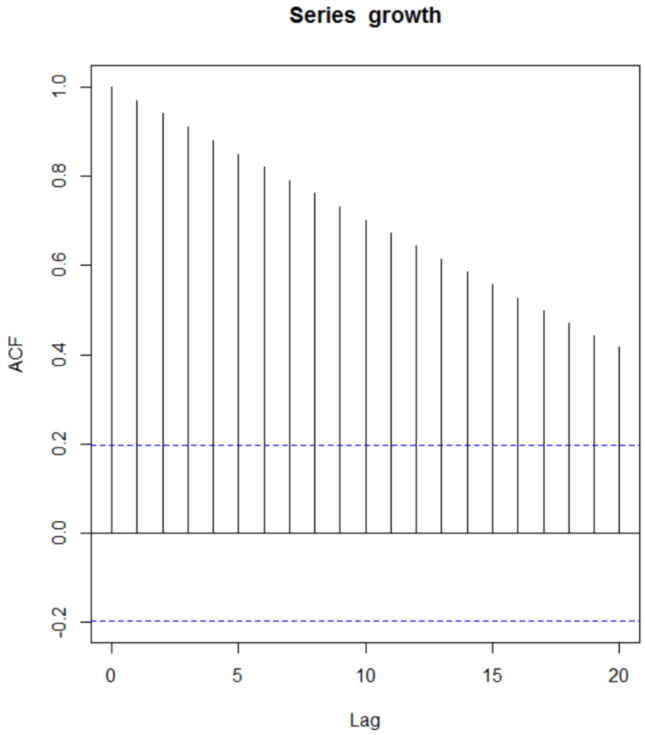
```
> milestimeseries <- ts(miles, frequency = 12, start = c(1996,1))
> milestimeseries
          Jan      Feb      Mar      Apr      May      Jun      Jul
1996 41972194 42054796 50443045 47112397 49118248 52880510
1997 45850623 42838949 53620994 49282817 51191842 54707221
1998 46514139 43769273 53361926 51968480 53515798 56460422
1999 47988560 45241211 56555731 53920855 54674958 59213000
2000 49045412 49306303 60443541 58286680 60533783 64903295
2001 52634354 49532578 61575055 59151645 59662416 64353323
2002 46224031 44615129 56897729 52542164 55116060 59745343
2003 51197175 47040806 56766580 51857453 54335598 60272900
2004 53979786 53179693 64035864 62340117 62530704 68866398
2005 59629608 55795165 70595861 65145552 68268899 72952959
2006 61035027 56729212 70799794 68120559 69352606 74099239
2007 63016013 57793832 72700241 69836156 71933109 76926452
2008 64667106 61504426 74575531 68906882 72725750 76162105
2009 58373786 53506580 66027341 65166300 65868254 71350227
2010 59651061 53240066 68307090 64953250 68850904 74474550
2011 61630362 55391206 70158268 67683558 71711448 76057910
2012 61940180 58243763 71696039 68669228 71887523 76760759
          Aug      Sep      Oct      Nov      Dec
1996 57723208 47035464 49263120 43937074 48539606
1997 59715433 49418190 51058879 47056048 49654209
```

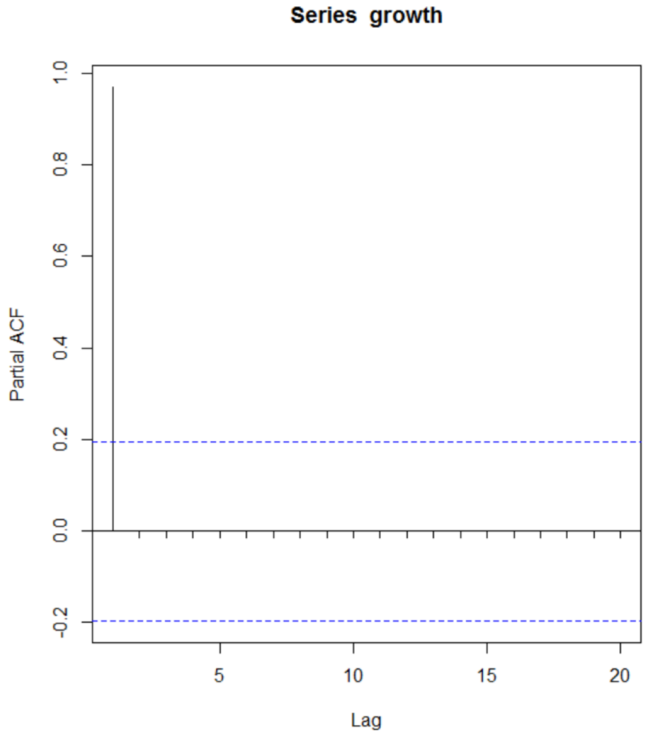


Data sources:
http://www.bts.gov/xml/air_traffic/src/index.xml
and https://datamarket.com/data/set/281x/us-air-carrier-traffic-statistics-revenue-passenger-miles
Last accessed: 31-Mar-2016

