글로벌투자전략팀

김동영, CFA
Quant Analyst
dy76.kim@samsung.com

안미성, Ph.D.
Economist
misung11.ahn@samsung.com

장준희
Research Associate
junhee259.jang@samsung.com

퀀트 모델링 A to Z

(1) 회귀분석

- 다양한 퀀트 모델링 기법에 대해 이론 설명과 구체적인 사용법을 설명할 계획
- 회귀분석 모델에 대한 설명 및 사용 방법 수록

서론: 우리는 이번 기획을 통해서 자산 운용에 도움이 되는 여러가지 모델링 방법을 설명하고, 예제를 통한 구체적인 사용법 또한 전달하려고 한다. 본 기획은 기초적인 회귀분석 모델부터 출발하여 10여가지 이상의 시리즈로 발간할 계획이다.

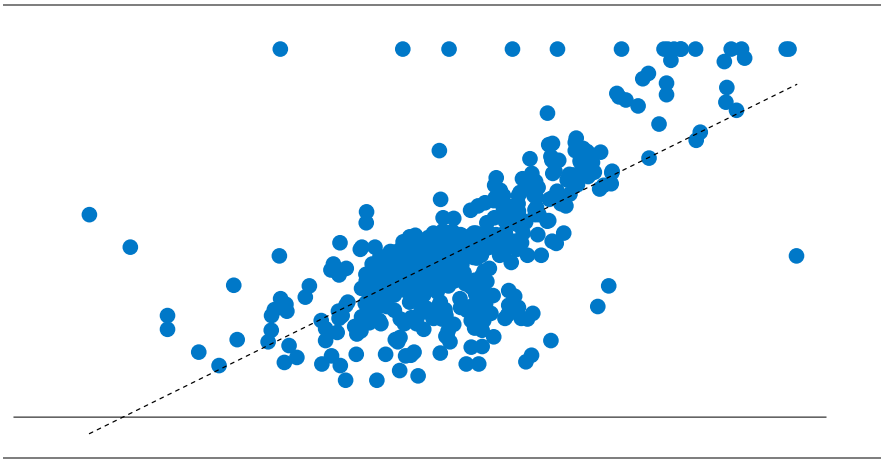
회귀분석: 회귀분석은 독립변수들과 종속변수의 선형관계를 분석한다. 일반적으로 회귀분석 모델은 1) 단순 선형 회귀분석 모델(simple linear regression model)과 2) 다중 회귀분석 모델(multiple regression model)을 지칭한다. 회귀분석 모델은 단순하다고 여겨진다. 하지만, 기본 회귀분석은 로짓(logit), 프로빗(probit)모델 혹은 머신러닝 기법으로 이야기되는 Ridge 회귀, Lasso 회귀 등의 이론적 기초가 되므로 원리를 이해하는 것이 필요하다.

(엑셀에서의 사용법, 파이썬에서의 사용법 수록)

[단순 선형 회귀분석 모델]

$$y = \beta_0 + \beta_1 x + u$$

여기서, y는 종속변수, x는 독립변수, u는 오차항, β_1는 기울기 모수(parameter), β_0는 절편 모수임



I. 서론

자산 운용과 주식 투자에서 감에 의한 투자를 지양하려면, 수치적인 분석 과정이 필요하다. 이 때 사용되는 것이 퀀트(계량적) 모델이다. 과거의 데이터로 현상을 파악하고, 이 패턴을 근거로 미래를 예측하는 퀀트 모델에는 회귀분석에서부터 시계열 분석, 머신러닝 기법 등이 다 포함된다. 과거에는 이론적인 학습 이후 실제 모델을 구현하는 데도 많은 노력이 들어갔다. 하지만 현재는 다양한 툴이 개발되면서, 모델의 실제 사용도 쉬운 환경으로 변화했다. 따라서, 분석의 정교함과 유용성을 올릴 수 있는 많은 길이 생긴 셈이다.

우리는 이번 기획을 통해서 자산 운용에 도움이 되는 여러가지 모델링 방법을 설명하고, 예제를 통한 구체적인 사용법 또한 전달하려고 한다. 본 기획은 기초적인 회귀분석 모델부터 출발하여 10여가지 이상의 시리즈로 발간할 계획이다.

II. 단순 선형 회귀분석

회귀분석은 독립변수들과 종속변수의 선형관계를 분석한다. 일반적으로 회귀분석 모델은 1) 단순 선형 회귀분석 모델(simple linear regression model)과 2) 다중 회귀분석 모델(multiple regression model)을 지칭한다. 회귀분석 모델은 단순하다고 여겨진다. 하지만, 기본 회귀분석은 로짓(logit), 프로빗(probit)모델 혹은 머신러닝 기법으로 이야기되는 Ridge 회귀, Lasso 회귀 등의 이론적 기초가 되므로 원리를 이해하는 것이 필요하다.

단순 회귀분석 모델은 1개의 독립변수(x)와 1개의 종속변수(y) 간의 관계를 조사하는 모형이다. 기술적으로, 회귀분석 모형은 x의 한 단위 변화가 y의 몇 단위 변화를 초래하는지를 추정(estimate)한다. 이를 이용하여 분석자는 정성적으로 수립한 x와 y간의 관계에 대한 가설을 회귀분석 모델을 통해 정량적으로 테스트할 수 있다. 모형식은 다음과 같이 종속변수와 독립변수 간의 일차 함수식이다.

