시계열 회귀분석(Time Series Regression - 2)

강의 : 김성범 교수님

November 19, 2022

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 $\hat{f} = F[\hat{\beta}_0, \hat{\beta}_1, \cdots, \hat{\beta}_n]$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;

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

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 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

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

따라서

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

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

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

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

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

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

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

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

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.

이와같은 방식으로 $\hat{f}$를 계산하는 방식을 최소제곱법(최소자승법, least square method, least square estimation)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 된다.

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 $|v| \cdot |w| \leq |v \cdot w|$으로 쉽게 증명된다.[2]

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

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

4

5

6

7