시계열 회귀분석(Time Series Regression - 3)

강의 : 김성범 교수님

정리 : 김선중

November 25, 2022

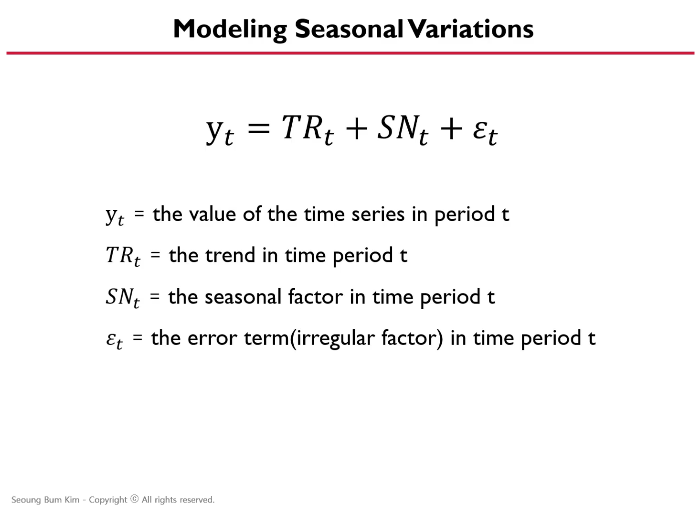
이것은 Time Series Regression - Part 3 강의에 대한 노트이다.

Contents

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.



이번 정리부터는 유튜브 캡쳐를 나열한 후 정리하려고 하고, 너무 많은 내용을 적기보다는 강의 내용 자체에만 충실하게 적어보려고 한다.

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

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

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

$$\vdots$$

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

L

Example - Binary Variable Models

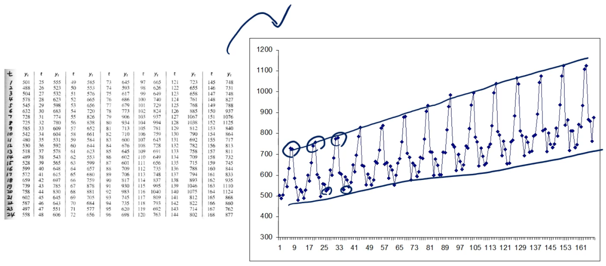
- Number of monthly hotel rooms occupied for 14 years.

24
12
168

t	y_t	t	y_t	t	y_t	t	y_t	t	y_t	t	y_t	t	y_t
1	501	25	555	49	585	73	645	97	665	121	723	145	748
2	488	26	523	50	553	74	593	98	626	122	655	146	731
3	504	27	532	51	576	75	617	99	649	123	658	147	748
4	578	28	623	52	665	76	686	100	740	124	761	148	827
5	545	29	598	53	656	77	679	101	729	125	768	149	788
6	632	30	683	54	720	78	773	102	824	126	885	150	937
7	728	31	774	55	826	79	906	103	937	127	1067	151	1076
8	725	32	780	56	838	80	934	104	994	128	1038	152	1125
9	585	33	609	57	652	81	713	105	781	129	812	153	840
10	542	34	604	58	661	82	710	106	759	130	790	154	864
11	480	35	531	59	584	83	600	107	643	131	692	155	717
12	530	36	592	60	644	84	676	108	728	132	782	156	813
13	518	37	578	61	623	85	645	109	691	133	758	157	811
14	489	38	543	62	553	86	602	110	649	134	709	158	732
15	528	39	565	63	599	87	601	111	656	135	715	159	745
16	599	40	648	64	657	88	709	112	735	136	788	160	844
17	572	41	615	65	680	89	706	113	748	137	794	161	833
18	659	42	697	66	759	90	817	114	837	138	893	162	935
19	739	43	785	67	878	91	930	115	995	139	1046	163	1110
20	758	44	830	68	881	92	983	116	1040	140	1075	164	1124
21	602	45	645	69	705	93	745	117	809	141	812	165	868
22	587	46	643	70	684	94	735	118	793	142	822	166	860
23	497	47	551	71	577	95	620	119	692	143	714	167	762
24	558	48	606	72	656	96	698	120	763	144	802	168	877

Seasonal Variations

- A linear trend and increasing seasonal variations



Example - Binary Variable Models

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

$$= \underbrace{\beta_0 + \beta_1 t}_{} + \underbrace{\beta_2 M_1 + \beta_3 M_2 + \beta_3 M_3 + \cdots + \beta_{12} M_{11}}_{} + \varepsilon_t$$

X_t 7 X_{L-1}

where $y_t^* = \ln y_t$ and
$M_1, M_2, \ldots, M_{11}$ are seasonal binary variables.