[단순 선형 회귀분석 모델]

$$y = \beta_0 + \beta_1 x + u$$

여기서, y는 종속변수, x는 독립변수, u는 오차항, β_1는 기울기 모수(parameter), β_0는 절편 모수임

이 식은 1) y라는 경제 변수의 변동이 x라는 경제 변수의 변동에 따라서 같이 움직이고, 2) 둘 간의 관계가 선형 관계로 나타낼 수 있다는 뜻이다.

회귀 분석을 한다는 것은, x와 y간 선형 관계를 결정하는 β_0, β_1라는 2개의 모수를 주어진 표본을 이용하여 추정하는 작업을 말한다. 이 때 보통최소제곱법(ordinary least squares, OLS)을 주로 사용한다. OLS는 오차항의 제곱의 합($\sum u^2$)을 최소화하는 두 모수(β_0, β_1)를 한정된 표본에서 추정한다. 표본에서 추정된 모수를 추정치(estimate)라고 부르며, $\hat{\beta_0}, \hat{\beta_1}$이라고 주로 표기한다. (이론 계량경제학은 한정된 표본에서 추정된 추정치($\hat{\beta_0}, \hat{\beta_1}$)가 특정 조건 하에서 모수($\beta_0, \beta_1$)에 근사함을 엄밀하게 증명한다.)

OLS의 장점은 1) 까다롭지 않은 조건 하에서 추정치가 모수에 근사하며, 2) 해당 추정치를 다음과 같이 공식 형태(estimator, 추정량)로 바로 구할 수 있다는 점이다.

[β_0, β_1의 추정 (OLS 방법)]

우선, $i = 1,2, \dots, n$개의 관측치(observations)로 구성되어 있는 표본에서 표본분산과 표본공분산은 다음과 같이 정의된다.

$$\textit{표본분산} var(x) = {s_x}^2 = \frac{\sum_{i=1}^{n}(x_i - \bar{x})^2}{n-1}$$

$$\textit{표본공분산} cov(x,y) = \frac{\sum_{i=1}^{n}(x_i - \bar{x})\,(y_i - \bar{y})}{n-1}$$

여기서 $\bar{x}$는 x의 평균, $\bar{y}$는 y의 평균임. s_x는 표본표준편차임

기울기 모수(β_1, 통상적인 베타)와 절편 모수(β_0)의 OLS 추정량(차례로 $\widehat{\beta_1}$, $\widehat{\beta_0}$)은 표본분산과 표본공분산을 이용하여 다음과 같이 표현할 수 있다.

$$\widehat{\boldsymbol{\beta_1}} = \frac{cov(x,y)}{var(x)} = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^{n}(x_i - \bar{x})^2}$$

$$\widehat{\boldsymbol{\beta_0}} = \bar{y} - \widehat{\beta_1}\bar{x}$$

현대 사회에서, 해당 공식을 직접 써서 회귀계수를 손으로 계산하는 경우는 거의 없다. 엑셀 프로그램에서 *slope*, *intercept* 함수를 쓰면 단순 선형 회귀분석 모델의 회귀계수를 즉각 구할 수 있다.

회귀분석 모형을 수립한 다음에는 과연 x가 y를 잘 설명하는가에 대한 의문을 가질 수 있다. 일반적으로 두 변수의 선형 관계의 강도를 나타내는 지표로 상관계수(coefficient of correlation)가 있고, 두 개 이상의 변수 간 선형 관계에도 적용 가능한 결정계수(coefficient of determination) 지표가 있다. 이 지표는 단순 선형 회귀분석에서도 동일하게 사용된다.

[상관계수 (coefficient of correlation)]

$$\textit{상관계수}\ r_{xy} = \frac{cov(x,y)}{s_x s_y} = \frac{\sum_{i=1}^{n}(x_i - \bar{x})\,(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2}\sqrt{\sum_{i=1}^{n}(y_i - \bar{y})^2}}$$

여기서 $cov(x,y)$는 표본공분산, s_x는 x 의 표본표준편차, s_y는 y 의 표본표준편차임

분산은 특정 한 개 변수가 고정된 위치(평균값)에서 얼마만큼 변동하는지를 나타내는 수치다. 공분산은 x, y 두 개 변수가 고정된 위치(평균값)에서 얼마나 동시에 변동하는지를 나타내는 수치다.

위 식에 있는 "상관계수"는 공분산을 x와 y의 표준편차로 나눠서, x와 y가 "같이" 변동하는 정도를 단위(unit)를 제거하여 나타낸 수치다. 단위를 제거하는 과정에서 상관계수는 1에서 -1 사이 값을 갖도록 조정된다. 1에 가까울수록 둘 간에 양의 선형 관계성이 강함을, -1에 가까울수록 음의 선형 관계성이 강함을 의미하고, 0에 가까울수록 선형 관계성이 없음을 의미한다.

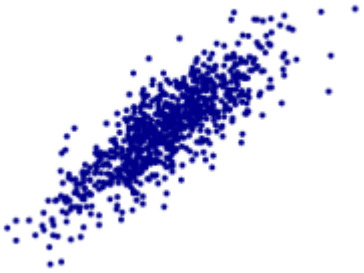
일견, 상관계수 공식($r_{xy} = \frac{cov(x,y)}{s_x s_y}$)과 기울기 모수 추정량($\hat{\beta}_1 = \frac{cov(x,y)}{{s_x}^2}$)은 서로 비슷해 보인다. 그러나, 1) 상관계수는 x와 y간 선형 관계의 정도를 나타내는 지표이고, 2) 기울기 모수 추정량(베타)은 y의 x에 대한 민감도 수준을 나타내는 지표라는 것을 서로 구별할 필요가 있다.

일반적인 모델의 유의성을 볼 때는 상관계수가 기준이 된다.

