회귀분석&시계열분석 간단정리

황정민

1. 회귀

1-1) 선형회귀

선형회귀: 종속변수와 설명변수(들)의 선형 상관관계를 나타내는 모형입니다.

$$y = f(\mathrm{x}) + \epsilon = \beta_0 + \beta_1 x_1 \ldots + \beta_p x_p + \epsilon = \beta_0 + \sum_{j=1}^{p} \beta_j x_j + \epsilon$$

이를 행렬식을 사용해 표현해보면 다음과 같이 간단하게 나타낼 수 있습니다.

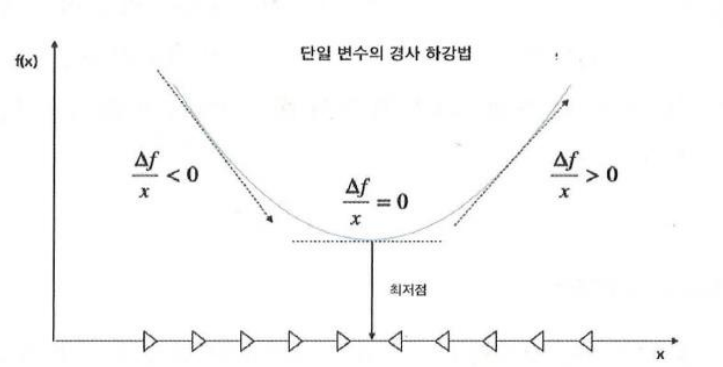
$$\underset{(N \times 1)}{\boldsymbol{y}} = \underset{(N \times P)}{\boldsymbol{X}} \underset{(P \times 1)}{\boldsymbol{\beta}} + \underset{(N \times 1)}{\boldsymbol{\epsilon}}$$

데이터에서 모델 파라미터를 학습할 때 사용할 수 있는 방법에는 최소자승법, 최대우도추정법, 경사하강법 등이 있습니다.

***최소자승법(OLS)**: 입력 데이터로부터 출력에 가장 근접한 초평면을 찾는 방법. 즉, 잔차제곱합을 최소화하는 ß 를 선택하는 방법

***최대우도추정법(MLE)**: 우도함수(Likelihood function)의 최대값이 나오도록 하는 ß 를 선택하는 방법

***경사하강법**: 함수 값이 낮아지는 방향으로 설명 변수의 값을 변형시켜가면서 최종적으로는 손실 함수의 값이 최소가 되도록 하는 ß 를 찾는 방법. 1 차 미분계수를 이용해 함수의 최소값을 찾아가는 iterative 한 방법입니다.



경사하강법을 사용하는 이유:

* 함수들 중, 닫힌 형태(closed form)가 아니거나 함수의 형태가 복잡해 (가령, 비선형함수) 미분계수와 그 근을 계산하기 어려운 경우가 많다.
* 실제 미분계수를 계산하는 과정을 컴퓨터로 구현하는 것에 비해 gradient descent 는 컴퓨터로 비교적 쉽게 구현할 수 있다.
* 데이터 양이 매우 큰 경우 gradient descent 와 같은 iterative 한 방법을 통해 해를 구하면 계산량 측면에서 더 효율적으로 해를 구할 수 있다.

***확률적 경사 하강법(SGD):**

속도 향상을 위해 무작위로 데이터 포인트를 선택하고, 이 데이터 포인트에 대한 그레디언트를 계산하는 방법

1-2) 규제화

규제화(Regularization): 회귀계수에 제약을 가함으로써 과적합을 방지하고 모델의 일반화 성능을 높이는 기법입니다. 즉, 편향(bias)을 조금 허용하는 대신, 분산(variance)을 줄이는 것입니다. 규제화의 대표적인 예로는 Ridge, Lasso, Elastic Net 등이 있습니다.

***Ridge**: 기존의 손실 함수에 ß 의 L2-norm 을 패널티로 적용하여 회귀계수를 추정합니다.

$$\hat{\beta} = \arg\min_{\beta}\left((Y - X\beta)^T (Y - X\beta) + \lambda ||\beta||_2^2 \right)$$

***Lasso**: Ridge 와 달리 ß 의 L1-norm 을 패널티로 적용하여 회귀계수를 추정합니다.

$$\hat{\beta} = \arg\min_{\beta}\left((Y-X\beta)^T(Y-X\beta)+\lambda||\beta||_1\right)$$

***Elastic Net**: L1-norm, L2-norm 패널티를 결합한 방법

$$\hat{\beta} = \arg\min_{\beta}\left((Y-X\beta)^T(Y-X\beta)+\lambda\left(\alpha||\beta||_2^2 + (1-\alpha)||\beta||_1\right)\right)$$

1-3) 평가

적합도(Goodness-of-fit)는 모델이 결과의 변동을 얼마나 잘 설명하는 지를 평가합니다.
R^2, AIC, BIC 등을 사용해 적합도를 평가할 수 있습니다.
***R^2(결정계수)**: 모델에 의해 설명되는 목표변수의 변동 비율을 측정합니다.

$$R^2 = 1 - \frac{\text{RSS}}{\text{TSS}}$$

***AIC(아카이케 정보 기준), BIC(베이지안 정보 기준)**: 최대 우도 추정을 기반으로 합니다.
이 값들이 작을수록 모델의 적합도가 높다는 것을 나타냅니다.

AIC = $-2\log(\mathcal{L}^*) + 2k$, 여기서 $\mathcal{L}^*$는 최대 우도 함수의 값이며, k는 파라미터의 수다.