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

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

M_2

12

	X_1 (t)	X_2 (M_1)	X_3 (M_2)	X_4 (M_3)	⋯	X_{12} (M_{11})	Y
1	1	1	0	0	⋯	0	501
2	2	0	1	0	⋯	0	488
3	3	0	0	1	⋯	0	504
4	4						
⋮	⋮						
168	168	0	0	0	⋯	0	877

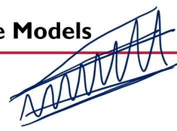
Example - Binary Variable Models

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

$$y_t = \underbrace{\beta_0 + \beta_1 t}_{TR_t} + \underbrace{\beta_2 M_1 + \beta_2 M_1 + \beta_3 M_2 + \cdots + \beta_{12} M_{11}}_{SN_t} + \varepsilon_t$$

$$SN_t = \beta_{s1}\underbrace{x_{s1,t}}_{M_1} + \beta_{s2}\underbrace{x_{s2,t}}_{M_2} + \cdots + \beta_{s(L-1)}\underbrace{x_{s(L-1),t}}_{M_{11}}$$

Example - Binary Variable Models



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

$$= \beta_0 + \beta_1 t + \beta_2 M_1 + \beta_2 M_1 + \beta_3 M_2 + \cdots + \beta_{12} M_{11} + \varepsilon_t$$

where $y_t^* = \ln y_t$ and $M_1, M_2, \ldots, M_{11}$ 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, β_2 is the difference between the level of time series in January and the level of the time series in December. + −
- A positive β_2 implies that the level of the time series in January can be expected to be greater than the level in December.

Example - Binary Variable Models

Parameter Estimates

Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|
Intercept	1	6.28756	0.00643	978.26	<.0001
Time	1	0.00273	0.00003379	80.65	<.0001
1	1	-0.04161	0.00802	-5.19	<.0001
2	1	-0.11208	0.00801	-13.98	<.0001
3	1	-0.08446	0.00801	-10.54	<.0001
4	1	0.03983	0.00801	4.97	<.0001
5	1	0.02040	0.00801	2.55	0.0119
6	1	0.14691	0.00801	18.34	<.0001
7	1	0.28902	0.00801	36.09	<.0001
8	1	0.31119	0.00801	38.86	<.0001
9	1	0.05599	0.00801	6.99	<.0001
10	1	0.03954	0.00801	4.94	<.0001
11	1	-0.11222	0.00801	-14.01	<.0001

Durbin-Watson D 1.190

$\hat{\beta}_0$

$$\hat{y}_{169}^* = b_0 + b_1(169) + b_2(1) + \cancel{0}$$

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

$$= 6.7065$$

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

	t X_1	M1 X_2	⋯	M11 X_{12}	Y
1					
2					
⋮					
168					
169	169	1	0 ⋯	0	?

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 t의 단위는 '월'로 주어져있고, 총 7년의 데이터가 있으므로 $t \in \{1, 2, \cdots, 168\}$이다. 종속변수는 y이고 독립변수는 x이지만, $x_t = t$라는 점에서 독립변수를 그냥 t라고 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_β을 설정할 때 추세 말고도 계절성(seasonal variation, SN_t)도 고려할 것이다. 즉

$$f_\beta(t) = TR_t + SN_t \tag{1}$$

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

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

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

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

이 된다. 각 월(1월 $\sim$ 11월)에 대한 정보를 담고 있는 항은 β_2, $\cdots$, β_{12}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 β_2, $\cdots$, β_{11}로 조정하면 된다. 하지만, y절편에 해당하는 β_0를 이미 설정해놓았으므로, 12월에는 β_0라는 기본값을 부여받게 되는 것이다. 12월에 대한 정보는 β_0를 통해 조정될 수 있는 것이다. 새로운 매개변수 β_{12}를 도입할 수도 있지만, 굳이 그렇게 하는 게 의미가 없는 것이다.