한편, 예를 들어 종속변수(y)에 A주식과 B주식이 각각 쓰인 2개의 회귀분석 모델을 서로 비교해서 독립변수의 영향 강도를 비교할 때는, 기울기 모수(베타) 추정치가 중요해진다.

[상관계수 높음, 베타 낮음]

[상관계수 낮음, 베타 높음]



두 변수 간 선형 관계의 강도를 나타내는 또 다른 지표는 결정계수다. 결정계수는 독립변수(x)의 변동에 의해 설명되는 종속변수(y)의 변동 정도를 측정한 지표다. 앞서 말했듯 결정계수는 여러 독립변수 들과 종속변수 간의 선형성을 수치화한다는 점에서 상관계수보다 더 일반적이다.

[결정계수 (coefficient of determination)]

회귀계수가 확정된 모든 회귀분석 모델에서,

[i 번째 종속변수 관측치 y_i]의 [평균($\bar{y}$)]으로부터의 변동량은

1) [관측치(y_i)]의 [회귀분석 모델 상 y_i의 추정치($\hat{y}_i$)]으로부터의 변동량과

2) [회귀분석 모델 상 y_i의 추정치($\hat{y}_i$)]의 [평균($\bar{y}$)]으로부터의 변동량으로 분리할 수 있다.

즉, $y_i - \bar{y} = (y_i - \hat{y}_i) + (\hat{y}_i - \bar{y})$ 이다.

여기서, 각 항의 제곱합 간의 관계는 다음과 같다.

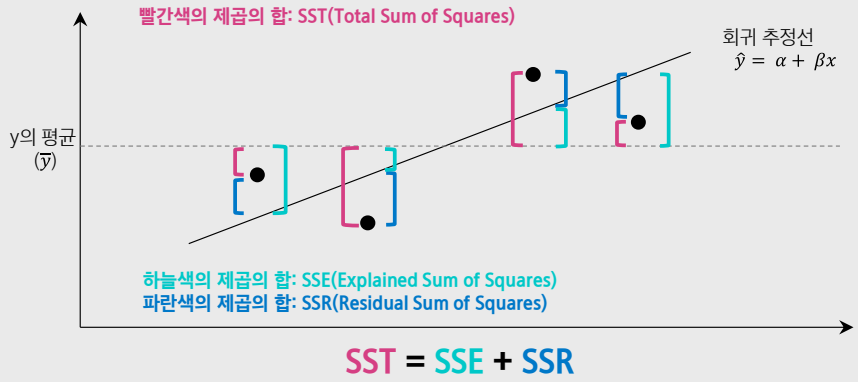
$$\textit{총제곱합}(SST, total\ sum\ of\ squares,) = \sum_{i=1}^{n}(y_i - \bar{y})^2$$

$$\textit{설명된 제곱합}(SSE, explained\ sum\ of\ squares) = \sum_{i=1}^{n}(\hat{y}_i - \bar{y})^2$$

$$\textit{잔차 제곱합}(SSR, residual\ sum\ of\ squares) = \sum_{i=1}^{n}(y_i - \hat{y}_i)^2 = \sum_{i=1}^{n}{\hat{u}_i}^2$$

여기서 $\bar{y}$는 y_i의 평균임. $\hat{y_i}$는 y_i의 회귀분석식을 통한 추정값임($\hat{y_i} = \hat{\beta_0} + \hat{\beta_1} x_i$)

셋 간에는 $SST = SSE + SSR$의 관계가 성립한다.



결정계수 R^2의 정의는 다음과 같다.

$$결정계수\ R^2 = y의\ 설명가능한\ 변동 / y의\ 총변동 = SSE/SST = 1 - SSR/SST$$

[단순 선형 회귀분석 상의 결정계수]

단순 선형 회귀분석모델에서 R^2와 r_{xy}의 관계를 본다면, 결정계수는 x와 y의 상관계수의 제곱과 같다.

(증명은 Appendix에 수록).

$$R^2 = {r_{xy}}^2$$

(또한, 결정계수는 y_i 실제값과 맞춘값 $\hat{y_i}$ 사이의 표본상관계수의 제곱과도 같다)

예를 들어 변동하는 y_i의 값이 선형 모델에 완벽히 부합하도록 움직인다면, 총제곱합(SST)와 설명된 제곱합(SSE)는 같은 값이 되고 결정계수는 1이 된다. 변동하는 y_i의 값이 선형 모델와 전혀 무관하게 움직인다면, SSE는 0이 되고 결정계수는 0이 된다. 결정계수는 0~1 사이의 값을 가지는데, 1에 가까울수록 선형 관계가 강하다는 뜻이며, 0에 가까울수록 선형 관계가 약하다는 뜻이다.

단순선형회귀분석모델에서 모델의 유의성은, 결정계수 혹은 상관계수로 판단할 수 있다.

III. 다중 회귀분석

복잡한 현실 세계에서, 두 변수의 선형 관계로만 세상을 설명하긴 어렵다. 여러 개의 독립변수들이 하나의 종속변수에 영향을 끼친다고 하면, 여러 변수의 선형 결합을 모형화하는 "다중 회귀분석 모델"이 유용할 수 있다.

[다중 회귀분석 모델]

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_k x_k + u$$

여기서, y는 종속변수임. $x_1, \cdots, x_k$는 독립변수들임. $\beta_0, \cdots, \beta_k$는 회귀계수임. u는 오차항임

위 다중 회귀분석 모델에서 β_0은 절편이다. i번째 회귀계수 β_i는 그외 다른 변수들이 고정되어 있을 때, x_i의 변화가 y에 미치는 효과(x_i가 한 단위 움직일 때 y가 움직이는 정도)를 의미한다.