	Aug	Sep	Oct	Nov	Dec
1996	57723208	47035464	49263120	43937074	48539606
1997	59715433	49418190	51058879	47056048	49654209
1998	59927214	48751280	52578217	48734375	50208641
1999	63003663	53131972	56653901	53215500	51746821
2000	66256804	55900504	58373996	55590325	54822970
2001	68377080	38601868	43964788	44915764	47836501
2002	62944816	49096035	54019748	50106814	56656594
2003	64989766	52121480	56724551	54128776	58739845
2004	70961522	57881042	63021142	59453943	62680310

ACF and PACF – Idealized Trend

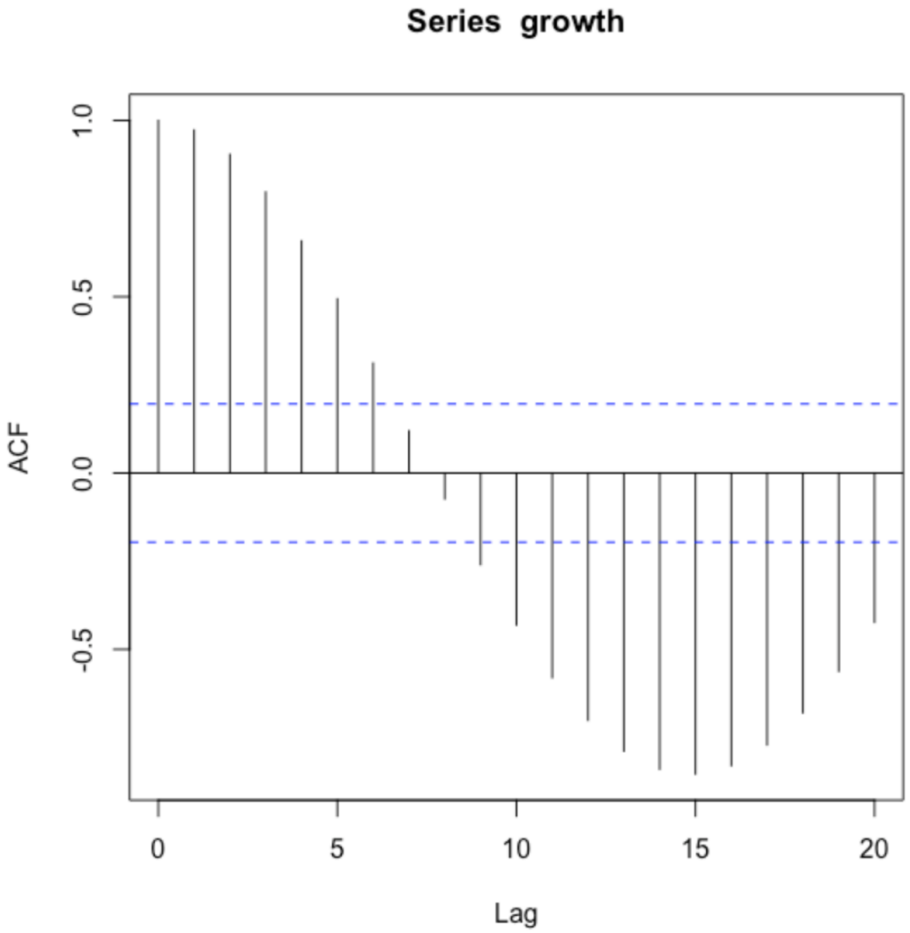


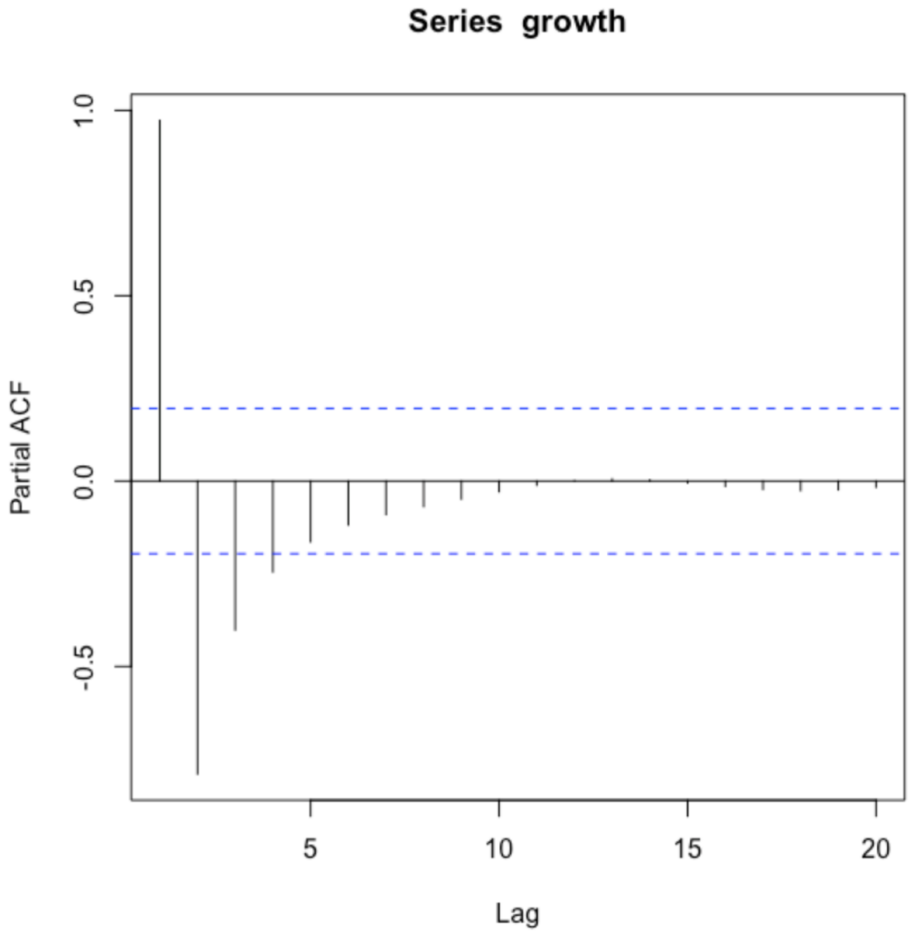


95% CI: $0 \pm \frac{1.96}{\sqrt{n}}$

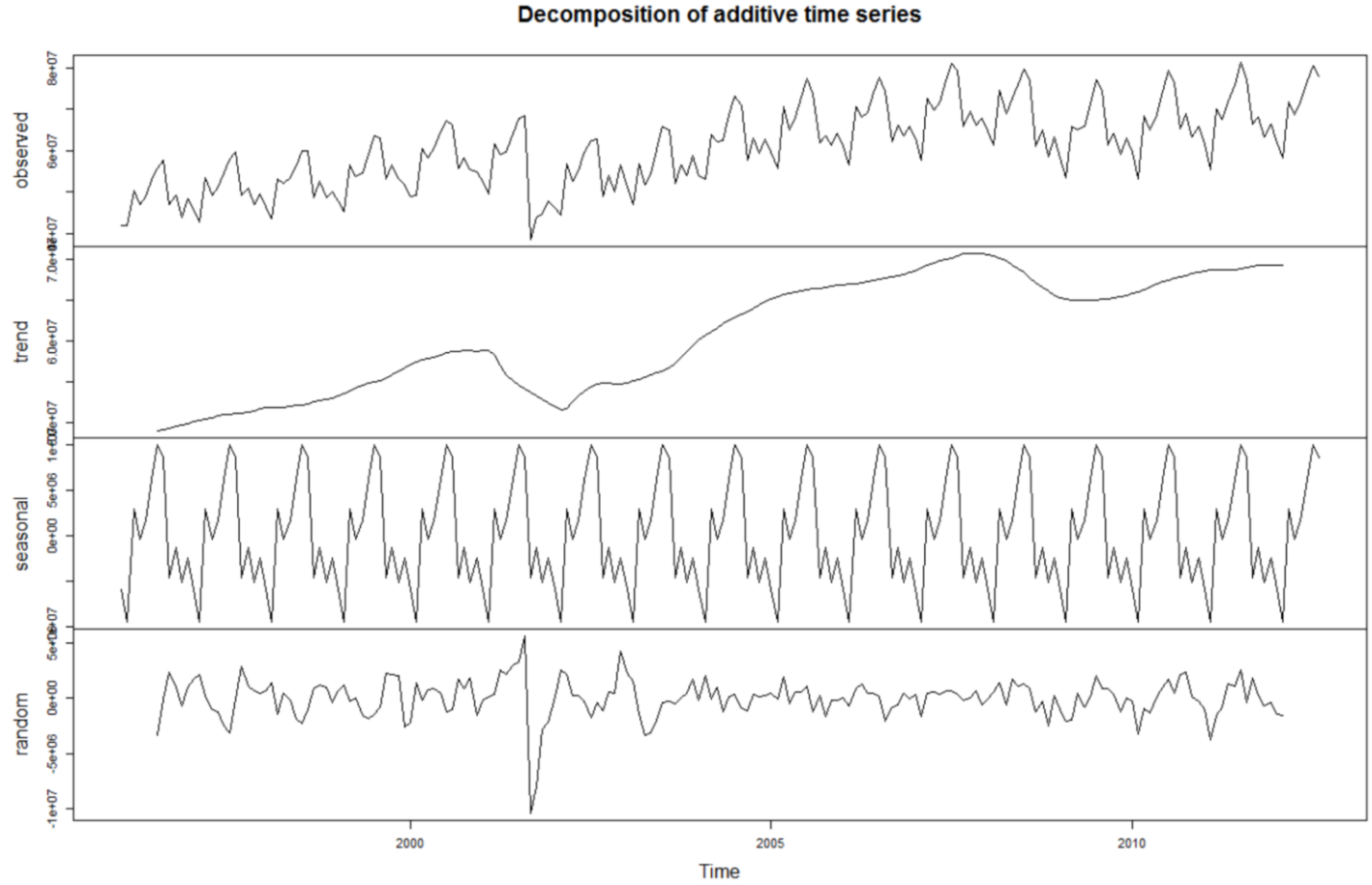
- ACF is a bar chart of correlation coefficients of the time series and its lags.
- PACF is a plot of the partial correlation coefficients of the time series and its lags.

