

# 시계열 회귀분석(Time Series Regression - 3)

강의 : 김성범 교수님

정리 : 김선중

November 25, 2022

이것은 Time Series Regression - Part 3 강의에 대한 노트이다.

## Contents

1 Binary Variable Models	2
2 Trigonometric Models	4
3 Growth Curve Models	5
4 First-Order Autoregressive Process	6

아래 식에서 보듯이, 많은 경우에 시계열은 추세(trend)와 계절성(seasonality), 그리고 오차로 이루어져있다. 이번 강의에서는 seaonsal variation을 모델링하는 네 가지 모델에 대해서 다루는 것 같다. 그것들은

- binary variable models
- trigonometric models
- growth curve models
- first-order autoregressive process

이다. 이 중에서 앞의 세 개는 고전적인 방법이고 마지막 방법이 앞으로 계속 사용할 방법인 것 같다.

### Modeling Seasonal Variations

---

$$y_t = TR_t + SN_t + \varepsilon_t$$

$y_t$  = the value of the time series in period t

$TR_t$  = the trend in time period t

$SN_t$  = the seasonal factor in time period t

$\varepsilon_t$  = the error term(irregular factor) in time period t

# 1 Binary Variable Models

## Modeling Seasonal Variations Using Binary Variables

The seasonal factor expressed using binary variables is

$$SN_t = \beta_{s1}x_{s1,t} + \beta_{s2}x_{s2,t} + \cdots + \beta_{s(L-1)}x_{s(L-1),t}$$

where  $x_{s1,t}, x_{s2,t}, \dots, x_{s(L-1),t}$  are binary variables that are defined as follows:

$$x_{s1,t} = \begin{cases} 1 & \text{if time period } t \text{ is season 1} \\ 0 & \text{otherwise} \end{cases}$$

$$x_{s2,t} = \begin{cases} 1 & \text{if time period } t \text{ is season 2} \\ 0 & \text{otherwise} \end{cases}$$

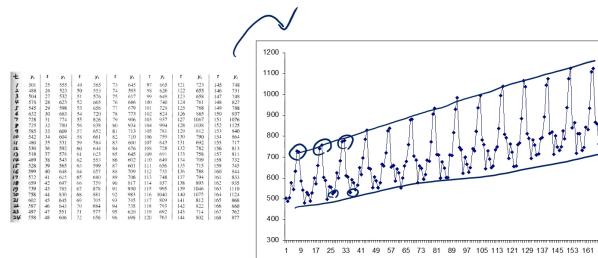
⋮

$$x_{s(L-1),t} = \begin{cases} 1 & \text{if time period } t \text{ is season } (L-1) \\ 0 & \text{otherwise} \end{cases}$$

Seoung Bum Kim - Copyright © All rights reserved.

## Seasonal Variations

- A linear trend and increasing seasonal variations



Seoung Bum Kim - Copyright © All rights reserved.

## Example - Binary Variable Models

$$y_t = TR_t + SN_t + \varepsilon_t$$

$$y_t = \beta_0 + \beta_1 t + \beta_2 M_1 + \beta_3 M_2 + \beta_4 M_3 + \cdots + \beta_{12} M_{12} + \varepsilon_t$$

$TR_t$

$SN_t$

$$SN_t = \beta_{s1}x_{s1,t} + \beta_{s2}x_{s2,t} + \cdots + \beta_{s(L-1)}x_{s(L-1),t}$$

$M_1$

$M_2$

$M_{12}$

Seoung Bum Kim - Copyright © All rights reserved.