다중 회귀분석 모델에서의 회귀계수도 OLS 방법을 사용하면 단순 회귀분석에서와 같이 공식 형태(estimator, 추정량)로 바로 계산할 수 있다. (행렬 표현식을 사용하면, 회귀계수 추정량의 공식을 간단히 표현 가능함. 증명은 Appendix에 수록.)

다중 회귀분석에서는 크게 두 가지를 주의해야 한다.

1) R^2는 [y 의 설명가능한 변동 / y 의 총변동] 수식을 통해 모델의 적합도를 나타내는 수치다. 그런데, 다중 회귀분석 모델에서는 독립변수를 추가할 때마다 항상 설명가능한 변동(SSE)은 커지고 잔차 변동(SSR)은 작아진다. 따라서 독립변수를 하나씩 추가할수록 R^2는 언제나 기계적으로 커진다. 즉, 독립변수들을 많이 넣을수록 과적합으로 인해 R^2가 좋아 보인다. 따라서, 다중 회귀분석에서는 독립변수 추가에 따른 패널티를 감안하기 위해서 수정결정계수라는 지표를 따로 본다.

$$\textit{수정결정계수 } \bar{R}^2 = 1 - [SSR/(n-k-1)]/[SST/(n-1)]$$

여기서, n 은 표본의 관측치수, k 는 독립변수의 수임

하나의 독립변수가 회귀에 추가되면 분자항의 SSR(잔차 제곱합)이 하락하지만, 분모항이 $(n-k-1)$도 하락한다. 추가 변수의 능력에 따라서 수정결정계수는 오를 수도, 내릴 수도 있다. 즉, 추가 변수의 설명력이 높을 때에만 수정결정계수가 상승한다.

2) 다중공선성의 문제를 피하는 것이 좋다. 우선 다중 회귀분석 모델의 가정에는 "독립변수 간에 완전한 공선성이 없다"는 조건이 들어있다. 만약 x_m이 x_n의 배수로 완벽히 표현된다면, 회귀분석 식에서 x_m기울기와 x_n기울기의 조합은 무한한 경우의 수가 나올 수 있다. 따라서 독립변수 간의 완전한 공선성이 없어야 한다.

또한 독립변수 간에 높은 공선성이 존재하면, OLS 기울기 추정량의 분산이 급격히 커지게 되어 모델의 견고성과 추정치의 신뢰성이 떨어진다. 실제로, 비슷비슷한 독립 변수들을 많이 넣은 회귀분석 모델에서는, 샘플 데이터가 조금씩만 업데이트되더라도 OLS 추정으로 나온 기울기 추정치가 그때 그때 크게 바뀌어 버리는 단점이 발생한다.

IV. 예제

매크로 변수의 주가 영향을 분석할 때, 가장 손쉽게 이용하는 것이 회귀분석 모델이다.

일례로, 올해 WTI 유가는 50% 가량 급등하면서 주식시장 영향이 큰 상황이다. 유류 제품을 판매하거나, 유류 소비가 큰 기업 들은 유가 변동에 따라서 실적 및 주가 또한 크게 변한다. 이에 투자자는 관심 종목의 주가가 유가 변동에 의해 얼마나 변화하는지를 회귀분석으로 추정할 수 있다.

WTI 유가의 월간 수익률을 독립변수, 개별 종목의 월간 주가수익률을 종속변수로 하는 회귀 분석 모델은 다음과 같다.

$$\left[A\ 종목의\ 월간\ 수익률\right] = \beta_0 + \beta_1 \left[WTI의\ 월간\ 수익률\right] + u$$

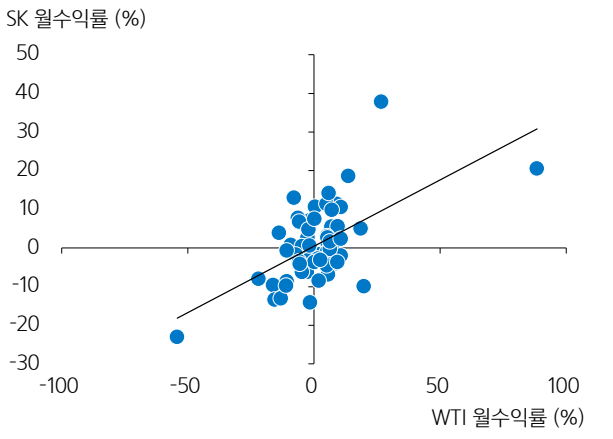
종속변수에 SK(A034730)의 월간 수익률을 넣고 샘플기간을 2016년~2020년의 월간자료 60개로 하면, 자료는 다음과 같다.

x, y 표본 데이터

월 (%)	y: SK 월수익률	x: WTI 월수익률
16-01-29	0.8	-9.2
16-02-29	10.7	0.4
16-03-31	18.6	13.6
	...	
20-10-30	-9.7	-11.0
20-11-30	37.8	26.7
20-12-31	9.8	7.0

자료: Bloomberg

산포도



자료: Bloomberg

회귀분석 모델에 대해 OLS 추정을 하면, β_1=0.34, β_0=0.49로 계산된다. 이에 따른 SK 주가 수익률과 유가 수익률의 회귀분석식은 다음처럼 나온다.

$$\left[SK의\ 월간\ 수익률\right] = 0.34 + 0.49 \left[WTI의\ 월간\ 수익률\right] + u$$

위 식의 의미는 WTI 월간 수익률이 1% 상승할 때, SK의 월간 수익률은 0.49% 상승한다는 뜻이다. 즉, SK는 WTI에 대해서 0.49의 베타를 가지는 유가 상승 수혜주임을 말한다.

