시계열 회귀분석(Time Series Regression - 2)

강의 : 김성범 교수님

November 20, 2022

이것은 Time Series Regression - Part 2 강의에 대한 노트이다.

1 오차와 잔차(errors and residuals)

1.1 추정에서의 오차와 잔차

임의추출을 통해 모평균을 추정하는 문제를 생각할 때, 어떤 표본 X_i에 대하여, X_i와 모평균 μ 사이의 차를 오차(error, e_i)라고 한다.

$$e_i = X_i - \mu.$$

하지만, 실제 상황에서는 모평균을 안다는 것은 불가능하다. 그래서 모평균 대신 표본평균인 $\bar{X} = \frac{X_1 + \cdots + X_n}{n}$을 사용해 그 차이를 계산할 수 있는데, 그 값이 잔차(residual, ϵ_i)이다.

$$\epsilon_i = X_i - \bar{X}.$$

1.2 회귀에서의 오차와 잔차

회귀(regression)란, 독립변수 x와 종속변수 y에 대하여 관측값(observed value)들이 주어져 있을 때, 두 변수 사이의 관계를 찾는 것이다. 조금 더 정확하게는, y를 x에 대한 함수 $y = g(x)$로서 표현하는 것이다. 관측값은 다음과 같이 주어질 것이다.

$$\text{관측값} = \{(x_i, y_i) : i = 1, 2, \cdots, N\}.$$

일부 x값들에 대해서만 y값들이 주어져있으므로 함수 g를 완벽하게 안다는 것은 불가능하다. 대신, 관측값들을 토대로 하여 만든 최적의 함수 $\hat{f}$를 생각하게 된다. 이때, 관측값 y_i에 대한 오차(error, e_i)와 잔차(residual, ϵ_i)는 다음과 같다.

$$e_i = y_i - g(x_i) \tag{1}$$

$$\epsilon_i = y_i - \hat{f}(x_i) \tag{2}$$

예측값인 $\hat{f}(x_i)$는 $\hat{y}_i$라고 쓰기도 하므로, 잔차를 다음과 같이 쓰기도 한다.

$$\epsilon_i = y_i - \hat{y}_i.$$

최적의 함수 $\hat{f}$를 찾아가는 과정은 다음과 같다. 먼저, 여러 개의 매개변수 $\beta_0, \beta_1, \beta_2, \cdots, \beta_k$를 통해 함수 $f = F[\beta_0, \beta_1, \cdots, \beta_k]$를 정의한다. 즉, 함수 f는 parametric function/model 이다. (resp. nonparametric function/model). 그리고, 그러한 f들 중에서 관측값들을 가장 잘 반영하는 함수 $\hat{f}$를 찾는다. 다시 말해, 전체적으로 y_i와 $f(x_i)$가 가장 비슷해지는 $\beta_0, \beta_1, \cdots, \beta_k$을 찾는데, 이때에 '비슷함'의 기준은 MSE 등을 통해 정할 수 있다. 가장 비슷해지는 때의 $\beta_0, \beta_1, \cdots, \beta_k$는, hat을 붙여서 $\hat{\beta}_0, \hat{\beta}_1, \cdots, \hat{\beta}_k$로 표현한다. 위의 표현대로 쓰면 $\hat{f} = F[\hat{\beta}_0, \hat{\beta}_1, \cdots, \hat{\beta}_k]$이 될 것이다.

2 여러가지 회귀모델

(앞서 말했듯이) 회귀란, 관측값들 $\{(x_i, y_i) : i = 1, 2, \cdots, N\}$로부터 x와 y 사이의 관계식 $y = \hat{f}(x)$을 찾아가는 것이다. 시계열 데이터를 다루는 만큼 index를 i 대신 t라고 쓰고, $\hat{f}(x)$를 추세(trend)를 뜻하는 TR이라는 기호로 고쳐 쓰면 식 (1)는 다음 식이 된다.

$$y_t = TR_t + e_t.$$

강의에서 위와 같은 식이 나오긴 했지만, 별로 마음에 들지 않으므로 다음 표현을 쓰는 것도 괜찮을 것 같다.

$$y_t = \hat{f}(x_t) + e_t.$$

2.1 상수회귀(no trend)

상수함수 $f(x) = \beta_0$를 생각하자. 아까 말했듯이, 이것은 parametric function이다. β_0에 의해 결정되는 함수이다. 이러한 f 중에서 최적의 $\hat{f}$를 구한다는 것은, 다음과 같은 MSE 값이 최소가 되는 때를 찾는 것이다.

$$\begin{aligned} MSE(\beta_0) &= \sum_{t=1}^{n} (y_t - f(x_t))^2 \\ &= \sum_{t=1}^{n} (y_t - \beta_0)^2 \end{aligned}$$

이때, MSE는 β_0의 값에만 의존한다. 조금 더 정확하게는, MSE는 β_0에 대한 이차함수이다. 이 이차함수의 최솟값은 중학교 3학년 과정의 방식으로 풀어도 되기는 하지만, 미분해서 0이 되는 때를 구해도 상관없다;

$$\begin{gathered} \frac{d}{d\beta_0} MSE(\beta_0) = 0 \\ 2\sum_{t=1}^{n} (y_t - \beta_0) = 0 \\ \sum_{t=1}^{n} y_t - n\beta_0 \\ \beta_0 = \frac{1}{n}\sum_{t=1}^{n} y_t. \end{gathered}$$

즉 $\hat{\beta}_0 = \frac{1}{n}\sum_{t=1}^{n} y_t$이고, $\hat{f}(x) = \frac{1}{n}\sum_{t=1}^{n} y_t$이다.