## Example - Binary Variable Models

$$y_t^* = TR_t + SN_t + \varepsilon_t$$

$$= \beta_0 + \beta_1 t + \beta_2 M_1 + \beta_3 M_2 + \cdots + \beta_{12} M_{12} + \varepsilon_t$$

where  $y_t^* = \ln y_t$  and  $M_1, M_2, \dots, M_{12}$  are seasonal binary variables.

$$M_1 = \begin{cases} 1 & \text{if period } t \text{ is January} \\ 0 & \text{otherwise} \end{cases}$$

	$X_1$ ( $M_1$ )	$X_2$ ( $M_2$ )	$X_3$ ( $M_3$ )	$X_4$ ( $M_4$ )	...	$X_{12}$ ( $M_{12}$ )	$Y$
1	1	0	0	0	...	0	50.1
2	2	0	1	0	...	0	48.8
3	3	0	0	1	...	0	50.4
4	4	0	0	0	...	0	87.7

Seoung Bum Kim - Copyright © All rights reserved.

## Example - Binary Variable Models

$$y_t^* = TR_t + SN_t + \varepsilon_t$$

$$= \beta_0 + \beta_1 t + \beta_2 M_1 + \beta_3 M_2 + \cdots + \beta_{12} M_{12} + \varepsilon_t$$

where  $y_t^* = \ln y_t$  and  $M_1, M_2, \dots, M_{12}$  are seasonal binary variables.

$M_{12}$

- Here we have arbitrarily set the seasonal parameter for season 12 (the last month, December) equal to zero. Thus, the other 11 seasonal parameters are defined with respect to December.
- For example,  $\beta_2$  is the difference between the level of time series in January and the level of the time series in December.
- A positive  $\beta_2$  implies that the level of the time series in January can be expected to be greater than the level in December.

Seoung Bum Kim - Copyright © All rights reserved.

## Example - Binary Variable Models

Variable	DF	Parameter Estimate	Standard Error	t Value	Pr >  t
Intercept	1	6.28756	0.00643	978.26	<.0001
Time	1	0.000373	0.000373	80.65	<.0001
1	1	-0.11671	0.00802	-14.39	<.0001
2	1	-0.11208	0.00801	-13.98	<.0001
3	1	-0.08446	0.00801	-10.54	<.0001
4	1	0.03893	0.00801	4.97	<.0001
5	1	0.14691	0.00801	18.54	<.0001
7	1	0.28902	0.00801	36.09	<.0001
8	1	0.31119	0.00801	38.86	<.0001
9	1	0.05201	0.00801	6.99	<.0001
10	1	-0.03954	0.00801	-4.94	<.0001
11	1	-0.11222	0.00801	-14.01	<.0001

$$\hat{y}_{169}^* = b_0 + b_1(169) + b_2(1) + \times$$

$$= 6.28756 + .00273(169) + (-.04161)(1)$$

$$= 6.7065$$

$$\hat{y}_{169} = e^{6.7065} = 817.70$$

Seoung Bum Kim - Copyright © All rights reserved.

seasonal variation을 다루기 위한 여러 모델들 중 첫번째로 다루는 모델은 binary variable model이다. 예시로 주어지고

있는 데이터셋은 한 호텔의 투숙된 객실 수에 대한 데이터이다. 시간  $t$ 의 단위는 ‘월’로 주어져있고, 총 7년의 데이터가 있으므로  $t \in \{1, 2, \dots, 168\}$ 이다. 종속변수는  $y$ 이고 독립변수는  $x$ 이지만,  $x_t = t$ 라는 점에서 독립변수를 그냥  $t$ 라고 봐도 될 것 같다. 다시 말해서, 몇 번째 달( $t$ )에 몇 개의 객실들( $y$ )이 투숙되어 있는지 하는 단변수 회귀 (univariate regression) 문제이다.

이전 강의의 예에서는 추세(trend,  $TR_t$ )만을 예측모델로 잡았었다. 하지만, 세 번째 슬라이드에서 보듯 계절성이 뚜렷이 드러나고 있으므로, 예측모델  $f_\beta$ 을 설정할 때 추세 말고도 계절성(seasonal variation,  $SN_t$ )도 고려할 것이다. 즉

$$f_\beta(t) = TR_t + SN_t \quad (1)$$

이고

$$y_t = f_\beta(t) + \epsilon_t$$

이다. 추세는 일차함수(affine function)로 표현할 것이어서

$$TR_t = \beta_0 + \beta_1 t$$

이고, 계절성은 각각의 계절에 대하여 상수회귀(no trend, constant regression)로 둔다. 식으로 표현하면

$$SN_t = \begin{cases} \beta_2 & (t = 12n + 1) \\ \beta_3 & (t = 12n + 2) \\ \vdots & (\text{단, } n = 0, 1, 2, \dots, 6) \\ \beta_{11} & (t = 12n + 10) \\ 0 & (t = 12n + 11) \end{cases}$$