ACF and PACF – Idealized Seasonality





ACF and PACF (Real-world): Decomposing Time Series into the 3 Components – Revenue Passenger Miles (RPM)



Stationary and Non-Stationary

- Stationary data has constant statistical properties – mean, variance, autocorrelation, etc. – over time

- If the data is stationary, forecasting is easier!

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.

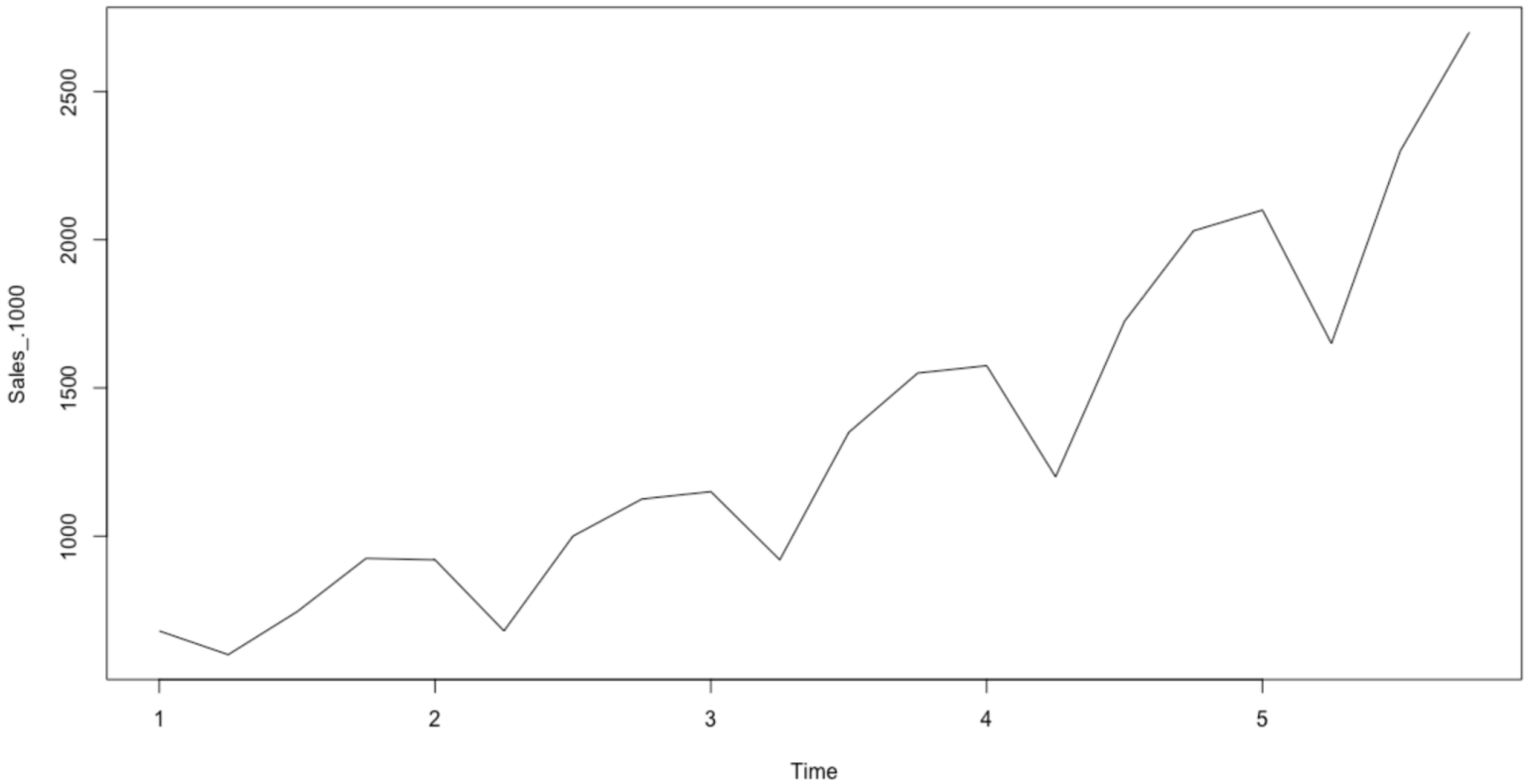
CURVE FITTING / REGRESSION ON TIME METHODS

Regression on Time

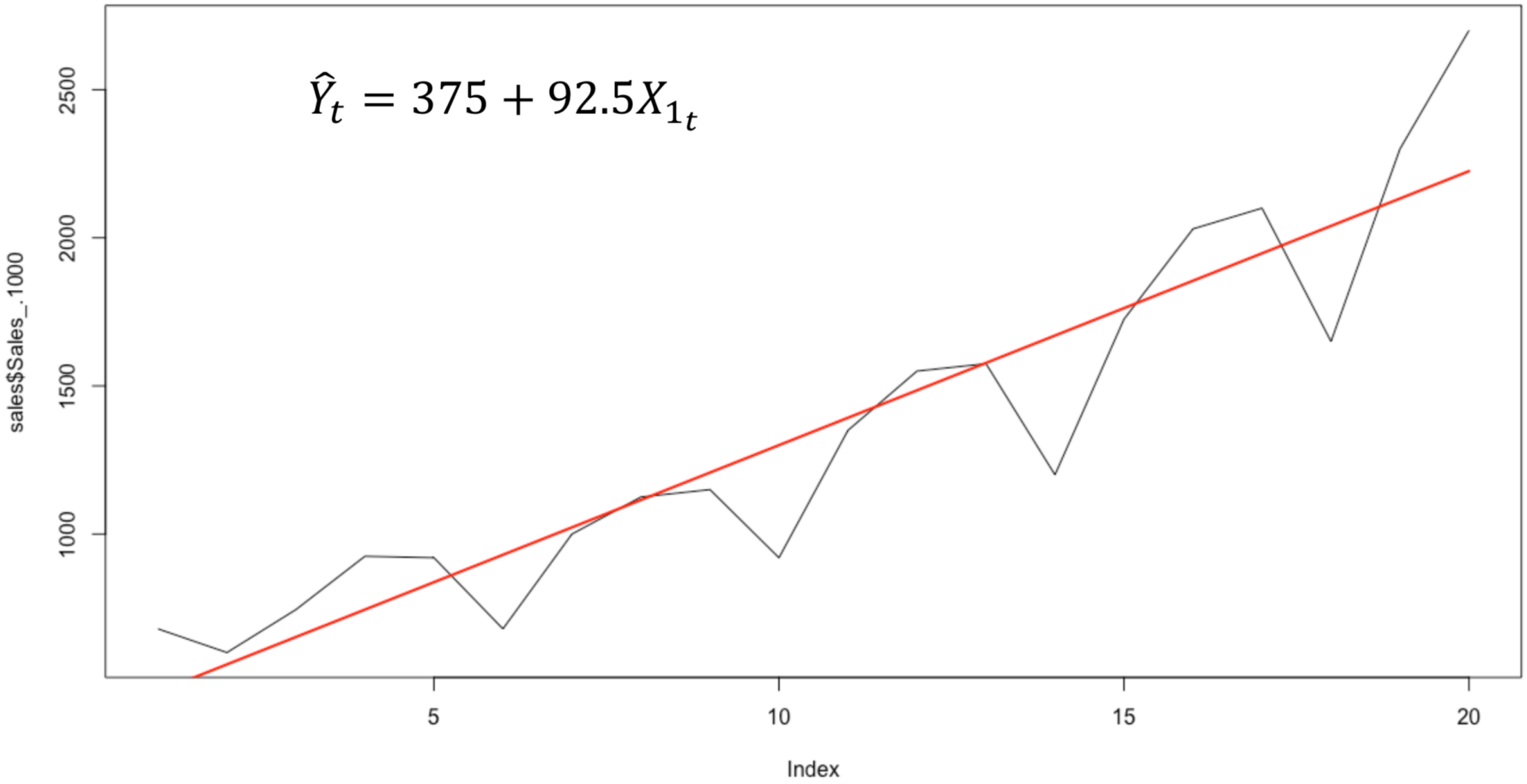
- Use when trend is the most pronounced

- ACF decays exponentially and PACF has very few spikes

Quarter	Sales_$1000
1	680
2	600
3	745
4	925
5	920
6	680
7	1000
8	1125
9	1150
10	920
11	1350
12	1550
13	1575
14	1200
15	1725
16	2030
17	2100
18	1650



