DATA ANALYTICS

Unit 3: Ljung Box and Theil's coefficient

Jyothi R.
Department of Computer Science and
Engineering

DATA ANALYTICS

Differencing

- In Figure 1 to 9: The Google stock price was non-stationary in panel (a)
- But the daily changes were stationary in panel (b). This shows one way to make a non-stationary time series stationary — compute the differences between consecutive observations. This is known as **differencing**.
- Transformations such as logarithms can help to stabilise the variance of a time series.
- Differencing can help stabilise the mean of a time series by removing changes in the level of a time series, and therefore eliminating or reducing trend and seasonality.
- By looking at the time plot of the data, the ACF plot is also useful for identifying non-stationary time series.
- For a stationary time series, the ACF will drop to zero relatively quickly, while the ACF of non-stationary data decreases slowly.
- Also, for non-stationary data, the value of r_1 is often large and positive.

Differencing

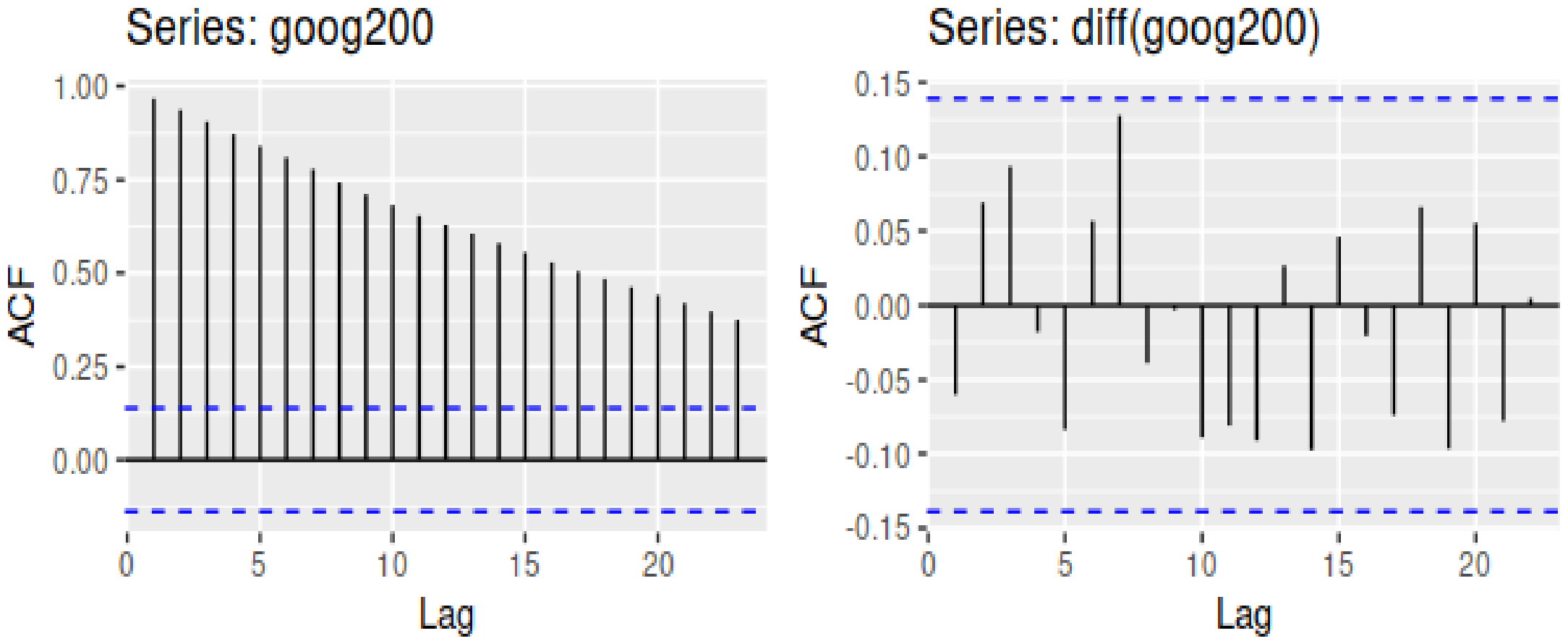

Figure 10:The ACF of the Google stock price (left) and of the daily changes in Google stock price (right).