그런데, 어차피 그렇게 할거면, trend를 설정할 때, y절편이 없는 일차함수로 잡은 다음 M_1, M_2, $\cdots$, M_{12}를 설정해도

같은 의미가 될 것이다. 그렇게 하는 편이 보기에 깔끔해보인다. 그러니까

$$SN_t = \begin{cases} \beta_1 & (t = 12n+1) \\ \beta_2 & (t = 12n+2) \\ & \vdots \\ \beta_{12} & (t = 12n+12) \end{cases} \qquad (\text{단, } n = 0, 1, 2, \cdots, 6)$$

로 두고

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

로 두는 것이 더 깔끔하면서도, 기존 방법과 완전히 같은 방법이지 않을까 한다. 하지만, 아마도 기존 방법이 통계 방면에서의 관습인 게 아닐까 싶기도 하다.

이렇게 parametric model f_β를 설정했다. 그 다음으로 하는 것은 기존의 회귀분석(ordinary regression analysis)을 진행하는 것이다. 강의에서는 45분 32초쯤에 '일반적인 최소제곱법(LSE, least square estimation ; OLS ordinary least square, ordinary least squares)'을 사용하는 것이다. 표에서 관측치가 주어져있었다.

$$\begin{aligned} \text{관측치} &= \{(t, y_t) : t = 1, 2, \cdots, 168\} \\ &= \{(1, 501), (2, 488), \cdots, (168, 877)\} \end{aligned} \tag{3}$$

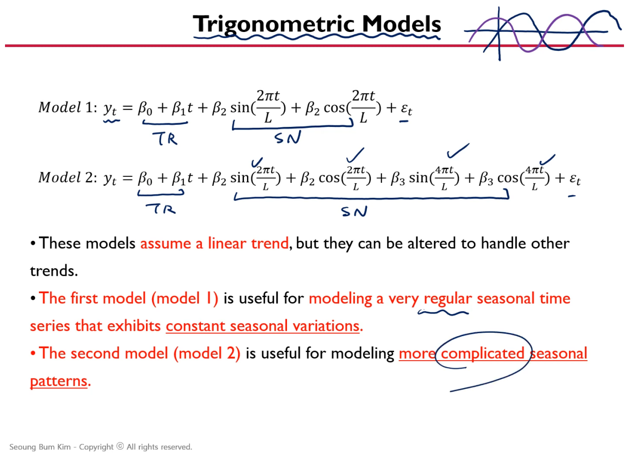
이걸 가지고 MSE를 계산하면

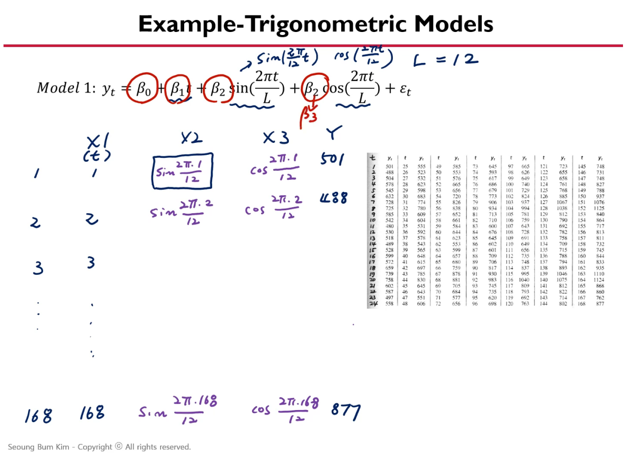
$$\text{MSE} = \frac{1}{168} \sum_{t=1}^{168} (y_t - f_\beta(t))^2 \tag{4}$$

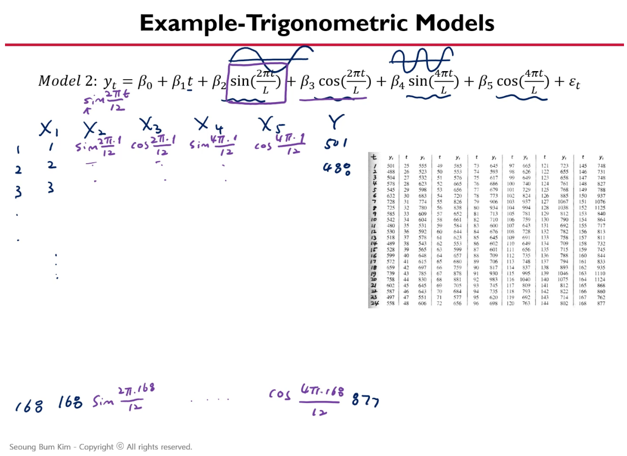
이 된다. MSE를 β_0, β_1, $\cdots$, β_{11}로 편미분한 것을 0으로 두면 미지수가 12개이고 식이 12개인 연립방정식이 나오는데, 그 연립방정식을 풀어 근을 $\beta_0 = \hat{\beta}_0$, $\beta_1 = \hat{\beta}_1$, $\cdots$, $\beta_{11} = \hat{\beta}_{11}$ 들을 가지고 최적의 함수 $\hat{f}$를 찾을 수 있다.