2.2 선형회귀(linear trend, linear regression)

일차함수 $f(x) = \beta_0 + \beta_1 x$를 생각하자.[1] 이것도 parametric function이다. 이 함수는 β_0와 β_1에 의해 결정되는 함수이다. 이번에도 $\hat{f}$를 구하기 위해 MSE가 최솟값이 되는 때를 찾는다. 다만, 이번에는 MSE가 두 매개변수 β_0, β_1의 함수이다.

$$MSE(\beta_0, \beta_1) = \sum_{t=1}^{n} (y_t - \beta_0 - \beta_1 x)^2$$

이것은 다변수함수라는 점에서는 아까보다 훨씬 복잡한 함수이지만, 이차함수라는 점에서는 간단하다. 아래로 볼록한 포물면을 그래프로 가지고, 단 하나의 극솟값을 가지므로 이 함수가 극솟값을 가지게 되는 (β_0, β_1)만 찾으면 그때가 최소점이 된다. 그리고

[1] 정확하게는 일차함수(linear function)라는 표현보다는 'affine 함수(affine function)'라고 써야 더 맞을 것 같다.

그건 MSE를 두 매개변수 β_0, β_1로 각각 편미분하여 0이 되는 때이다. 먼저 $\frac{\partial}{\partial\beta_0}MSE(\beta_0,\beta_1)=0$를 간단히하면

$$\frac{\partial}{\partial\beta_0}MSE(\beta_0,\beta_1)=0$$
$$-2\sum_{t=1}^{n}(y_t-\beta_0-\beta_1 t)=0$$
$$\sum_{t=1}^{n}y_t-n\beta_0-\frac{n(n+1)}{2}\beta_1=0$$

이고, $\frac{\partial}{\partial\beta_1}MSE(\beta_0,\beta_1)=0$를 간단히하면

$$\frac{\partial}{\partial\beta_1}MSE(\beta_0,\beta_1)=0$$
$$-2\sum_{t=1}^{n}t(y_t-\beta_0-\beta_1 t)=0$$
$$\sum_{t=1}^{n}ty_t-\frac{n(n+1)}{2}\beta_0-\frac{n(n+1)(2n+1)}{6}\beta_1=0$$

따라서

$$\begin{bmatrix} n & \frac{n(n+1)}{2} \\ \frac{n(n+1)}{2} & \frac{n(n+1)(2n+1)}{6} \end{bmatrix}\begin{bmatrix}\beta_0\\ \beta_1\end{bmatrix}=\begin{bmatrix}\sum y_t\\ \sum ty_t\end{bmatrix}$$

를 풀면 $\hat\beta_0$와 $\hat\beta_1$을 구할 수 있다. 그리고 좌변의 정사각행렬은 항상 역행렬이 존재하므로, 위의 연립방정식은 항상 풀린다. 이렇게 하면, 최적의 함수 $\hat f(x)=\hat\beta_0+\hat\beta_1 x$를 얻어낼 수 있다.

2.3 이차회귀(quadratic trend, quadratic regression)

이차함수 $f(x)=\beta_0+\beta_1 x+\beta_2 x^2$를 생각하자. 위와 마찬가지로

$$MSE(\beta_0,\beta_1,\beta_2)=\sum_{t=1}^{n}(y_t-\beta_0-\beta_1 x-\beta_2 x^2)^2$$

이다. 이것은 β_0, β_1, β_2에 대한 함수이고, 각각의 매개변수에 대한 편미분값을 0으로 둠으로써, 다음 연립방정식을 얻고,

$$\sum_{t=1}^{n}y_t-n\beta_0-\frac{n(n+1)}{2}\beta_1-\frac{n(n+1)(2n+1)}{6}\beta_2=0$$
$$\sum_{t=1}^{n}ty_t-\frac{n(n+1)}{2}\beta_0-\frac{n(n+1)(2n+1)}{6}\beta_1-\frac{n^2(n+1)^2}{4}\beta_2=0$$
$$\sum_{t=1}^{n}t^2y_t-\frac{n(n+1)(2n+1)}{6}\beta_0-\frac{n^2(n+1)^2}{4}\beta_1-\frac{n(n+1)(2n+1)(3n^2+3n-1)}{30}\beta_2=0$$

이 연립방정식을 풀면, MSE가 최소가 되는 β_0, β_1, β_2를 계산할 수 있다.