당사의 전망처럼 유가의 중장기 상승 예상을 하고 있다면, SK와 같은 양(+)의 유가 베타를 가진 종목으로의 우선 투자를 고려해 볼만하다.

V. 사용법

1. 엑셀 사용법

엑셀에서는 다음의 함수를 통해서, 가장 기본이 되는 단순선형회귀분석모델을 편하게 쓸 수 있다.

`slope(known_y's, known_x's)`: 단순선형회귀분석모델의 기울기(β_1)를 반환
`intercept(known_y's, known_x's)`: 단순선형회귀분석모델의 절편(β_0)을 반환
`correl(array1, array2)`: 두 변수 간의 상관계수를 반환

또한, 2차원의 분산형 차트에서 "추세선 추가" 기능을 통해 회귀추세선, 회귀분석식, 결정계수 등을 바로 표시할 수 있다.

엑셀 상에서 다중회귀분석모델은 1) linest 배열함수 혹은 2) 데이터 분석 메뉴의 회귀 분석 기능을 통해서 사용할 수 있다.

linest 배열함수는 배열수식 형태 [ctrl+shift+enter]로 입력하는 함수로서, 데이터 분석 메뉴 활용법보다 핸들링이 간편하다.

`linest(known_y's, known_x's, const=TRUE, stats=FALSE)`:
다중회귀분석의 결과(기울기계수, 절편, 표준오차, F통계량 등)을 반환

`stats=FALSE`인 경우에는 1×(k+1) 배열만 필요함(여기서, k는 독립변수 개수임). 이 때는 한 행에 기울기계수와 절편만 반환함
stats=TRUE인 경우에는 5×(k+1) 배열이 필요함. 이 때는 첫 행에 기울기계수 및 절편, 둘째 행에 표준오차 등을 반환함

	A	B	C	D	E	F
1	m_n	m_{n-1}	. . .	m_2	m_1	b
2	se_n	se_{n-1}	. . .	se_2	se_1	se_b
3	r_2	se_y				
4	F	d_f				
5	ss_{reg}	ss_{resid}				

※ 엑셀 예제는 다음의 링크에서 다운로드가 가능합니다.

https://bit.ly/31POUnC

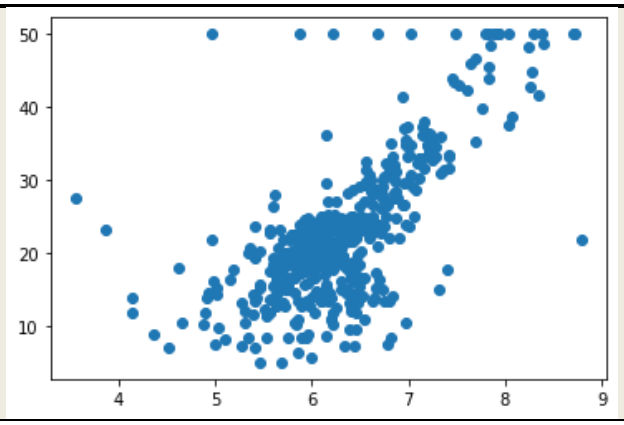
2. 파이썬(Python) 사용법

파이썬은 실무와 교육 모두에서 높은 활용도를 보이는 프로그래밍 언어로 현재 세계에서 가장 널리 쓰이는 프로그래밍 언어이다. PYPL Index에 따르면 파이썬은 2018년 6월을 기점으로 Java의 인기도를 역전하여 지속적으로 격차를 확대하며 독보적인 프로그래밍 언어의 일인자로 도약하였다. 파이썬은 다른 언어보다 작성과 해석이 쉽고 직관적일 뿐만 아니라 다양한 확장성을 보유해 개발자부터 입문자까지 모든 수준의 사용자에게 선호되고 있다.

파이썬은 오픈 소스로, 누구나 무료로 사이트에서 다운로드하여 사용할 수 있다. 특히, 아나콘다(Anaconda)를 설치하면 파이썬과 함께 머신러닝, 통계분석 등 데이터 분석에 필요한 패키지를 같이 설치할 수 있어 파이썬의 효율성을 극대화할 수 있다. 따라서, 퀀트 모델링을 포함한 다양한 데이터 분석에 파이썬을 활용하고자 하는 사용자라면 통합 패키지인 아나콘다(Anaconda)를 설치하여 파이썬을 사용하기를 추천한다.

