Ridge, LASSO, Elastic Net 회귀를 이용한 전기동 전망

2018.07.20

박소윤 연구원
문의 : 02-565-3492
soyoon@koreapds.com

SUMMARY

원자재 가격을 예측하는 다양한 통계적 방법론이 존재하나 모든 시계열을 예측하기에는 한계가 존재한다. 따라서 Korea PDS에서는 계속적으로 변화하는 시장상황을 예측하고자, 정규화 선형 회귀 모형을 통해 예측을 시도해 보았다. 계수의 크기를 제한하여 모형 계수의 크기가 과도하게 증가하지 않도록 정규화 하는 모델이다. 이는 Ridge, LASSO, Elastic Net 회귀 모형이 있다.

Lasso와 Elastic Net 모형은 시계열 모형인 Arima, Winters 모형의 성능에 견주어도 비슷한 수준인 우수한 예측력을 나타냈다.

목차

정규화 선형 회귀 방법 Regularized Regression Model

정규화 선형 회귀 모형

모형 계수 크기의 과도한 증가를 막기 위해 제약 조건을 추가하는 방법론임. Ridge, LASSO, Elastic Net 회귀 모형이 있음.

정규화(regularized) 선형 회귀 방법은 일반적으로 회귀 계수(weight)에 대한 제약 조건을 추가함으로써 모형이 과도하게 최적화되는 현상을 막는 방법론이다.

모형이 과도하게 최적화되면 모형 계수의 크기도 과도하게 증가하는 경향이 나타나기 때문이다. 따라서 일반적으로 계수의 크기를 제한하여 정규화 하는 모델을 사용하기도 하며 이에는 Ridge, LASSO, Elastic Net 회귀 모형이 있다.

1) 능형 회귀 Ridge Regression

기존의 모형 추정 방법론들의 **회귀계수를 과잉 추정하는 문제를 피하기 위해서 회귀계수의 평균 쪽으로 줄어들게 고안된 방법론**이 능형회귀이다. LASSO처럼 회귀계수가 0으로 수렴되지는 않으며, 상관성이 있는 변수들에 대해서 적절한 가중치 배분을 통해 축소 시킨다. Ridge 회귀 모형에서는 **가중치들의 제곱합(squared sum of weights)을 최소화**하는 것을 제약 조건으로 한다. 능형회귀는 라그랑즈 승수법(Lagrange multiplier)에 의해서 다음과 같이 표현할 수 있다.

$$\hat{\beta}^{ridge} = min_{\beta}\left\{\sum_{i=1}^{n}(y_i - \beta_0 - \sum_{j=1}^{p} x_{i,j}\beta_j)^2 + \lambda\sum_{j=1}^{p}{\beta_j}^2\right\} = RSS + \lambda\sum_{j=1}^{p}{\beta_j}^2$$

이때, λ는 제약 조건의 비중을 조절하는 수축 패널티이다($\lambda \geq 0$). λ가 크면 정규화 정도가 커지고 가중치의 값들이 작아져 계수 추정치가 작아진다. 이를 통해 얻어진 추정치는 최소제곱의 추정치보다 축소되어 **분산(variance)을 줄이면서 추정된 모형의 예측오차를 줄일 수 있다는 장점**이 있다.