이다. 이것을 표현하기 위해서 강의에서는  $M_1, M_2, \dots, M_{11}$ 을 사용하고 있는데, 이건 각 월에 대한 characteristic function (indicator function)으로 이해하면 될 것 같다. 여하튼, 식 (1)을 다시 정리하면

$$f_\beta(t) = \begin{cases} \beta_0 + \beta_2 + \beta_1 t & (t = 12n + 1) \\ \beta_0 + \beta_3 + \beta_1 t & (t = 12n + 2) \\ \vdots & (\text{단, } n = 0, 1, 2, \dots, 6) \\ \beta_0 + \beta_{11} + \beta_1 t & (t = 12n + 10) \\ \beta_0 + \beta_1 t & (t = 12n + 11) \end{cases} \quad (2)$$

이 된다. 각 월(1월 ~ 11월)에 대한 정보를 담고 있는 항은  $\beta_2, \dots, \beta_{11}$ 이다. 그리고 12월에 대한 정보를 담고 있는 항은 없다. 하지만 문제가 되지 않는다.

다시 말해서, 1월부터 11월까지의 월들에 각각 어떤 기본값을 부여할 지에 대해서는 매개변수  $\beta_2, \dots, \beta_{11}$ 로 조정하면 된다. 하지만,  $y$ 절편에 해당하는  $\beta_0$ 를 이미 설정해놓았으므로, 12월에는  $\beta_0$ 라는 기본값을 부여받게 되는 것이다. 12월에 대한 정보는  $\beta_0$ 를 통해 조정될 수 있는 것이다. 새로운 매개변수  $\beta_{12}$ 를 도입할 수도 있지만, 굳이 그렇게 하는 게 의미가 없는 것이다.

그런데, 어차피 그렇게 할거면, trend를 설정할 때,  $y$ 절편이 없는 일차함수로 잡은 다음  $M_1, M_2, \dots, M_{12}$ 를 설정해도 같은 의미가 될 것이다. 그렇게 하는 편이 보기에 깔끔해보인다. 하지만, 지금과 같은 방법을 취하는 것이 아마도 통계 방면에서의 관습인 게 아닐까 싶기도 하다.

이렇게 parametric model  $f_\beta$ 를 설정했다. 그 다음으로 하는 것은 기존의 회귀분석(ordinary regression analysis)을 진행

하는 것이다. 강의에서는 45분 32초쯤에 ‘일반적인 최소제곱법(LSE, least square estimation ; OLS ordinary least square, ordinary least squares)’을 사용하는 것이다. 표에서 관측치가 주어져있었다.

$$\begin{aligned} \text{관측치} &= \{(t, y_t) : t = 1, 2, \dots, 168\} \\ &= \{(1, 501), (2, 488), \dots, (168, 877)\} \end{aligned}$$