BIC = $-2\log(\mathcal{L}^*) + \log(N)^k$, 여기서 N은 표본 크기다.

(BIC 의 화살표는 -입니다)

이 밖에도
수정결정계수(Adjusted R^2)-클수록 모델의 적합도가 높습니다.
맬로우스의 C_p(Mallows's Cp)- 작을수록 모델의 적합도가 높습니다.
등이 모델의 적합도를 평가하는 데 쓰일 수 있습니다.

2. 시계열모형

2-1) 평활법

평활법: 과거 및 현재 자료의 불규칙 변동을 부드럽게 평활(smoothing)시켜 미래의 값을 예측하는 기법입니다. 이동평균법, 지수이동평균법 등이 있습니다.

***(단순)이동평균법**:

$$MA_n = \frac{Z_n + Z_{n-1} + \cdots + Z_{n-m+1}}{m} = \frac{1}{m}\sum_{t=n-m+1}^{n} Z_t$$

***지수이동평균법**: 과거의 모든 기간을 계산대상으로 하며 최근의 데이터에 더 높은 가중치를 두는 기법입니다.

$$\boldsymbol{S_{n+1} = \alpha Z_{n+1} + (1-\alpha)S_n}$$

(S_n 은 n 시점에서의 지수이동평균, α 는 평활 계수)

1-2) 정상성

***정상성**: 시계열의 확률적인 성질들이 시간의 흐름에 따라 불변함을 의미합니다. 만약, 시계열이 정상시계열이라면 뚜렷한 추세가 존재하지 않고, 시계열의 진폭이 시간의 흐름에 따라 변화하지 않습니다. 정상시계열의 대표적인 예로는 백색잡음과정(White Noise Process)가 있습니다.

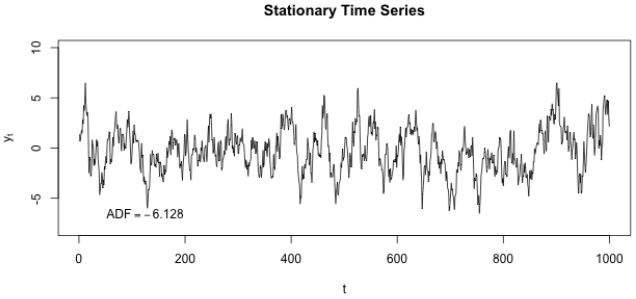
***강정상성**: 모든 n 에 대하여 결합확률밀도함수가 시간대를 바꿔도 동일하다는 가정입니다. 즉, 어떠한 시간대이든 분포가 항상 동일합니다. 강정상성을 만족하는 시계열 자료는 거의 없습니다.

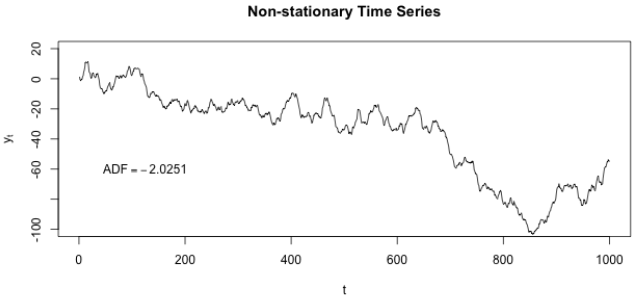
$$f(Z_{t_1}, \cdots, Z_{t_n}) = f(Z_{t_1+k}, \cdots, Z_{t_n+k})$$

***약정상성**: 시계열이 다음의 3 가지 조건을 만족시키면 됩니다.

1) 모든 t 에 대해 평균이 일정
2) 모든 t 에대해 분산이 일정하고, 유한하다.
3) 두 시점 사이의 자기공분산은 시점 t 와는 무관하며, 시차에만 의존한다. (이를 시간불변성이라고도 한다)

이후에 사용할 모델들이 성립하기 위해서는 시계열이 정상시계열이라는 가정이 필요합니다. 만약, 시계열이 비정상시계열이라면 적절한 변환을 통해 시계열을 정상시계열로 만들어 줘야합니다.