Regression Analysis

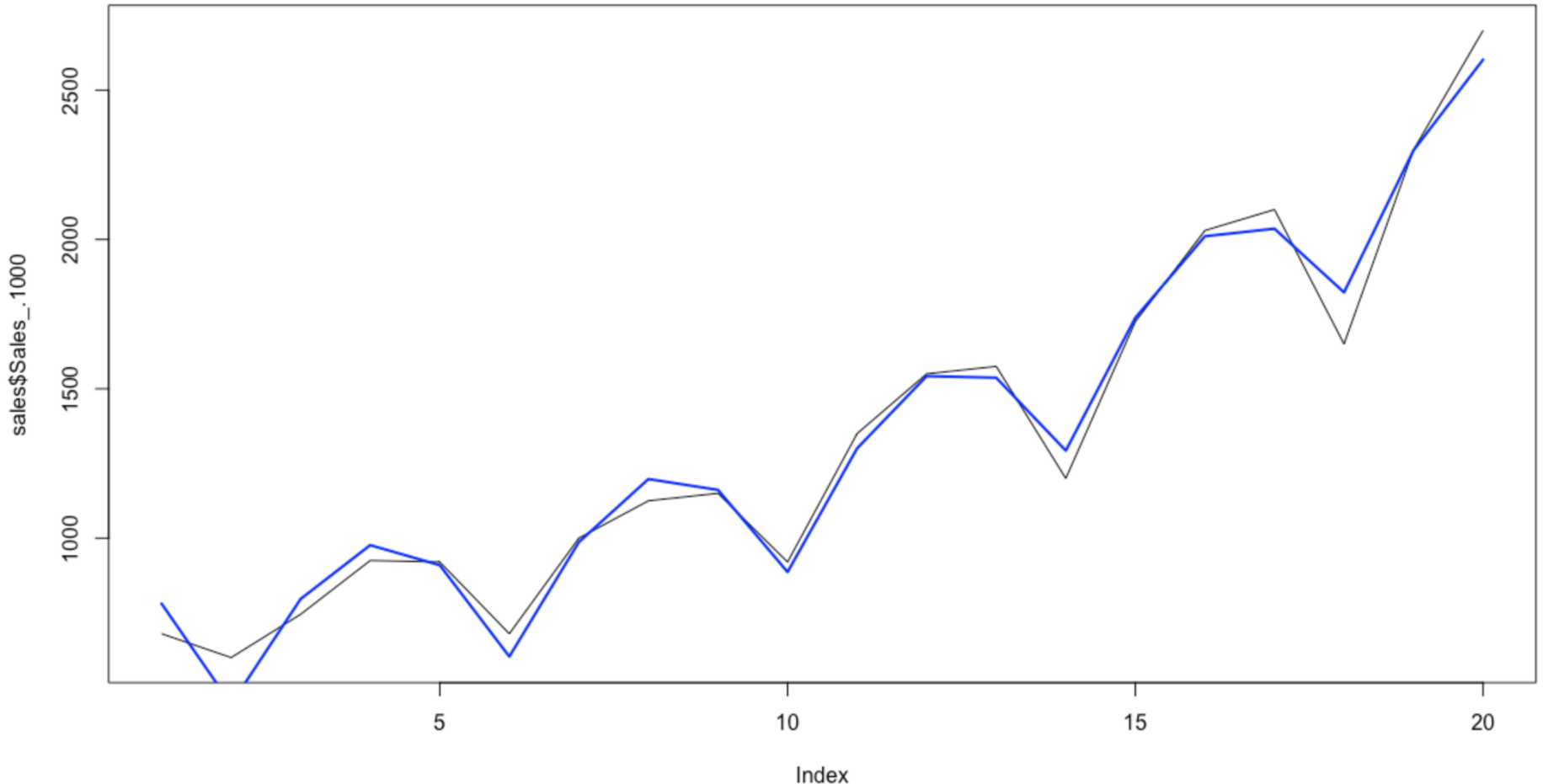


Seasonal Regression Models

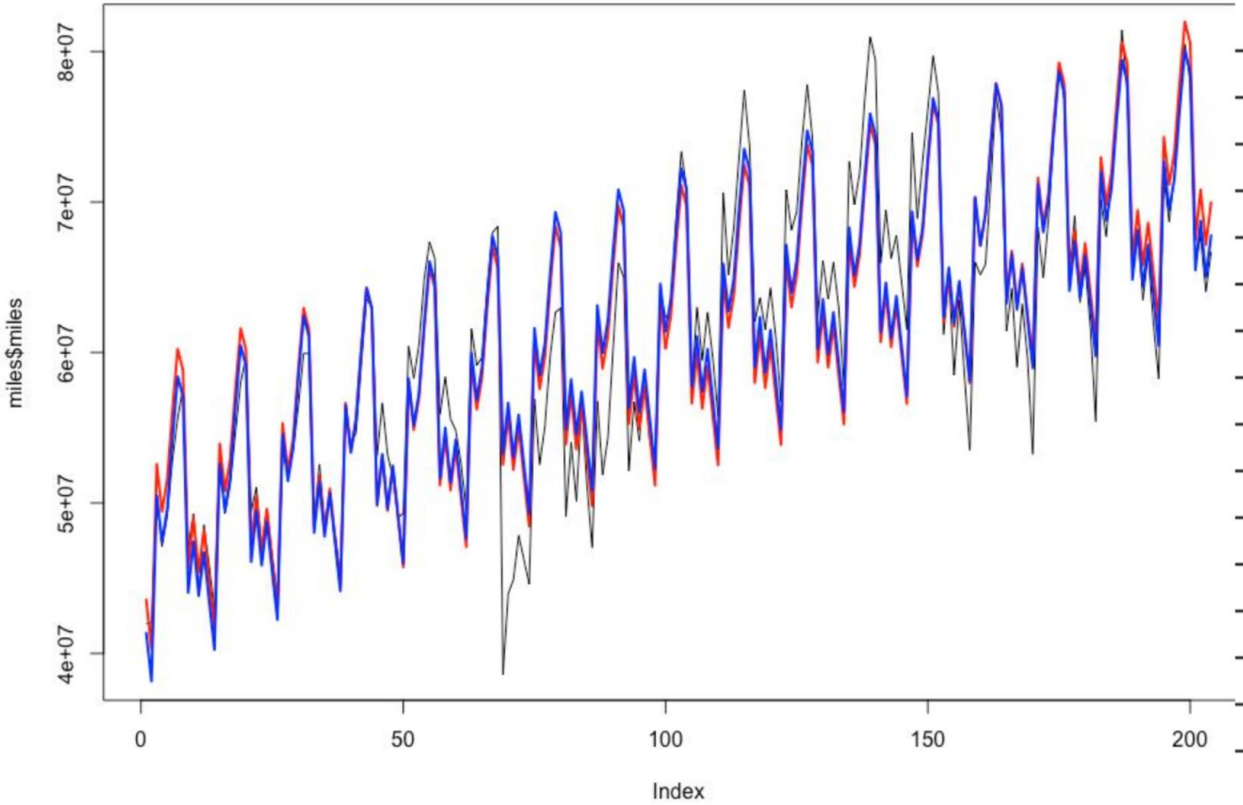
	Value of		
Quarter	X_{3_t}	X_{4_t}	X_{5_t}
1	1	0	0
2	0	1	0
3	0	0	1
4	0	0	0

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

Seasonal Regression Models



Seasonal Regression Models - RPM



	miles	time	seasonal
1	41972194	1	1
2	42054796	2	2
3	50443045	3	3
4	47112397	4	4
5	49118248	5	5
6	52880510	6	6
7	55664750	7	7
8	57723208	8	8
9	47035464	9	9
10	49263120	10	10
11	43937074	11	11
12	48539606	12	12
13	45850623	13	1
14	42838949	14	2
15	53620994	15	3

Another 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.