Differencing

Figure 8.2:The ACF of the Google stock price (left) and of the daily changes in Google stock price (right).

- The ACF of the differenced Google stock price looks just like that of a white noise series.
- There are no autocorrelations lying outside the 95% limits, and
- The Ljung -Box Q∗statistic has a *p*-value of 0.355 (for h=10).
- This suggests that the *daily change* in the Google stock price is essentially a random amount which is uncorrelated with that of previous days.

Ljung-Box Test for Auto-Correlations

- Ljung−Box is a test of lack of fit of the forecasting model and checks whether the auto-correlations for the errors are different from zero.
- The null and alternative hypotheses are given by
- H_0: The model does not show lack of fit
- H_0: The model exhibits lack of fit

Ljung-Box Test for Auto-Correlations

- The Ljung−Box statistic (*Q*-Statistic) is given by (Ljung and Box, 1978)

$$Q(m) = n(n+2)\sum_{k=1}^{m}\frac{\rho_k^2}{n-k}$$

- where n is the number of observations in the time series,
- k is the number of lag,
- ρ_k is the auto-correlation of lag k, and
- m is the total number of lags.

Ljung-Box Test for Auto-Correlations

- *Q*-statistic is an approximate chi-square distribution with $m - p - q$ degrees of freedom where p and q are the AR and MA lags.
- The *Q*-statistic for ARIMA(1, 1, 1) is 10.216 (Table 1) and the corresponding *p*-value is 0.855 and thus we fail to reject the null hypothesis.
- Table 1: ARIMA (1, 1, 1) model summary for Omelette demand

Model	Model Fit Statistics			Ljung–Box *Q*(18)		
	R-Squared	RMSE	MAPE	Statistics	*Df*	Sig.
Omellette-Model_1	0.584	3.439	20.830	10.216	16	0.855

- *Q*(*m*) measures accumulated auto-correlation up to lag *m*.

POWER OF FORECASTING MODEL: THEIL'S COEFFICIENT

- The power of forecasting model is a comparison between
- Naive forecasting model and the model developed.
- In the Naive forecasting model, the forecasted value for the next period is same as the last period's actual value
- $F_{t+1} = Y_t$.

- Theil's coefficient (*U*-statistic) is given by (Theil, 1965)

POWER OF FORECASTING MODEL: THEIL'S COEFFICIENT

- For the data shown in Table 213.14 (demand for avionic system spares),
- The *U*-statistic calculations are shown in Table 3.
- TABLE 3:U-statistic calculation

Day	Y_t	ARMA (1,2) Forecast	$(Y_t - F_t)^2$	Naïve Forecast ($F_{t+1} = Y_t$)	$(Y_t - F_t)^2$
31	503	464.8107	1458.423	443	3600
32	688	378.5341	95769.15	503	34225
33	602	444.6372	24763.04	688	7396
34	629	685.8851	3235.909	602	729
35	823	743.5124	6318.281	629	37636
36	671	630.7183	1622.614	823	23104
37	487	649.3491	26357.22	671	33856
		Total	159524.6	Total	140546

POWER OF FORECASTING MODEL: THEIL'S COEFFICIENT

- Theil's coefficient (*U*-statistic) is given by (Theil, 1965)

$$U = \frac{\sum_{t=1}^{n} (Y_{t+1} - F_{t+1})^2}{\sum_{t=1}^{n} (Y_{t+1} - Y_t)^2}$$

- Theil's coefficient is the ratio of the mean squared error of the forecasting model to the MSE of the Naïve model.
- The value of $U < 1$ indicates that forecasting model is better than the Naive forecasting model.
- $U > 1$ indicates that the forecasting model is not better than Naive model.