1-3) 시계열 모델 1

***AR 모형(자기회귀모형)**: 시계열 Z_t를 그 이전 시점의 시계열(Z_{t-1}, Z_{t-2},...)로 회귀시킨 모형으로 p 차 자기회귀모형 AR(p)는 다음과 같습니다. 즉, Z_t가 Z_{t-1}, Z_{t-2},..., Z_{t-p} 만의 선형결합, 즉(t-p) 시차 까지만 영향을 받는다고 가정하고 설정된 모형입니다.

$$Z_t = \phi_1 Z_{t-1} + \phi_2 Z_{t-2} + \phi_3 Z_{t-3} + \cdots + \phi_p Z_{t-p} + a_t, \quad (a_t\text{는 } iid\text{인 } white\ noise)$$

***MA 모형(이동평균모형)**: 시계열 Z_t를 그 이전 시점의 백색잡음과정(White Noise Process)로(a_{t-1}, a_{t-2},...)로 회귀시킨 모형으로 q 차 자기회귀모형 MA(q)는 다음과 같습니다.

$$Z_t = a_t - \theta_1 a_{t-1} - \theta_2 a_{t-2} - \theta_3 a_{t-3} - \cdots - \theta_q a_{t-q}$$

***ARMA 모형(자기회귀이동평균모형)**: AR 모형과 MA 모형을 결합시킨 모형으로, ARMA(p, q)는 다음과 같이 나타낼 수 있습니다.

$$Z_t = \phi_1 Z_{t-1} + \phi_2 Z_{t-2} + \phi_3 Z_{t-3} + \cdots + \phi_p Z_{t-p} + a_t - \theta_1 a_{t-1} - \theta_2 a_{t-2} - \theta_3 a_{t-3} - \cdots - \theta_q a_{t-q}$$

***ARIMIA(누적자기회귀이동평균모형)**: 원시계열이 비정상시계열이고, 평균이 일정하지 않은 경우, 추세를 제거(평균 안정화)하기 위해 차분을 사용할 수 있습니다. 이 때, 원시계열 Z_t를 d 번 차분한 시계열이 ARMA(p,q)를 따를 때, Z_t는 ARIMA(p,d,q)를 따른다고 합니다.

***SARIMA(계절 ARIMA):** 시계열 자료에 계절성을 해결하기 위해 ARIMA 모델에 추가적으로 계절성 항을 포함하여 구성한 모델입니다.

$$\text{ARIMA} \quad \underbrace{(p,d,q)}_{\substack{\uparrow \\ \text{모델의} \\ \text{비 계절성 부분}}} \quad \underbrace{(P,D,Q)_m}_{\substack{\uparrow \\ \text{모델의} \\ \text{계절성 부분}}}$$

1-4) 시계열 모델 2

대부분의 경제, 금융 관련 시계열에서 전차는 White Noise Process 처럼 보이지만 잔차의 절대값 또는 잔차 제곱은 자기상관 관계를 갖습니다. 이러한 시계열의 이분산성을 반영한 모형이 ARCH, GARCH 입니다.

***ARCH(자귀회귀 조건부 이분산 모형)**: 시계열 r_t는 같은시점에서의 표준편차와 잡음의 곱으로 표시되고, 분산은 하나의 상수를 포함하여 r_t의 과거 관측값들의 제곱의 일차결합으로 표시됩니다. ARCH(p)모형은 다음과 같이 나타낼 수 있습니다.

$$r_t = \sigma_t \varepsilon_t, \quad \varepsilon_t \sim iid\ N(0,1)$$

$$\sigma_t^2 = \alpha_0 + \alpha_1 r_{t-1}^2 + \alpha_2 r_{t-2}^2 + \cdots + \alpha_p r_{t-p}^2 \quad \text{<-분산방정식(변동방정식)}$$

(단, $\alpha_0 > 0, \alpha_i \geq 0, \forall i > 0$)

***GARCH(일반화된 자귀회귀 조건부 이분산 모형)**: ARCH 모델을 확장시킨 것으로, 분산을 r_t의 과거 관측값들의 제곱과 이전 분산 값들의 일차결합으로 표시됩니다. GARCH(p, q)모형은 다음과 같이 나타낼 수 있습니다.

$$r_t = \sigma_t \varepsilon_t, \quad \varepsilon_t \sim iid\ N(0,1)$$

$$\sigma_t^2 = \alpha_0 + \alpha_1 r_{t-1}^2 + \alpha_2 r_{t-2}^2 + \cdots + \alpha_p r_{t-p}^2 + \beta_1 \sigma_{t-1}^2 + \beta_2 \sigma_{t-2}^2 + \cdots + \beta_q \sigma_{t-q}^2$$

$$= \alpha_0 + \sum_{i=1}^{p} \alpha_i r_{t-i}^2 + \sum_{j=1}^{q} \beta_j \sigma_{t-j}^2$$

(단, $\alpha_0 > 0, \alpha_i \geq 0, \beta_i \geq 0, \sum_{i=1}^{p} \alpha_i + \sum_{j=1}^{q} \beta_j < 1, \forall i > 0$)

이 밖에도 VAR, ARMAX 등의 모형이 시계열 자료를 분석하는데 사용되고 있습니다.