2.4 다항회귀(polynomial trend, polynomial regression)

마찬가지로, 다항함수 $f(x)=\beta_0+\beta_1 x+\beta_2 x^2+\cdots+\beta_k x^k$를 parametric model로 잡으면

$$MSE(\beta_0,\beta_1,\beta_2,\cdots,\beta_k)=\sum_{t=1}^{n}(y_t-\beta_0-\beta_1 x-\beta_2 x^2-\cdots-\beta_k x^k)^2$$

이고, 똑같은 작업을 통해 $\hat f(x)=\hat\beta_0+\hat\beta_1 x+\hat\beta_2 x^2+\cdots+\hat\beta_k x^k$를 계산할 수 있다.

2.1 **2.4**에서와 같이, $\hat{f}$를 계산하는 방식을 최소제곱법(최소자승법, least square method, least square estimation, the method of least square)이라고 한다.

3 자기상관성(autocorrelation)

회귀문제에서 각각의 오차(e_t)들에 대하여 전제되어야 하는 몇가지 가정들이 있었다.

- $e_t \sim N(0, \sigma^2)$: 오차는 평균이 0인 정규분포를 따른다.
- $t \neq s$이면 e_t와 e_s는 서로 독립이다.

그런데 우리가 다룰 시계열데이터의 경우, 이러한 전제가 성립되지 않을 가능성이 있다. 즉, 각각의 e_t에 대하여, e_t들이 서로 독립이 아닌 종속의 관계를 따를 수 있고, 그렇게 된다면, 최소제곱법과 같은 풀이법을 사용할 수 없을 지도 모른다. 여기에서는 공분산과 피어슨 상관계수에 대해 복습하고, 자기상관성에 대해 알아본다.

3.1 공분산(covariance)

두 확률변수 X, Y에 대하여, X와 Y의 공분산(correlation) $\mathrm{Cov}(X, Y)$는 다음과 같이 정의된다.

$$\mathrm{Cov}(X, Y) = \mathbb{E}[(X - \mathbb{E}[X])(Y - \mathbb{E}[Y])] \tag{3}$$

$\mathbb{E}$는 linear operator이므로

$$\begin{aligned}
\mathrm{Cov}(X, Y) &= \mathbb{E}[(X - \mathbb{E}[X])(Y - \mathbb{E}[Y])] \\
&= \mathbb{E}[(X - \mu)(Y - \nu)] \\
&= \mathbb{E}[XY] - \mu\mathbb{E}[Y] - \nu\mathbb{E}[X] + \mathbb{E}[\mu\nu] \\
&= \mathbb{E}[XY] - \mu\nu - \mu\nu + \mu\nu \\
&= \mathbb{E}[XY] - \mathbb{E}[X]\mathbb{E}[Y]
\end{aligned}$$

라고 쓸 수도 있다. 만약 X, Y가 독립이면, $\mathbb{E}(XY) = \mathbb{E}(X)\mathbb{E}(Y)$이므로, $\mathrm{Cov}(X, Y) = 0$이 되고 그 역도 성립한다.

관측치 $\{(x_i, y_i) : i = 1, 2, \cdots, N\}$에 대한 공분산은

$$\mathrm{Cov}(X, Y) = \frac{1}{n-1} \sum_{t=1}^{n} (x_t - \mu)(y_i - \nu)$$

으로 계산하면 되는 것 같다. 공분산에 대한 일반적인 설명과 여러가지 예제들은 StatQuest:Covariance, Clearly Explained에 가장 잘 되어 있는 것 같다.

3.2 피어슨 상관계수(correlation)

두 확률변수 X, Y에 대하여 X, Y의 피어슨 상관계수는

$$\mathrm{Cor}(X, Y) = \frac{\mathbb{E}[(X - \mathbb{E}[X])(Y - \mathbb{E}[Y])]}{\sigma[X]\sigma[Y]} \tag{4}$$

이다. $\mathrm{Cor}(X,Y)$는 -1과 1 사이의 값을 가지는데 그것은 다음과 같은 Cauchy-Schwarz 부등식 $|v|\cdot|w| \le |v\cdot w|$으로 쉽게 증명된다. [2]

$$\begin{aligned}\left|\mathbb{E}\left[(X-\mathbb{E}[X])\,(Y-\mathbb{E}[Y])\right]\right| &= \frac{1}{n-1}\left|\sum_{t=1}^{n}(x_i-\mu)(y_i-\nu)\right| \\ &= \frac{1}{n-1}\left|(x_1-\mu,\cdots,x_n-\mu)\cdot(y_1-\nu,\cdots,y_n-\nu)\right| \\ &\le \frac{1}{n-1}\left|(x_1-\mu,\cdots,x_n-\mu)\right|\cdot\left|(y_1-\nu,\cdots,y_n-\nu)\right| \\ &= \sigma(X)\sigma(Y)\end{aligned}$$