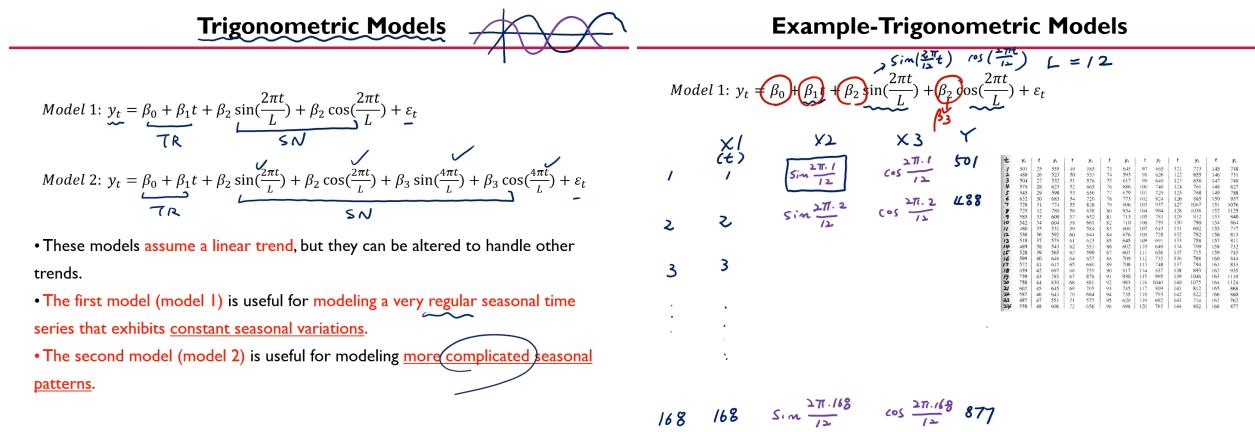
이걸 가지고 MSE를 계산하면

$$\text{MSE} = \frac{1}{168} \sum_{t=1}^{168} (y_t - f_\beta(t))^2$$

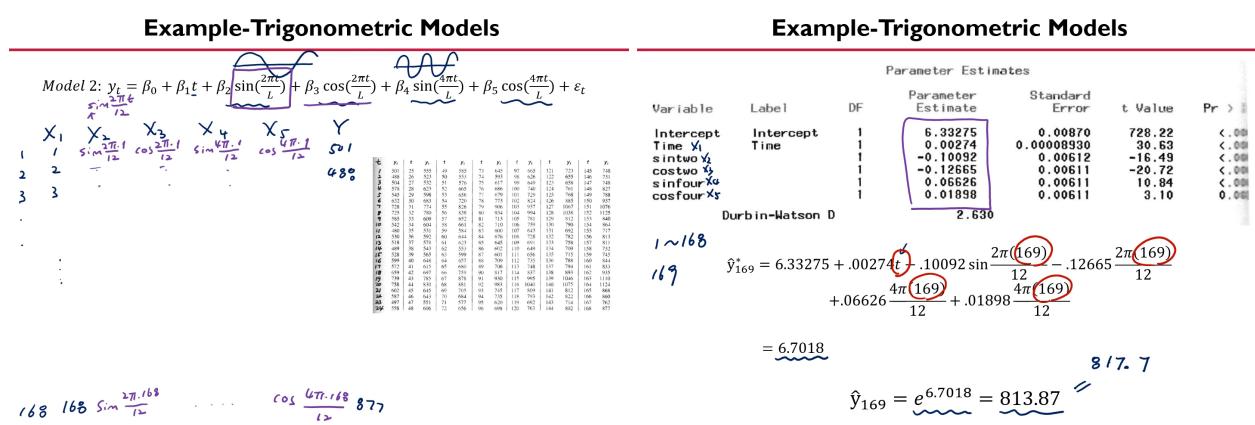
이 된다.  $MSE$ 를  $\beta_0, \beta_1, \dots, \beta_{11}$ 로 편미분한 것을 0으로 두면 미지수가 12개이고 식이 12개인 연립방정식이 나오는데, 그 연립방정식을 풀어 근을  $\beta_0 = \hat{\beta}_0, \beta_1 = \hat{\beta}_1, \dots, \beta_{11} = \hat{\beta}_{11}$  들을 가지고 최적의 함수  $\hat{f}$ 를 찾을 수 있다.

이 계산들은 보통은 컴퓨터를 통해, 몇개의 명령어를 입력하여 계산하는 것 같고, 마지막 캡쳐의 표에 이  $\hat{\beta}_i$ 들의 값이 적혀있는 것으로 보인다.

## 2 Trigonometric Models



Seoung Bum Kim - Copyright © All rights reserved.



Seoung Bum Kim - Copyright © All rights reserved.

### 3 Growth Curve Models

#### Growth Curve Models

- Useful for the models, not linear in the parameters.