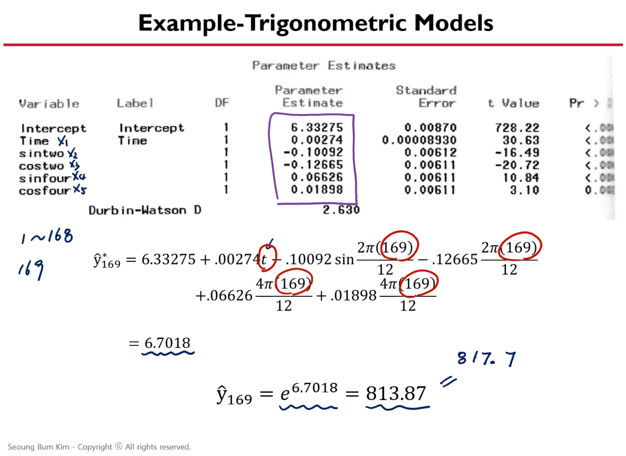
이 계산들은 보통은 컴퓨터를 통해, 몇개의 명령어를 입력하여 계산하는 것 같고, 마지막 캡쳐의 표에 이 $\hat{\beta}_i$들의 값이 적혀있는 것으로 보인다.

2 Trigonometric Models









$y_t = TR_t + SN_t + \epsilon_t$에서 SN_t는 '계절성'을 의미했고, '이것은 일정한 주기를 가지는 패턴'정도로 이해할 수 있었다. 그러니까 SN_t는 일정한 주기를 가지는 함수, 즉 주기함수이다. 이러한 주기함수를 사인함수와 코사인함수의 일차결합인

$$\beta_2 \sin\left(\frac{2\pi t}{L}\right) + \beta_3 \sin\left(\frac{2\pi t}{L}\right) \tag{5}$$

로 표현하는 것은 적당해보인다. 식 (5)은 적당한 실수 $A = \sqrt{{\beta_2}^2 + {\beta_3}^2} > 0$, B에 대하여

$$A \cos\left(\frac{2\pi t}{L} - B\right) \tag{6}$$

로 바꿀 수 있으니, 이것은 일반적인 사인함수(코사인함수)를 나타낸다. 그런데 Model 2에서는 주기가 절반인 사인함수와 코사인함수의 일차결합을 더 더해서 표현하고 있다. 이건 아무래도 푸리에 이론과 관련이 있을 것 같다.

푸리에 이론에 따르면, 주기가 L인 임의의 주기함수 $g : \mathbb{R} \to \mathbb{R}$는 사인함수와 코사인함수의 급수로 표현할 수 있다.

$$g(t) = \frac{a_0}{2} + \sum_{k=1}^{\infty} a_k \sin\left(\frac{2k\pi t}{L}\right) + b_k \cos\left(\frac{2k\pi t}{L}\right)$$

그러니까, 함수 g_n을 $(n = 1, 2, \cdots)$

$$g_k(t) = a_k \sin\left(\frac{2n\pi t}{L}\right) + b_k \cos\left(\frac{2k\pi t}{L}\right)$$

로 두면 $\left(g_0(t) = \frac{a_0}{2}\right)$, Model 1에서 SN_t는 $g_0(t) + g_1(t)$를 의미하고, Model 2에서 SN_t는 $g_0(t) + g_1(t) + g_2(t)$를 의미한다. $\sum g_k(t)$에서 n을 더 늘려가면, $\sum g_k(t)$는 점점 더 SN_t을 잘 표현할 수 있을 것이다.