관측치 $\{(x_i,y_i): i=1,2,\cdots,N\}$에 대한 피어슨 상관계수는

$$\mathrm{Cor}(X,Y) = \frac{\frac{1}{n-1}\sum_{t=1}^{n}(x_t-\mu)(y_i-\nu)}{\sqrt{\frac{1}{n-1}\sum_{t=1}^{n}(x_t-\mu)^2}\sqrt{\frac{1}{n-1}\sum_{t=1}^{n}(x_t-\mu)^2}} = \frac{\sum_{t=1}^{n}(x_t-\mu)(y_i-\nu)}{\sqrt{\sum_{t=1}^{n}(x_t-\mu)^2}\sqrt{\sum_{t=1}^{n}(x_t-\mu)^2}}$$

으로 계산하면 되는 것 같다. 피어슨 상관계수에 대한 일반적인 설명과 여러가지 예제들은 StatQuest:Pearson's Correlation, Clearly Explained에 가장 잘 되어 있는 것 같다. 관측치의 N개 점들이 모두 한 직선 위에 있으면, 피어슨 상관계수는 (기울기가 양수일 경우) 1이거나 (기울기가 음수일 경우) -1이다.

3.3 자기상관성(autocorrelation)

시계열 데이터 $\{x_t : t = 1,2,\cdots,n\}$는 길이가 n인 수열이라고 생각할 수 있다. 이 시계열데이터를 shift 시켜서 $\{x_t : t = 2,3,\cdots,n+1\}$를 만들면

$$\{(x_t, x_{t+1}) : t = 1,2,\cdots,n-1\}$$

는 하나의 '관측치'가 된다. 이것에 대한 피어슨 상관계수를 (lag=1인) 자기상관성이라고 한다. [3] 강의에서는, 관측치에서의 첫번째 성분을 X, 뒤의 성분을 X'이라고 썼다. 그러니까,

$$\begin{aligned}X &= \{x_t : t = 1,2,\cdots,n\}\\ X' &= \{x_{t+1} : t = 1,2,\cdots,n\}\end{aligned}$$

이라고 생각하자. 그러면 자기상관성($\rho_1(X)$, autocorrelation of order 1)은[4]

$$\rho_1(X) = \mathrm{Cor}(X, X')$$

인 것이다.

이게 뭔지 잘 와닿지가 않아서, 몇개의 예를 만들어보았다. $n=6$이고 X가 선형적으로 증가하는 경우를 생각해보았다.

X	1	2	3	4	5	6	
X'		1	2	3	4	5	6

[2] X, Y가 이산확률변수인 경우만 고려했는데, 연속확률변수인 경우에도 비슷하게 증명될 수 있을 것이다.

[3] 그런데 이런 자기상관성이 왜 나오는지 잘 모르겠(었)다. **3.1.**의 앞에서 설명한 것은 오차 e_t들 사이의 독립/종속성이었지 X와 X'의 독립/종속성은 아니었다. 다시 생각해보면, 강의에서는 '일반적인' 자기상관성의 정의에 대해 설명하고 있지만, 오차 e_t에 대한 자기상관성은, 여기서 설명하는 자기상관성의 정의에 쉽게 적용할 수 있다. 강의의 27분 15초 부터 이에 대한 설명이 있다.

우리가 관심있는 것은, error term에 대한 자기상관성이다. 만약, error term의 자기상관성이 0이어서, 각 timestep마다의 error term e_t가 서로 독립적이면, 기존의 방법인 '다변수회귀'의 방법을 그대로 써도 되고, 만약 자기상관성이 0이 아니어서 e_t가 서로 종속적이라면, 다변수회귀가 아닌 다른 방법을 강구해야 한다. 하지만, 실제 상황에서 오차 e_t는 계산할 수 없으므로, 대신 잔차 ϵ_t를 고려한다. (the residuals are point estimates of the error terms.) 그리고 잔차의 자기상관성을 계산함으로써, 오차가 자기상관성을 가지는 지의 여부를 판단하게 된다. (a residual plot against time can be used to detect the violence of the independence assumption.)

[4] autocorrelation을 표현하는 표기 방식은 책마다 문헌마다 다 다른 것 같다. 여기에서는 영상의 29분 45초경에 나오기도 하고, 위키피디아에 쓰여져있기도 한 ρ(rho)를 사용했다. 아래첨자로 1을 넣은 것은 order=1이라는 의미이고 여기에서의 order는 lag를 의미한다. lag=2이면 second order, lag=3이면 third order와 같이 쓴다고 강의에 설명되어 있다.

자기상관성을 계산하기 위해서는, 관측치

$$\{(2,1),(3,2),(4,3),(5,4),(6,5)\}$$

에 대한 피어슨 상관계수를 계산하면 된다. (lag=1) 이 다섯 개의 점들은 한 직선 위에 있으므로 $\mathrm{Cor}(X, X') = 1$이 되어야 한다. 실제 계산을 해보아도

$$\begin{aligned}\mathrm{Cor}(X, X') &= \frac{(2-4)(1-3)+(3-4)(2-3)+(4-4)(3-3)+(5-4)(4-3)+(6-4)(5-3)}{\sqrt{(2-4)^2+(3-4)^2+(4-4)^2+(5-4)^2+(6-4)^2}\sqrt{(1-3)^2+(2-3)^2+(3-3)^2+(4-3)^2+(5-3)^2}} \\ &= \frac{10}{\sqrt{10}\sqrt{10}} = 1.\end{aligned}$$

이 된다. 다시 말해서, 선형적인 시계열과 그것을 shift시킨 것은 1의 상관성을 가진다.