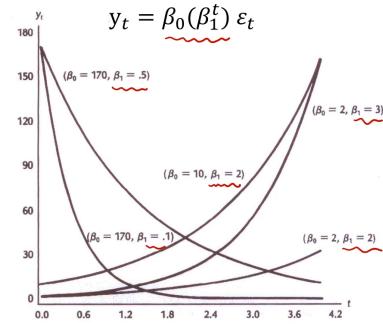
$$\log y_t = \beta_0 + \beta_1 t + \varepsilon_t$$

- Transform a nonlinear model to one that is linear in parameter.

$$\log(y_t) = \log(\beta_0) + t \log(\beta_1) + \log(\varepsilon_t)$$

$$\begin{aligned} \log(AB) &= \log(A) + \log(B) \\ \log(A^r) &= r \log(A) \end{aligned}$$

#### Growth Curve Models



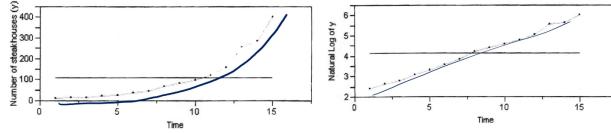
Seoung Burn Kim - Copyright © All rights reserved.

Seoung Burn Kim - Copyright © All rights reserved.

#### Growth Curve Models - Example

- Number of branches of Western Steakhouses for 15 years.

Year (t)	$y_t$	$\ln y_t$	Year (t)	$y_t$	$\ln y_t$
1	11	2.398	9	82	4.407
2	14	2.639	10	99	4.595
3	16	2.773	11	119	4.779
4	22	3.091	12	156	5.050
5	28	3.332	13	257	5.549
6	36	3.584	14	284	5.649
7	46	3.829	15	403	5.999
8	67	4.205			



Seoung Burn Kim - Copyright © All rights reserved.

#### Growth Curve Models - Example

Ln y = 2.07 + 0.257Year					
Predictor	Coeff	SE Coef	T	P	
Constant	2.07012	0.04103	50.45	0.000	
Year	0.25688	0.00451	56.92	0.000	
S = 0.07552	R-Sq = 99.6%	R-Sq(adj) = 99.6%			

New Obs	Fit	SE Fit	95.0% CI	95.0% PI
1	6.1802	0.0410	(6.0916, 6.2684)	(5.9945, 6.3659)

The point prediction of  $\ln y_{16}$ , where  $y_{16}$  is the number of steakhouses that will be in operation in period 16, is

$$\ln \hat{y}_{16} = \hat{\beta}_0 + \hat{\beta}_1 t = 2.07012 + 0.25688(16) = 6.1802$$

Thus a point prediction of  $y_{16}$  is

$$\hat{y}_{16} = e^{6.1802} = 483.09$$

The 95% prediction interval for  $\ln y_{16}$  is

$$[5.9945, 6.3659]$$

And thus a 95% prediction interval for  $y_{16}$  is

$$[e^{5.9945}, e^{6.3659}] = [401.22, 581.67]$$

Seoung Burn Kim - Copyright © All rights reserved.

## 4 First-Order Autoregressive Process

### Time Series Regression with Autocorrelation

- How to model a time series that possesses a first-order autocorrelated error structure.
- If we ignore autocorrelated error terms, we will pay a penalty in terms of wider prediction interval.
- By taking autocorrelation into account, we can obtain more precise prediction intervals.
- We can check this using the residual plots or the Durbin-Watson test.

$$Y = \beta_0 + \beta_1 X + \varepsilon_t$$

Seoung Bum Kim - Copyright © All rights reserved.