아래 코드처럼 파이썬에서는 *plt.scatter* 메소드로 산포도를, *pd.corr* 메소드로 상관계수를 구할 수 있다.

```
# -*- coding: utf-8 -*-

# 모듈 가져오기   # 파이썬에서는 샵(#)기호로 주석을 작성
import pandas as pd          # pandas를 가져와 pd로 사용
import matplotlib.pyplot as plt    # matplotlib 내 pyplot을 가져와 plt로 사용
from sklearn.datasets import load_boston    #load_boston 가져오기
from sklearn.linear_model import LinearRegression   # LinearRegression 가져오기

# 보스턴 주택가격(파이썬 내재 데이터) 샘플 가져오기
# 13개 독립변수 샘플을 X에, 방개수 독립변수 샘플을 x_rm에, 종속변수 샘플을 y에 저장
boston_dataset = load_boston() # load_boston 객체를 boston_data 변수에 선언

"""
<변수 X>
Boston_dataset 객체의 data 키에 대한 array를 행렬로, feature_names 키 값 list를 열 이름으로 하는 DataFrame 객체를 생성. 생성된 DataFrame 객체를 X 변수에 선언
"""
X = pd.DataFrame(boston_dataset.data, columns=boston_dataset.feature_names)

x_rm = pd.DataFrame(X['RM'])   # X의 'RM' 열로 DataFrame을 만들어 x_rm 변수에 선언

"""
<변수 y>
Boston_dataset 객체의 target 키에 대한 array를 행렬로, 'MEDV'를 열 이름으로 하는 DataFrame을 생성하여 y 변수에 선언
"""
y = pd.DataFrame(boston_dataset.target, columns=[['MEDV']])

# x: 방개수(RM), y: 주택가격(MEDV)의 산포도 만들기
plt.scatter(x_rm, y)       #x_rm과 y로 산점도 생성
```




```
# x: 방개수(RM), y: 주택가격(MEDV)의 상관계수 구하기
pd.concat([y, x_rm], axis=1).corr()     # y와 x_rm의 상관계수 계산
```

```
          (MEDV,)        RM
(MEDV,)   1.00000   0.69536
RM        0.69536   1.00000
```

파이썬의 많은 모듈에서 회귀분석모델을 제공한다. *sklearn*의 *LinearRegression* 클래스가 대표적이다. 해당 클래스에서는 *fit, score, predict* 메소드와 *coef_, intercept_* 속성을 제공한다.

```
## 단순선형회귀분석

# 단순선형회귀분석 모델
lr = LinearRegression()     # LinearRegression 클래스를 생성하여 lr에 선언

# 단순선형회귀분석 실행
lr.fit(x_rm, y)             # x_rm(506×1)과 y(506×1)로 단순선형회귀분석 실행

# 순서대로 기울기계수, 절편, 결정계수 확인
print(lr.coef_)               # 기울기 계수 출력
print(lr.intercept_)          # 절편 출력
print(lr.score(x_rm, y))      # 결정계수 출력
```

```
[[9.10210898]]
[-34.67062078]
0.48352545599133423
```

```
# 단순선형회귀분석 예측 (임의의 x_rm_test 변수 준비)
lr.predict(x_rm_test)    # 변수 x_rm_test에 대한 회귀분석 모델(lr) 예측값 계산
```

```
## 다중회귀분석

# 다중회귀분석 모델
lr2 = LinearRegression()     # LinearRegression 클래스를 생성하여 lr2에 선언

# 다중회귀분석 실행
lr2.fit(X, y)                # X(506×13)와 y(506×1)로 다중회귀분석 실행

# 순서대로 기울기계수, 절편, 결정계수 확인
print(lr2.coef_)             # 기울기 계수 출력
print(lr2.intercept_)        # 절편 출력
print(lr2.score(X, y))       # 결정계수 출력
```

```
[[-1.08011358e-01  4.64204584e-02  2.05586264e-02  2.68673382e+00
  -1.77666112e+01  3.80986521e+00  6.92224640e-04 -1.47556685e+00
   3.06049479e-01 -1.23345939e-02 -9.52747232e-01  9.31168327e-03
  -5.24758378e-01]]
[36.45948839]
0.7406426641094095
```

```
# 다중회귀분석 예측 (임의의 X_test 변수 준비)
lr2.predict(X_test)  # 변수 X_rm_test에 대한 회귀분석 모델(lr2) 예측값 계산
```

VI. Appendix

[단순선형회귀분석모델 상의 $R^2 = {r_{xy}}^2$ 증명]

우선, $SSE = {\hat{\beta_1}}^2 \sum_{i=1}^{n}(x_i - \bar{x})^2$이다.

$$SSE = \sum_{i=1}^{n}(\hat{y_i} - \bar{y})^2$$

$$= \sum_{i=1}^{n}(\hat{\beta_0} + \hat{\beta_1}x_i - \bar{y})^2 \ (\because \hat{y_i} = \hat{\beta_0} + \hat{\beta_1}x_i)$$

$$= \sum_{i=1}^{n}(\bar{y} - \hat{\beta_1}\bar{x} + \hat{\beta_1}x_i - \bar{y})^2 \ (\because \hat{\beta_0} = \bar{y} - \hat{\beta_1}\bar{x})$$

$$= \sum_{i=1}^{n}(\hat{\beta_1}x_i - \hat{\beta_1}\bar{x})^2 = \sum_{i=1}^{n}\{\hat{\beta_1}(x_i - \bar{x})\}^2$$

$$= {\hat{\beta_1}}^2 \sum_{i=1}^{n}(x_i - \bar{x})^2$$

결정계수(R^2)의 정의에서 출발하면 다음과 같다.

$$R^2 = SSE/SST$$

$$= \frac{\sum_{i=1}^{n}(\hat{y_i} - \bar{y})^2}{\sum_{i=1}^{n}(y_i - \bar{y})^2}$$

$$= \frac{{\hat{\beta_1}}^2 \sum_{i=1}^{n}(x_i - \bar{x})^2}{\sum_{i=1}^{n}(y_i - \bar{y})^2} \ (\because SSE = {\hat{\beta_1}}^2 \sum_{i=1}^{n}(x_i - \bar{x})^2)$$

$$= \frac{[\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})]^2}{[\sum_{i=1}^{n}(x_i - \bar{x})^2]^2} \frac{\sum_{i=1}^{n}(x_i - \bar{x})^2}{\sum_{i=1}^{n}(y_i - \bar{y})^2} \ (\because \hat{\beta_1} = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^{n}(x_i - \bar{x})^2})$$

$$= \frac{[\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})]^2}{\sum_{i=1}^{n}(x_i - \bar{x})^2 \sum_{i=1}^{n}(y_i - \bar{y})^2}$$

$$= \left[\frac{\sum_{i=1}^{n}(x_i - \bar{x})\,(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2}\sqrt{\sum_{i=1}^{n}(y_i - \bar{y})^2}}\right]^2$$

$$= {r_{xy}}^2$$

즉, $R^2 = {r_{xy}}^2$

[다중 회귀분석의 행렬 접근법]

y변수를 하나의 x변수로 설명하는 단순회귀분석 모형은 다음과 같이 표시된다.