이번에는 $n = 6$이고 X가 quadratic하게 증가하는 경우를 생각해보았다.

X	0	1	4	9	16	25	
X'		0	1	4	9	16	25

자기상관성을 계산하기 위해서는, 관측치

$$\{(1,0),(4,1),(9,4),(16,9),(25,16)\}$$

에 대한 피어슨 상관계수를 계산하면 된다. 이 다섯 개의 점들은 양의 상관관계가 있기는 해도 한 직선 위에 있지는 않으므로 $0 < \mathrm{Cor}(X, X') < 1$이면서 꽤 1에 가까운 값이 나올 것이다. 실제 계산은, 아까보다 복잡하다. 먼저 X와 X'의 표준편차에 해당하는 값을 계산하기 위해서 편차들을 적어보면 ($\mu = 11$, $\nu = 6$)

X	0	1	4	9	16	25	
X'		0	1	4	9	16	25
$X - \mu$	·	-10	-7	-2	5	14	·
$X' - \nu$	·	-6	-5	-2	3	10	·

이다. 표준편차에 해당하는 값을 계산하면

$$\sqrt{\sum_{t=1}^{5}(x_t-\mu)^2} = \sqrt{(-10)^2+(-7)^2+(-2)^2+5^2+14^2} = \sqrt{374}$$

$$\sqrt{\sum_{t=1}^{5}(x_{t-1}-\nu)^2} = \sqrt{(-6)^2+(-5)^2+(-2)^2+3^2+10^2} = \sqrt{174}$$

이다. 공분산에 해당하는 값을 계산하면

$$\sum_{t=1}^{5}(x_t-\mu)(x_{t-1}-\nu) = (-10)(-6)+(-7)(-5)+(-2)(-2)+(5)(3)+(14)(10) = 254.$$

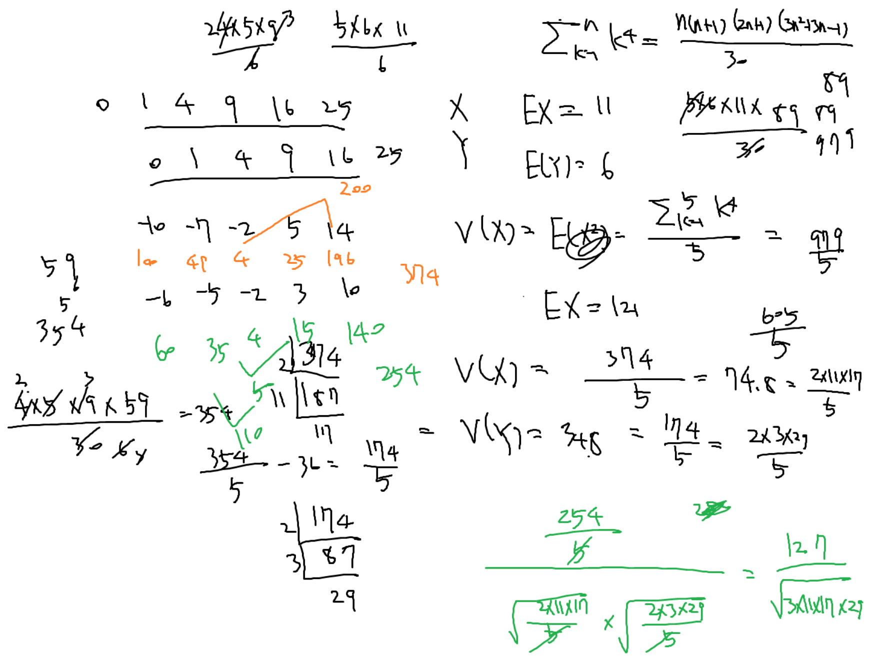
따라서, 자기상관성은

$$\mathrm{Cor}(X, X') = \frac{254}{\sqrt{374}\sqrt{174}} = 0.9957$$

와 같이 계산되고 autocorelation은 1에 거의 가까운 값을 가진다. [5]

아래는, 이걸 손으로 계산해본 것이다. 4로 나눴어야 했는데 5로 나누었기 때문에 과정은 조금 틀렸다. 하지만, 어차피 약분되기 때문에 결과적으로 autocorrelation을 계산하는 데에는 영향을 미치지는 않는다.

[5]아무리 그렇다고 해도, 1에 너무 가까운 값이 구해진 것 같다. 왜 그런지 잘 모르겠는데, 상대적으로 절댓값이 작은 0 1 정도에서 차이를 냈기 때문에 그런 걸까? 그렇지도 않을 것 같다. 다른 예를 들어보고 싶기는 한데 계산이 더 복잡할 것 같아서 안하겠다.