POWER OF FORECASTING MODEL: THEIL'S COEFFICIENT

Table 2: Monthly demand (quantity of 200 gram packets) along with average price

Period	Month	Demand in Units	Average Price	Period	Demand in Units	Average Price
1	January	10500472	37	25	10658309	36
2	February	10123572	34	26	8677622	38
3	March	7372141	36	27	7330354	37
4	April	7764303	38	28	8115471	37
5	May	6904463	40	29	8481936	34
6	June	10068862	34	30	8778999	37
7	July	6436190	40	31	10145039	32
8	August	9898436	34	32	8497839	38
9	September	6803825	39	33	8792138	34
10	October	8333787	36	34	8485358	36
11	November	7541964	39	35	8575904	36
12	December	8540662	37	36	9885156	32
13	January	10229437	37	37	11023467	35
14	February	8453201	38	38	7942451	40
15	March	7997459	35	39	12492798	32
16	April	8557825	35	40	9756258	32
17	May	7818397	36	41	8992741	32
18	June	8944499	37	42	7397807	40
19	July	8904086	36	43	9710611	32
20	August	8463682	39	44	8328379	39
21	September	7723957	37	45	11873063	32
22	October	7731422	39	46	10642507	32
23	November	8441834	35	47	10635075	32
24	December	7485122	40	48	10578547	32

POWER OF FORECASTING MODEL: THEIL'S COEFFICIENT

- For the data shown earlier on the demand for avionic system spares, the *U*-statistic calculations are shown in the Table below:

Day	Y_t	ARMA (1,2) Forecast	$(Y_t - F_t)^2$	Naïve Forecast ($F_{t+1} = Y_t$)	$(Y_t - F_t)^2$
31	503	464.8107	1458.423	443	3600
32	688	378.5341	95769.15	503	34225
33	602	444.6372	24763.04	688	7396
34	629	685.8851	3235.909	602	729
35	823	743.5124	6318.281	629	37636
36	671	630.7183	1622.614	823	23104
37	487	649.3491	26357.22	671	33856
		Total	159524.6	Total	140546

The *U*-statistic value = 159524.6 / 140546 = 1.1350.

That is, ARMA(1, 2) model is not better thanNaive forecasting.

DATA ANALYTICS

The ‘X’ Factor (ARX, ARIMAX, etc.)

- ‘X’ = exogenous variables or explanatory variables

Important considerations:

- What other factors influence the forecast?
- How do we process this additional data to make it amenable for inclusion in our model?

DATA ANALYTICS

Practice Quiz

1. Seasonality in time-series data is caused due to

(a) Changes in macro-economic factors such as recession, unemployment, and so on

(b) Festivals and customs in a society

(c) Random events that occur over a period of time

(d) Changes in customer behaviour driven by new products and promotions

DATA ANALYTICS

Practice Quiz

2. In a simple exponential smoothing method, the low value of smoothing constant *a* is chosen when

(a) The data has high fluctuations around the trend line

(b) There is seasonality in the data

(c) The data is smooth with low fluctuations

(d) There are variations in the data due to cyclical component

DATA ANALYTICS

Practice Quiz

3. White noise is

(a) Uncorrelated errors with expected value 0.

(b) Uncorrelated errors that are constant and do not change with time.

(c) Uncorrelated errors that follow normal distribution with mean 0 and constant standard deviation

(d) Errors that follow normal distribution with constant mean and standard deviation

DATA ANALYTICS

Practice Quiz

4. A stationary process in a time series is a process for which

(a) Mean and variance are constant at different time points

(b) The time series follows normal distribution with zero mean and constant standard deviation

(c) The covariance of the time series depends only on the lag

(d) Mean and standard deviation are constant at different time points and the covariance depends only on the lag between the values and is constant for a given lag

References

Text Book:

"Business Analytics, The Science of Data-Driven Making", U. Dinesh Kumar, Wiley 2017 Ch. 13.14.5 and 13.15

Image Courtesy

https://www.abs.gov.au/websitedbs/D3310114.nsf/home/Time+Series+Analysis:+The+Basics

THANK YOU

Jyothi R

Assistant Professor, Department of Computer Science

jyothir@pes.edu