Example							
year	tfr	partic	degrees	fconvict	ftheft	mconvict	mtheft
1 1931	3200	234	12.4	77.1	NA	778.7	NA
2 1932	3084	234	12.9	92.9	NA	745.7	NA
3 1933	2864	235	13.9	98.3	NA	768.3	NA
4 1934	2803	237	13.6	88.1	NA	733.6	NA
5 1935	2755	238	13.2	79.4	20.4	765.7	247.1
6 1936	2696	240	13.2	91.0	22.1	816.5	254.9
...							
37 1967	2586	339	80.4	115.2	70.6	781.1	272.0
38 1968	2441	338	90.4	122.9	73.0	849.7	274.7

- year, 1931–1968.
- tfr, the total fertility rate, births per 1000 women.
- partic, women's labor-force participation rate, per 1000.
- degrees, women's post-secondary degree rate, per 10,000.
- fconvict, women's indicable-offense conviction rate, per 100,000.
- ftheft, women's theft conviction rate, per 100,000.
- mconvict, men's indicable-offense conviction rate, per 100,000.
- mtheft, theft conviction rate, per 100,000.

Seoung Bum Kim - Copyright © All rights reserved.

### First-Order Autoregressive Process

$$\varepsilon_t = \phi_1 \varepsilon_{t-1} + a_t$$

$a_t$  is assumed to be an error term with mean 0 that satisfies the constant variance, independence, and normality assumption.

- $\phi_1$  is the correlation coefficient between error terms separated by one time period.
- If  $\phi_1 > 0$ , this indicates that the error terms are positively autocorrelated. That means a positive error term  $\varepsilon_t$  tends to produce another positive error term  $\varepsilon_{t-1}$ .
- If  $\phi_1 < 0$ , the error terms are negatively autocorrelated. In this case, a positive error term  $\varepsilon_{t-1}$  tends to produce a negative error term  $\varepsilon_t$  and vice versa.

Seoung Bum Kim - Copyright © All rights reserved.

### Example

#### • Ordinary multiple regression result

```
> mod.ols <- lm(fconvict ~ tfr + partic + degrees + mconvict, data=Hartnagel)
> summary(mod.ols)
```

Coefficients:

	Estimate	Std. Error	t value	Pr(> t )
(Intercept)	127.64000	59.95704	2.13	0.041
tfr	-0.04657	0.00803	-5.80	1.8e-06
partic	0.25342	0.11513	2.20	0.035
degrees	-0.21205	0.21145	-1.00	0.323
mconvict	0.05910	0.04515	1.31	0.200

Residual standard error: 19.2 on 33 degrees of freedom  
Multiple R-Squared: 0.695, Adjusted R-squared: 0.658  
F-statistic: 18.8 on 4 and 33 DF, p-value: 3.91e-008

Seoung Bum Kim - Copyright © All rights reserved.

### Example

#### • Time series regression with second-order autoregressive process

Correlation Structure: ARMA(2,0)  
Formula: ~1  
Parameter estimate(s):  
Phi1 Phi2  
1.0683 -0.5507

Coefficients:

	Value	Std. Error	t-value	p-value
(Intercept)	83.34	59.47	1.401	0.1704
tfr	-0.04	0.01	-4.309	0.0001
partic	0.29	0.11	2.568	0.0150
degrees	-0.21	0.21	-1.016	0.3171
mconvict	0.08	0.04	2.162	0.0380

Seoung Bum Kim - Copyright © All rights reserved.

#### ① Ordinary regression analysis

	Estimate	Std. Error	t value	Pr(> t )
(Intercept)	127.64000	59.95704	2.13	0.041
tfr	-0.04657	0.00803	-5.80	1.8e-06
partic	0.25342	0.11513	2.20	0.035
degrees	-0.21205	0.21145	-1.00	0.323
mconvict	0.05910	0.04515	1.31	0.200

#### ② Time series regression analysis

	Value	Std. Error	t-value	p-value
(Intercept)	83.34	59.47	1.401	0.1704
tfr	-0.04	0.01	-4.309	0.0001
partic	0.29	0.11	2.568	0.0150
degrees	-0.21	0.21	-1.016	0.3171
mconvict	0.08	0.04	2.162	0.0380

- In our example, ML estimate of the regression parameters under the AR(2) error-correlation model are not terribly different from the OLS estimates.
- The coefficient for mconvict is statistically significant.

Seoung Bum Kim - Copyright © All rights reserved.