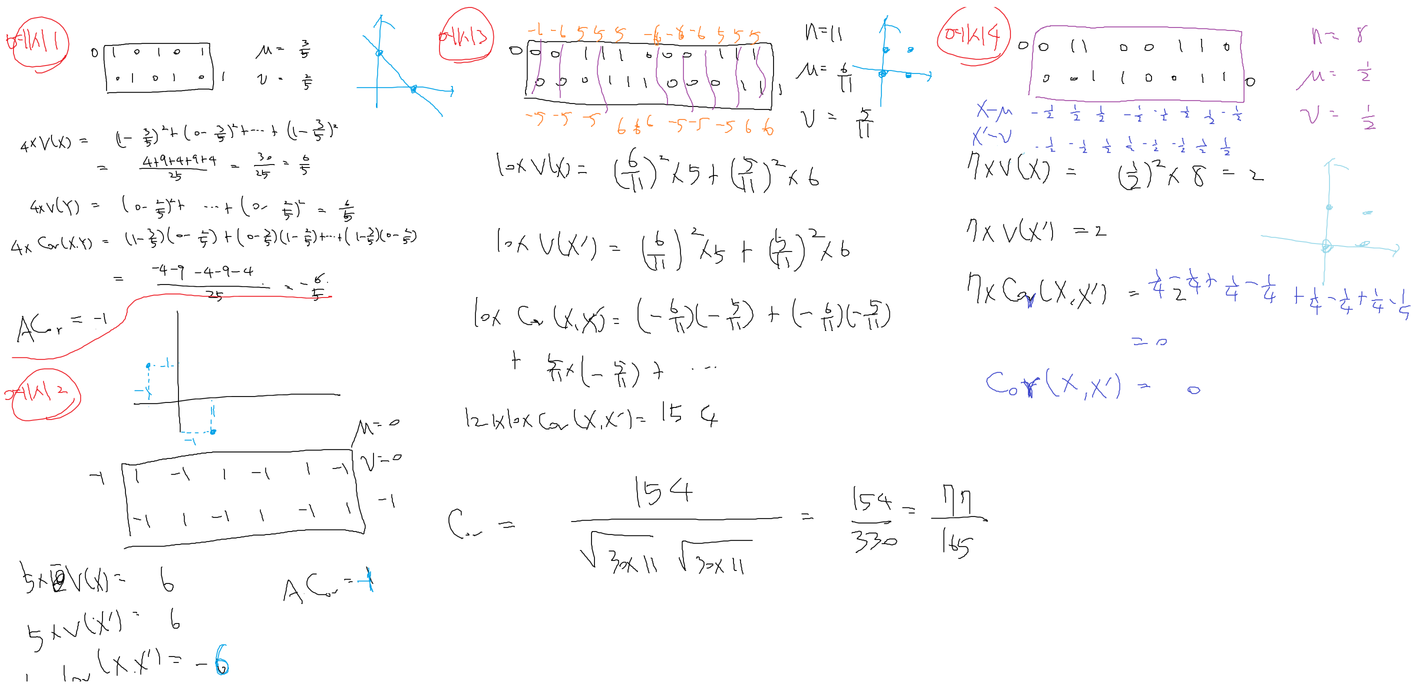


다음과 같은 생각이 들었다.

(질문) 어떤 시계열 데이터가 autocorrelation=0을 만족할 수 있을까?

위의 두 예에서는 X와 X'가 아주 강한 상관관계를 가지고 있었다. 상관계수 값이 작아지려면 어떤 시계열이어야 할까? 자기상관계수(autocorrelation)가 정확히 0이 되는 시계열이 존재할까? 그러니까, 어떤 timestep에 대해서도, 그리고 어떤 lag에 대해서도 상관관계가 0이 되도록 하는 시계열이 존재할 수 있을까? 구글에 아무리 검색을 해봐도 (e.g. “not autocorrelated time series”) 이것에 대한 말이 없다. stackoverflow에 물어봐야 할지도 모르겠다.

아래 그림과 같이 여러가지로 시도해봤는데, autocorrelation이 0이 되는 (lag=1) 시계열을 발견했다.