Year	Quarter	Time variable (this is created)	Revenues (in $M)
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
```

What is the Regression equation?

$$y = 10.0393 + 0.9440x$$

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	15.0	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	18.0	17.591	1.023

* T: Trend; S: Seasonal; I: Irregular

Quarterly Seasonality

Time	Average seasonality factor
Q1	$0.844 \left(= \frac{0.929+0.759}{2}\right)$
Q2	0.975
Q3	1.127
Q4	1.054

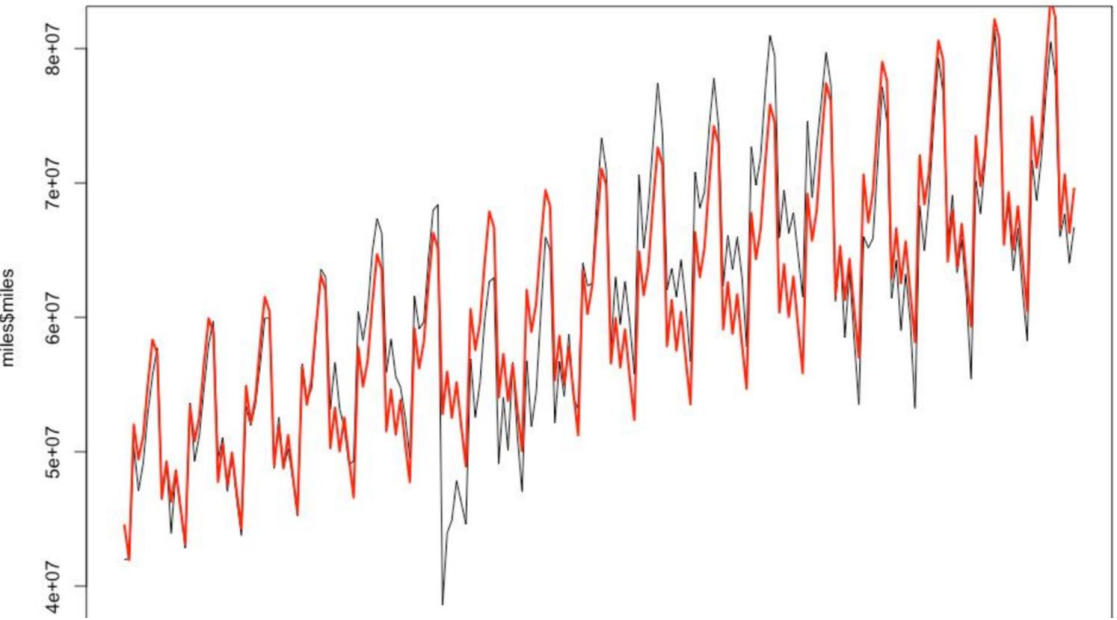
Time	Observed values TSI* (assuming no impact of cyclicality)	Predicted values (per the regression) T*	SI* = TSI/T
1	10.2	10.983	0.929
2	12.4	11.927	1.040
3	14.8	12.871	1.150
4	15.0	13.815	1.086
5	11.2	14.759	0.759
6	14.3	15.703	0.911
7	18.4	16.647	1.105
8	18.0	17.591	1.023

Computations

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

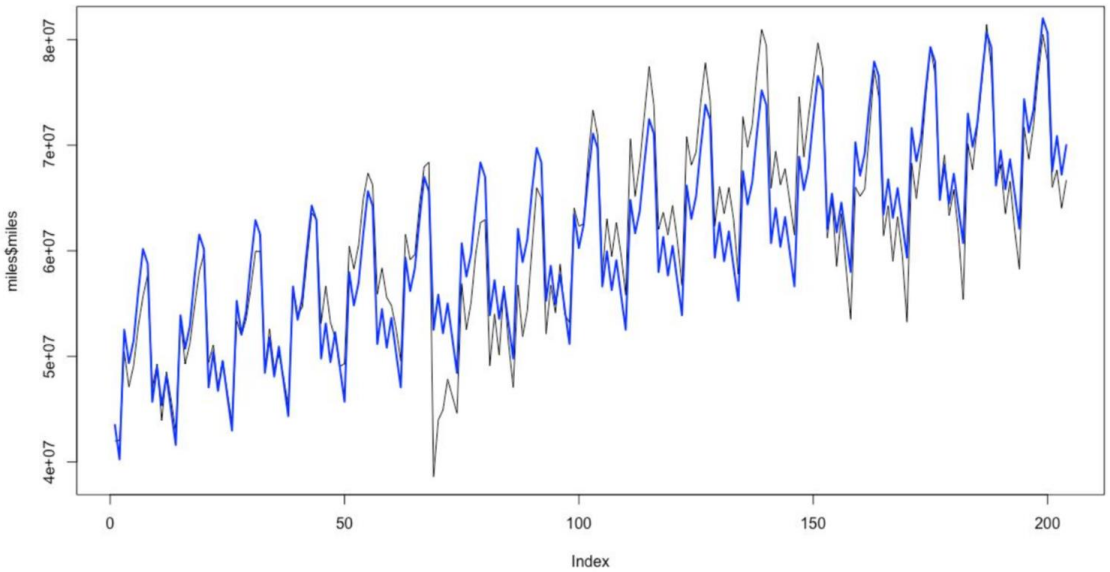
- Corrected for seasonality and randomness: 18.535 * 0.844 = 15.643

Seasonality: Multiplicative



	miles	time	seasonal	mae
1	41972194	1	1	0.849386
2	42054796	2	2	0.8491019
3	50443045	3	3	1.016129
4	47112397	4	4	0.946865
5	49118248	5	5	0.9849257
6	52880510	6	6	1.057953
7	55664750	7	7	1.111125
8	57723208	8	8	1.149602
9	47035464	9	9	0.9346292
10	49263120	10	10	0.9766855
11	43937074	11	11	0.8691307
12	48539606	12	12	0.9580177
13	45850623	13	1	0.9029174
14	42838949	14	2	0.8417232
15	53620994	15	3	1.051224

Seasonality: Additive



	miles	time	seasonal	mae
1	41972194	1	1	-7442550
2	42054796	2	2	-7473763
3	50443045	3	3	800670.5
4	47112397	4	4	-2643793
5	49118248	5	5	-751757.1
6	52880510	6	6	2896690
7	55664750	7	7	5567114
8	57723208	8	8	7511757
9	47035464	9	9	-3289802
10	49263120	10	10	-1175962
11	43937074	11	11	-6615823
12	48539606	12	12	-2127106
13	45850623	13	1	-4929905
14	42838949	14	2	-8055394
15	53620994	15	3	2612836

Issues with Regressing on Time

- If there is no trend or if seasonality and fluctuations are more important than trend, then the coefficients behave weirdly

Goodness of Fit

- MAE (Mean absolute error) $\frac{\sum|y_i-\hat{y}_i|}{n}$
- MSE (Mean square error) $\frac{\sum(y_i-\hat{y}_i)^2}{n}$
- RMSE (Root mean square error) $\sqrt{\frac{\sum(y_i-\hat{y}_i)^2}{n}}$
- MAPE (Mean absolute percent error) $\frac{1}{n}\left(\frac{\sum|y_i-\hat{y}_i|}{y_i}\right)*100$