$$y_i = \beta_0 + \beta_1 x_i + u_i \quad i = 1,2,3, \dots, n$$

독립변수가 k개인 다중회귀분석 모형은 다음과 같이 표시된다.

$$y_i = \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + \cdots + + \beta_k x_{ki} + u_i \quad i = 1,2,3, \dots, n$$

여기서, β_0는 절편, $\beta_1 \sim \beta_k$는 기울기모수, u_i는 오차항, i 는 i 번째 관측치, n은 표본의 크기다.

이 식에서 n개의 관찰치를 모두 넣은 연립방정식은 다음과 같다.

$$y_1 = \beta_0 + \beta_1 x_{11} + \beta_2 x_{21} + \cdots + \beta_k x_{k1} + u_1$$

$$y_2 = \beta_0 + \beta_1 x_{12} + \beta_2 x_{22} + \cdots + \beta_k x_{k2} + u_2$$

$$\dots \dots \dots$$

$$y_n = \beta_0 + \beta_1 x_{1n} + \beta_2 x_{2n} + \cdots + \beta_k x_{kn} + u_n$$

이의 행렬 형태 표시는 다음과 같이 할 수 있다.

$$\begin{bmatrix} y_1 \\ \vdots \\ y_n \end{bmatrix} = \begin{bmatrix} \beta_0 \\ \vdots \\ \beta_0 \end{bmatrix} + \begin{bmatrix} x_{11} & \cdots & x_{k1} \\ \vdots & \ddots & \vdots \\ x_{1n} & \cdots & x_{kn} \end{bmatrix} \begin{bmatrix} \beta_1 \\ \vdots \\ \beta_k \end{bmatrix} + \begin{bmatrix} u_1 \\ \vdots \\ u_n \end{bmatrix} \quad \text{(A)}$$

$$\underset{(n \times 1)}{\boldsymbol{y}} = \underset{(n \times 1)}{\boldsymbol{\beta_0}} + \underset{(n \times k)}{\boldsymbol{X}} \underset{(k \times 1)}{\boldsymbol{\beta}} + \underset{(n \times 1)}{\boldsymbol{u}}$$

식 (A)가 다중회귀분석을 행렬로 표현한 식이다.

여기에서 절편에 해당하는 $\boldsymbol{\beta_0}$를 $\boldsymbol{X}$ 부분에 합칠 수 있다. 이를 적용하면 다음과 같다.

$$\begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{bmatrix} = \begin{bmatrix} 1 & x_{11} & x_{21} & \dots & x_{k1} \\ 1 & x_{12} & x_{22} & \cdots & x_{k2} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 1 & x_{1n} & x_{2n} & \cdots & x_{kn} \end{bmatrix} \begin{bmatrix} \beta_0 \\ \beta_1 \\ \vdots \\ \beta_k \end{bmatrix} + \begin{bmatrix} u_1 \\ u_2 \\ \vdots \\ u_n \end{bmatrix} \quad \text{(B)}$$

$$\underset{(n \times 1)}{\boldsymbol{y}} = \underset{(n \times (k+1))}{\boldsymbol{X}} \underset{((k+1) \times 1)}{\boldsymbol{\beta}} + \underset{(n \times 1)}{\boldsymbol{u}}$$

식 (B)가 일반적인 다중회귀분석 모형의 행렬 표현식이다. 이 때 $\boldsymbol{y}$와 $\boldsymbol{X}$가 주어진 데이터이며 이를 가지고 회귀분석을 통해서 $\boldsymbol{\beta}$를 추정하게 된다.

[다중 회귀분석의 OLS 추정법 증명]

일반적인 다중회귀분석의 행렬 표현식은 다음과 같다.

$$\begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{bmatrix} = \begin{bmatrix} 1 & x_{11} & x_{21} & \cdots & x_{k1} \\ 1 & x_{12} & x_{22} & \cdots & x_{k2} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 1 & x_{1n} & x_{2n} & \cdots & x_{kn} \end{bmatrix} \begin{bmatrix} \beta_0 \\ \beta_1 \\ \vdots \\ \beta_k \end{bmatrix} + \begin{bmatrix} u_1 \\ u_2 \\ \vdots \\ u_n \end{bmatrix} \quad \text{(B)}$$

$$\underset{(n \times 1)}{\boldsymbol{y}} = \underset{(n \times (k+1))}{\boldsymbol{X}} \underset{((k+1) \times 1)}{\boldsymbol{\beta}} + \underset{(n \times 1)}{\boldsymbol{u}}$$

$\boldsymbol{\beta}$의 추정에 사용하는, 보통최소제곱법(최소자승법, ordinary least squares)은 잔차의 제곱의 합을 최소화하는 방식이다.

$$\hat{\boldsymbol{u}} = \boldsymbol{y} - \boldsymbol{X}\hat{\boldsymbol{\beta}}$$

잔차 제곱 합은 다음과 같이 쓸 수 있다.