맨 오른쪽 계산(예시 4)이다. 이걸 다시 정리해보면 ($n = 8$, $\mu = \nu = \frac{1}{2}$)

X	0	0	1	1	0	0	1	1	0	
X'		0	0	1	1	0	0	1	1	0
$X-\mu$	$\cdot$	$-\frac{1}{2}$	$\frac{1}{2}$	$\frac{1}{2}$	$-\frac{1}{2}$	$-\frac{1}{2}$	$\frac{1}{2}$	$\frac{1}{2}$	$-\frac{1}{2}$	$\cdot$
$X'-\nu$	$\cdot$	$-\frac{1}{2}$	$-\frac{1}{2}$	$\frac{1}{2}$	$\frac{1}{2}$	$-\frac{1}{2}$	$-\frac{1}{2}$	$\frac{1}{2}$	$\frac{1}{2}$	$\cdot$

이다. 표준편차에 해당하는 값을 계산하면

$$\sqrt{\sum_{t=1}^{8}(x_t-\mu)^2}=\sqrt{\frac{1}{4}\times 8}=\sqrt{2}$$
$$\sqrt{\sum_{t=1}^{8}(x_{t-1}-\nu)^2}=\sqrt{\frac{1}{4}\times 8}=\sqrt{2}$$

이다. 공분산에 해당하는 값을 계산하면

$$\sum_{t=1}^{8}(x_t-\mu)(x_{t-1}-\nu)=\frac{1}{4}-\frac{1}{4}+\frac{1}{4}-\frac{1}{4}+\frac{1}{4}-\frac{1}{4}+\frac{1}{4}-\frac{1}{4}=0.$$

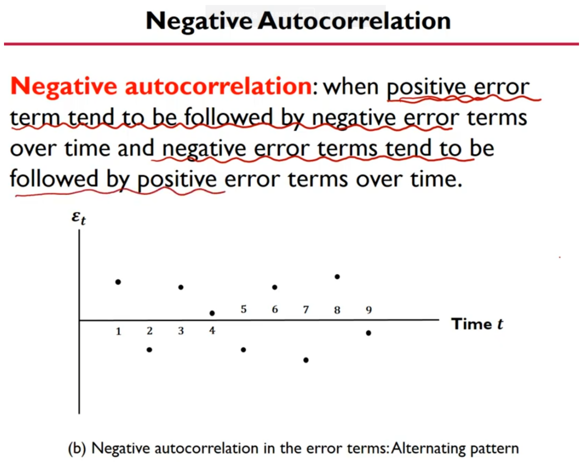
따라서,

$$\mathrm{Cor}(X,X')=\frac{0}{\sqrt{2}\sqrt{2}}=0$$

와 같이 계산되고 autocorelation은 0이다.

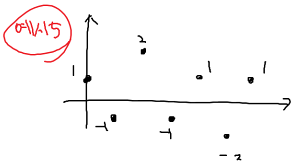
이것은 조금 더 쉽게 해석될 수도 있다. 위 그림의 예시 1은, 좌표평면 상에 두 점 $(1,0)$, $(0,1)$을 찍어 그 두 점에 대한 correlation을 계산하는 것이다. 그러면 당연히, 두 점을 잇는 직선이 존재하고 그 기울기가 음수니까, correlation은 -1이 될 수밖에 없었다. 마찬가지로, 예시 2는, 두 점 $(1,-1)$, $(-1,1)$을 찍어 그 두 점에 대한 correlation을 계산하는 것이고, 이번에도 correlation은 -1이다. 예시 3에서는 네 점 $(0,0)$, $(1,0)$, $(1,1)$, $(0,1)$의 네 점에 대하여 correalation을 계산하는 것이라고 해석할 수 있는데, 조금 더 정확하게는 4개의 점이라기 보다는, 11개의 점이고, $(0,0)$은 4개, $(1,0)$은 2개, $(0,1)$는 1개, $(1,1)$는 4개를 고려하는 것이다. 다시 말해, 네 개의 점이기는 하지만 가중치가 부여된 네 개의 점이기 때문에, 이 네 개의 점에 대해서는 correlation이 0인 것이 아니다. $(0,0)$과 $(1,1)$에 가중치가 상대적으로 많이 들어간 셈이므로 correlation은 양수가 되는 것이 맞고, 실제로도 계산이 그렇게 나온다. 마지막으로, 예시 4는 네 점 $(0,0)$, $(1,0)$, $(1,1)$, $(0,1)$이 equally weighted된 것이라고 볼 수 있다. 이렇게 찍힌 네 점에 대하여는 상관계수가 0이 되는 것이 당연하다. 그래서, 실제 계산도 0이 되었다.

그러면, 실제 강의에서 negative autocorrelation에 대한 내용이 얼추 설명이 된 것 같기도 하다. 그 부분의 캡쳐를 떠오면



(b) Negative autocorrelation in the error terms: Alternating pattern

인데, 이런 패턴(alternating pattern)의 시계열을 예시 1과 예시 2에서 봤었고, autocorrelation이 -1이 된다는 걸 확인했었다. 예시 1과 2에서처럼 일정한 진폭으로 진동하는 게 아니라, 진폭이 조금 변하면서 진동하는 예를 임의로 만들어보면