다시 돌아와서 말하자면, model 1에서는

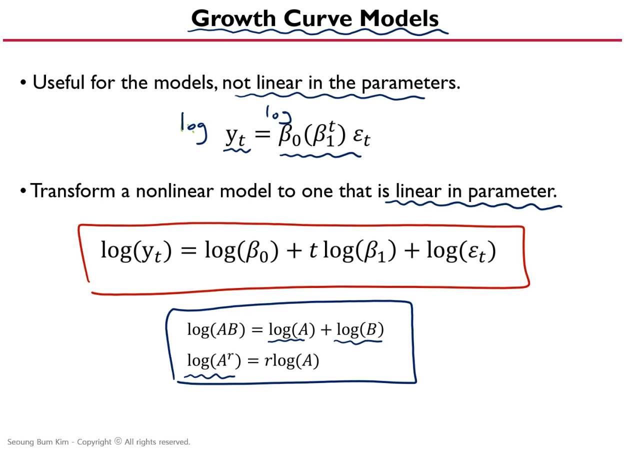
$$f_\beta(t) = \beta_0 + \beta_1 t + \beta_2 \sin\left(\frac{2\pi t}{L}\right) + \beta_3 \cos\left(\frac{2\pi t}{L}\right)$$

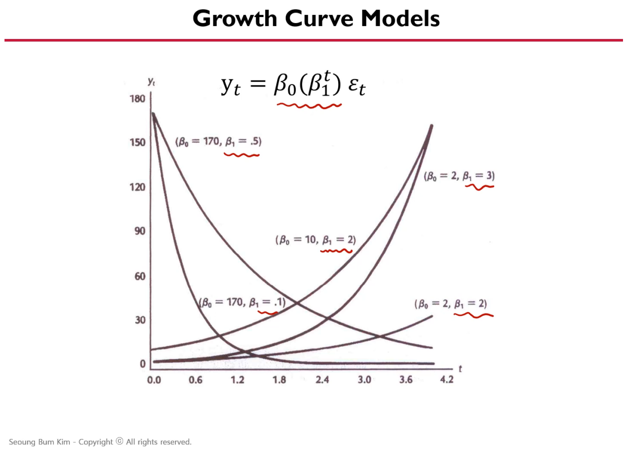
로 parametric model을 잡는 것이고, model 2에서는

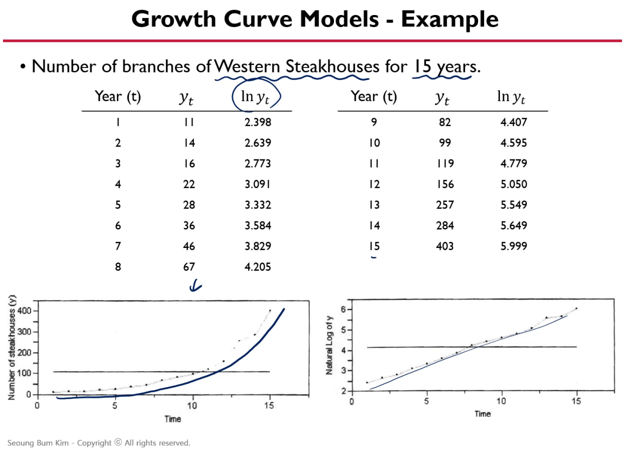
$$f_\beta(t) = \beta_0 + \beta_1 t + \beta_2 \sin\left(\frac{2\pi t}{L}\right) + \beta_3 \cos\left(\frac{2\pi t}{L}\right) + \beta_4 \sin\left(\frac{4\pi t}{L}\right) + \beta_5 \cos\left(\frac{4\pi t}{L}\right)$$

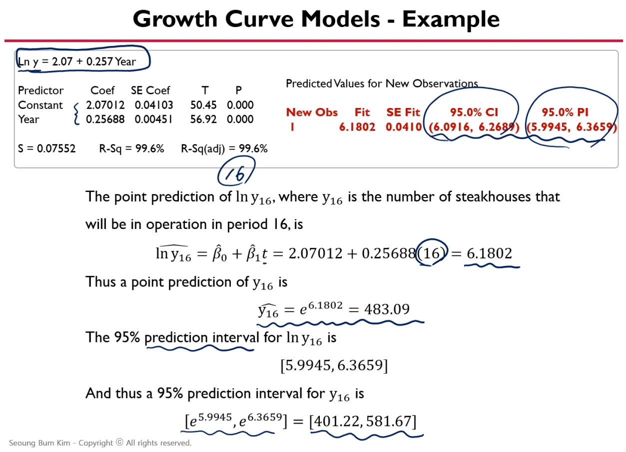
로 두는 것이다. 이 f_β를 가지고 관측치 (3)를 MSE (4)에 적용하여 최적화문제를 풀면 $\hat{f}$를 찾을 수 있을 것이다. 네번째 슬라이드에서는 관측치에 로그변환을 적용한 다음 SN_t을 모델링했으므로 나중에 다시 지수함수를 취해준다.