$$\sum \hat{u}_i{}^2 = \hat{\boldsymbol{u}}'\hat{\boldsymbol{u}} = (\boldsymbol{y} - \boldsymbol{X}\hat{\boldsymbol{\beta}})'(\boldsymbol{y} - \boldsymbol{X}\hat{\boldsymbol{\beta}})$$

$$= (\boldsymbol{y}' - (\boldsymbol{X}\hat{\boldsymbol{\beta}})')(\boldsymbol{y} - \boldsymbol{X}\hat{\boldsymbol{\beta}}) = (\boldsymbol{y}' - \hat{\boldsymbol{\beta}}'\boldsymbol{X}')(\boldsymbol{y} - \boldsymbol{X}\hat{\boldsymbol{\beta}})$$

$$= \mathbf{y}'\mathbf{y} - \mathbf{y}'\boldsymbol{X}\hat{\boldsymbol{\beta}} - \hat{\boldsymbol{\beta}}'\boldsymbol{X}'\boldsymbol{y} - \hat{\boldsymbol{\beta}}'\boldsymbol{X}'\boldsymbol{X}\hat{\boldsymbol{\beta}}$$

여기서, $\hat{\boldsymbol{\beta}}'\boldsymbol{X}'\boldsymbol{y}$는 스칼라이므로, $\hat{\boldsymbol{\beta}}'\boldsymbol{X}'\boldsymbol{y}$와 $\mathbf{y}'\boldsymbol{X}\hat{\boldsymbol{\beta}} = (\hat{\boldsymbol{\beta}}'\boldsymbol{X}'\boldsymbol{y})'$는 동일하게 바꿔 쓸 수 있다.

$$\hat{\boldsymbol{u}}'\hat{\boldsymbol{u}} = \mathbf{y}'\mathbf{y} - 2\hat{\boldsymbol{\beta}}'\boldsymbol{X}'\boldsymbol{y} - \hat{\boldsymbol{\beta}}'\boldsymbol{X}'\boldsymbol{X}\hat{\boldsymbol{\beta}}$$

잔차 제곱 합이 최소가 되기 위해서는 $\hat{\boldsymbol{u}}'\hat{\boldsymbol{u}}$를 $\hat{\boldsymbol{\beta}}$로 미분한 값이 0이 되어야 한다.

$\hat{\boldsymbol{u}}'\hat{\boldsymbol{u}}$의 미분값을 보자.

$$\frac{\partial(\hat{\boldsymbol{u}}'\hat{\boldsymbol{u}})}{\partial\hat{\boldsymbol{\beta}}} = \frac{\partial(\mathbf{y}'\mathbf{y} - 2\hat{\boldsymbol{\beta}}'\boldsymbol{X}'\boldsymbol{y} - \hat{\boldsymbol{\beta}}'\boldsymbol{X}'\boldsymbol{X}\hat{\boldsymbol{\beta}})}{\partial\hat{\boldsymbol{\beta}}}$$

여기서, 임의의 행렬 $\boldsymbol{A}$, 열벡터 $\boldsymbol{x}$가 있을 때, $\frac{\partial(\boldsymbol{x}'\boldsymbol{A}\boldsymbol{x})}{\partial\boldsymbol{x}} = (\boldsymbol{A} + \boldsymbol{A}')\boldsymbol{x}$ 이다(행렬의 정의에 따라 유도 가능).

그리고, $\boldsymbol{A}$가 대칭행렬($\boldsymbol{A} = \boldsymbol{A}'$)일 경우에는, $\frac{\partial(\boldsymbol{x}'\boldsymbol{A}\boldsymbol{x})}{\partial\boldsymbol{x}} = 2\boldsymbol{A}\boldsymbol{x}$ 로 쓸 수 있다.

위의 미분식에서 $\boldsymbol{X}'\boldsymbol{X}$은 대칭행렬이므로, $\frac{\partial(\hat{\boldsymbol{\beta}}'\boldsymbol{X}'\boldsymbol{X}\hat{\boldsymbol{\beta}})}{\partial\hat{\boldsymbol{\beta}}} = 2\boldsymbol{X}'\boldsymbol{X}\hat{\boldsymbol{\beta}}$ 로 쓸 수 있다. 따라서 위의 미분식은 다음과 같다.

$$\frac{\partial(\hat{\boldsymbol{u}}'\hat{\boldsymbol{u}})}{\partial\hat{\boldsymbol{\beta}}} = 0 - 2\boldsymbol{X}'\boldsymbol{y} - 2\boldsymbol{X}'\boldsymbol{X}\hat{\boldsymbol{\beta}}$$

위 방정식을 0이라고 놓으면,

$$\boldsymbol{X}'\boldsymbol{X}\hat{\boldsymbol{\beta}} = \boldsymbol{X}'\boldsymbol{y}$$

역행렬이 존재한다는 전제 하에, $\boldsymbol{\beta}$의 OLS 추정량은 다음과 같다.

$$\hat{\boldsymbol{\beta}} = (\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{y} \quad \text{(C)}$$

〈참고문헌〉

Wooldridge(2013), "계량경제학 I", 박영사

Gujarati and Porter(2009), "Gujarati의 계량경제학", 지필미디어

Compliance notice

- 보고서는 철저히 계량적 분석에 근거한 의견을 제시합니다. 따라서 당사의 대표 투자의견과 다를 수 있습니다.

신뢰에 가치로 답하다

삼성증권 SAMSUNG

삼성증권주식회사
서울특별시 서초구 서초대로74길 11(삼성전자빌딩)
Tel: 02 2020 8000 / www.samsungpop.com
삼성증권 지점 대표번호: 1588 2323 / 1544 1544
고객 불편사항 접수: 080 911 0900