X	1	-1	2	-1	1	-2	1	
X'		1	-1	2	-1	1	-2	1

$\mu = 0$ $n = 6$

$\nu = 0$

$5 \times V(X) = 1+4+1+1+4+1 = 12$

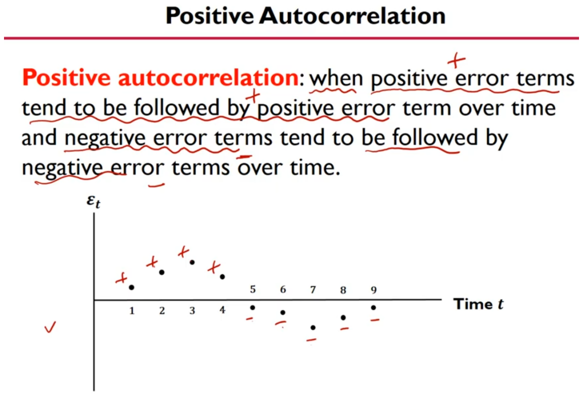
$5 \times V(X') = 1+1+4+1+1+4 = 12$

$5 \times Cov(X, X') = -1-2-2-1-2-2 = -10$

$Cor(X, X') = \frac{-10}{\sqrt{12}\sqrt{12}} = -\frac{5}{6}$

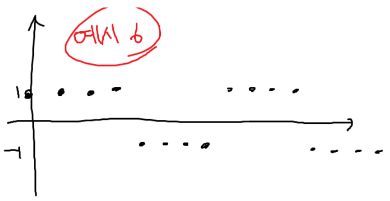
처럼 만들 수도 있다.

실제 강의에서 positive autocorrelation에 대해 설명한 내용에 관해서는 (아래 캡쳐)



(a) Positive autocorrelation in the error terms: Cyclical pattern

예시를 적절하게 든 적이 없다. 마치 사인함수나 코사인함수 같은 예를 들고 있는데 사인함수 같은건 계산하기가 좀 귀찮을 것 같..으니 간단한 예를 먼저 들면 (예시6)



1	1	1	1	-1	-1	-1	-1	1	1	1	1	-1	-1	-1	1	
	1	1	1	1	-1	-1	-1	-1	1	1	1	1	-1	-1	-1	1

$n = 12$

$\mu = 0$

$\nu = 0$

$11 \times V(X) = 12$

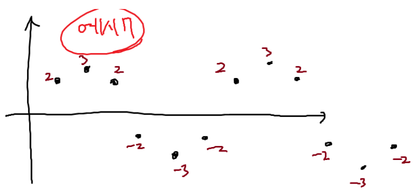
$11 \times V(X') = 12$

$11 \times Cov(X, X') = 3-1+3-1+3-1+3-1 = 8$

$Cor(X, X') = \frac{8}{\sqrt{12}\sqrt{12}} = \frac{2}{3}$

와 같이 된다. 이건 예시 4와 예시 3을 조금 연장한 것이다. lag를 1로 고정한 상태에서, 예시 4에서는 alternating 주기가 4였고, 예시 3에서는 alternating 주기가 6이었다면 이번 예시 6에서는 alternating 주기가 8로 늘어났다. 각각의 자기상관성은 예시 4에서 0, 예시 3에서 $\frac{77}{165}$쯤 되는 것 같더니 예시 6에서는 $\frac{2}{3}$으로 증가했다. 아마 alternating 주기를 더 늘려가면 자기상관성은 더 커질 것이다. 이것은 강의에서 설명하고 있는, "양수 이후에 양수가 나오고, 음수 이후에 음수가 나오는" 경향을 말하는 것 같다.

이번에는 강의 그림에서처럼 사인곡선과 조금 비슷하게 만들어보았다.



2	3	2	-2	-3	-2	2	3	2	-2	-3	-2	2	
	2	3	2	-2	-3	-2	2	3	2	-2	-3	-2	2

$n = 12$

$\mu = 0$

$\nu = 0$

$11 \times V(X) = 9 \times 4 + 4 \times 8 = 68$

$11 \times V(X') = 68$

$Cov(X, X') = (6+6-4-6+6-4) \times 2$

$= 8$

$Cor(X, X') = \frac{8}{\sqrt{68}\sqrt{68}} = \frac{2}{17}$

이 때에도 positive autocorrelation이 계산된다.

4 Durbin-Watson Test

4.1 autocorrelation applied to the residual term

지금까지, 어떤 시계열 데이터에 대한 자기상관성(autocorrelation)에 대해 말했다. 하지만, 3번 각주에서 말했듯이, 우리가 관심있는 것은 error term(오차, e_t)의 자기상관성이고, 실제로는 error term을 알 수 없으니 대신에 residual term(잔차, ϵ_t)의 자기상관성을 다룬다. residual term들의 수열을 그냥 ϵ이라고 하자.

$$\epsilon = \{\epsilon_t : t = 1, 2, \cdots, n\}$$

그리고 ϵ을 한 timestep 민 것을 ϵ'이라고 하자 (first order; lag=1)

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

그러면, 여기에서 구해야 하는, residual term의 autocorrelation $\rho_1(\epsilon)$은 다음과 같이 표현된다.

$$\begin{aligned}\rho_1(\epsilon) &= \mathrm{Cor}(\epsilon, \epsilon') \\ &= \frac{\sum_{t=1}^{n}(\epsilon_t - \epsilon_{t+1})^2}{}\end{aligned}$$

5 Increasing Seasonal Variations