3 Growth Curve Models









이번 부분은 왜 있는지 모르겠다. 기존의 모델이 additive한 형식의 모델이었다면, 이건 multipicative model을 설명하려는가본데, 사실 이전 Trigonometric Model에서 로그를 취하여 모델링하는 것을 이미 공부했다. 그리고 지금 위의 캡쳐에 보이는 모델링 방법은 정확히 로그를 취한 linear model과 같아보인다.

4 First-Order Autoregressive Process

Time Series Regression with Autocorrelation

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

$\varepsilon_i \cdot \varepsilon_j$

- 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.

First-Order Autoregressive Process

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

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

- ϕ_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 ε_t tends to produce another positive error term ε_{t-1}.

ϕ_1

- If $\phi_1 < 0$, the error terms are negatively autocorrelated. In this case, a positive error term ε_{t-1} tends to produce a negative error term ε_t and vice versa.

Example

X, Y

	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 indictable-offense conviction rate, per 100,000. → 기소 가능한 범죄 유죄 판결 비율
- ftheft, women's theft conviction rate, per 100,000.
- mconvict, men's indictable-offense conviction rate, per 100,000.
- mtheft, theft conviction rate, per 100,000.

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
```


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
```


Example

(1) Ordinary regression analysis

```
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
```

(2) Time series regression analysis

```
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
```

< 0.05

- 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.

자세히 읽지 못했다. 아마도, 이전까지는 ϵ_t와 ϵ_{t-1}이 서로 독립이라는 조건 하에서 논의를 진행시켰던 것 같은데, 여기에서는 두 값이 서로 종속일 수밖에 없으니, ϵ_t가 ϵ_{t-1}로 표현될 수 있다고 보는 것 같다. 정확하게는 ϵ_t가 ϵ_{t-1}에 대한 선형함수로 모델링 된다고 보는 것이고, 그 모델링에 대한 error term을 a_t로 두었다.

$$\epsilon_t = \phi_1 \epsilon_{t-1} + a_t \tag{7}$$

ϕ_1은 선형함수에 대한 계수(coefficient) 혹은 기울기이다.

이것을 Durbin-Watson test와 연관지으려면 이전 강의의 노트에 썼던

$$\epsilon = \{\epsilon_t : t = 1, 2, \cdots, n\}$$
$$\epsilon' = \{\epsilon_{t+1} : t = 1, 2, \cdots, n\}.$$

을 가져오고, 모든 $t \in \{2, 3, \cdots, n\}$에 대하여 위의 모델링 (7)을 적용시키겠다는 말이다.

강의에서는 Durbin-Watson test에 대해 언급하면서 $\phi_1 > 0$이면 error term들이 positively autocorrelated하다고 말하고, $\phi_1 < 0$이면 negatively autocorrelated하다고 말한다. 하지만 Durbin-Watson test보다는 autocorrelation에 대한 아래 식

$$\rho_1(\epsilon)\rho_1(\epsilon) = \frac{\sum_{t=2}^{n} \epsilon_t \epsilon_{t+1}}{\sum_{t=2}^{n} {\epsilon_t}^2}$$

에서 보는 게 더 편할 것 같다. 만약 $\phi_1 > 0$이면, (a_t의 절댓값이 작을 것이므로) ϵ_t와 ϵ_{t-1}은 서로 같은 부호인 경우가 많을 것이다. 그러니까 $\epsilon_t \epsilon_{t-1}$은 양수가 되는 경우가 많고, 그걸 다 더한 $\rho_1(\epsilon)$ 값은 양수가 될 것이다. 따라서 ϵ은 postively autocorrelated하다.

반대로, 만약 $\phi_1 < 0$이면, ϵ_t와 ϵ_{t-1}은 서로 다른 부호인 경우가 많을 것이고, $\epsilon_t \epsilon_{t-1}$은 음수가 되는 경우가 많다. 따라서 ϵ은 negatively autocorrelated하다.