2) LASSO (Least Absolute Shrinkage and Selection Operator) 회귀

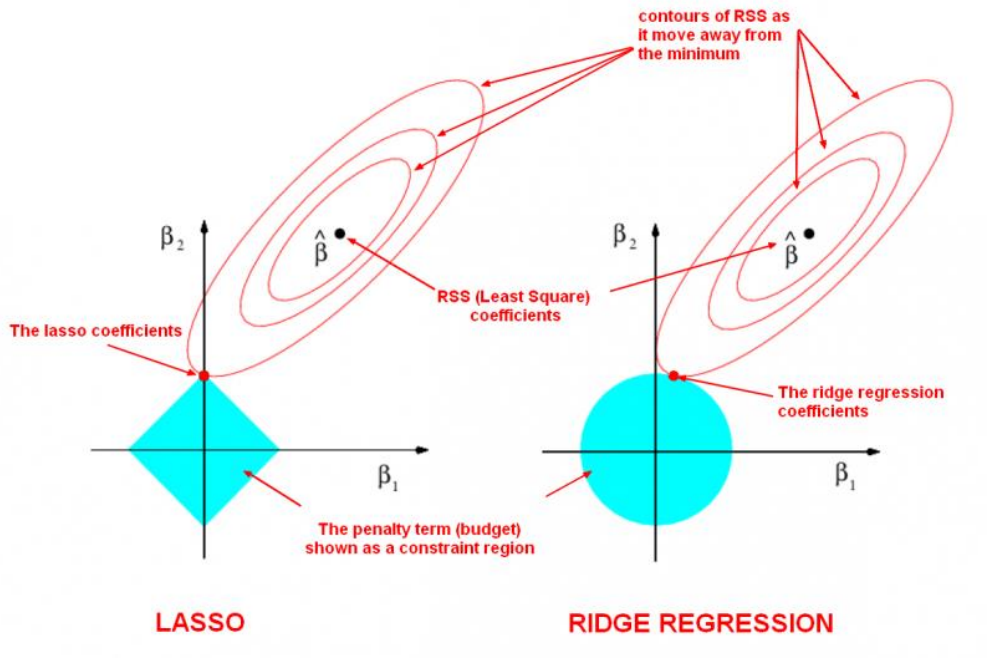
회귀분석에서 회귀계수를 추정할 때 일반적으로 사용되는 최소자승법의 단축된 형태인 LASSO는영향력이 적은 회귀계수 값을 0으로 수축(Shrink)시킨다. 능형회귀와 다르게 **축소와 변수선택을 동시에 실행**함으로써 **예측력을 향상**시키고, 적은 수의 변수만을 선택하기 때문에 **추정된 모형에 대한 해석력이 높다.** Lasso 회귀 모형은 **가중치의 절대값의 합을 최소화**하는 것을 제약 조건으로 한다.

$$\hat{\beta}^{lasso} = min_{\beta}\left\{\sum_{i=1}^{n}(y_i - \beta_0 - \sum_{j=1}^{p} x_{i,j}\beta_j)^2 + \lambda\sum_{j=1}^{p}|\beta_j|\right\} = RSS + \lambda\sum_{j=1}^{p}|\beta_j|$$

3) 신축망 회귀 Elastic Net Regression

Elastic Net 회귀 모형은 **가중치의 절대값의 합과 제곱합을 동시에 제약 조건으로 가지는 모형**이다. Trevor (2005)가 2005년에 **Ridge와 LASSO를 동시에 이용하는 방법**을 고안한 것으로 수식은 아래와 같다. λ_1 , λ_2 두 개의 하이퍼 모수를 가지고 있고, 이를 조정하여 Ridge와 LASSO를 구현 가능하다.

$$\hat{\beta} = \text{argmin}\left\{\sum_{i=0}^{p}(Y_i - \alpha - \beta_1 X_{i1} - \cdots - \beta_n X_{in})^2 + \lambda_1\right\}$$



[그림1] LASSO, Ridge Regression 그림

출처: Robert Tibshirani(1996)

Ridge 모형과 LASSO 모형의 비교

위의 수식과 같이 능형회귀는 계수의 제곱, Lasso는 절대값을 사용한다. 또한, 그림을 통해 쉽게 비교할 수 있는데, 능형회귀(우)는 모든 설명 변수 p를 포함한다. 이것은 수축페널티가 크다고 해도 각각의 설명변수의 계수가 0에 가까워지지 0은 아니기 때문이다(수축페널티가 무한대일 경우는 제외).

반면, Lasso(좌)는 일부 변수만 포함하고 나머지 변수의 계수는 0이 된다. 이는 수축 페널티가 클수록 일부 변수의 계수는 0이 되기 때문이다. **즉, Ridge 모형은 가중치 계수를 한꺼번에 축소시키는데 반해 Lasso 모형은 일부 가중치 계수가 먼저 0으로 수렴하는 특성이 있다.**

변수 선정 단계

[독립 변수 선정 단계]

사용변수: 주요변수 55개 중 28개 품목
데이터 기간: 2005.01~2017.12
선정 방법: 2^n개의 회귀 모델 생성 후 비교를 통한 변수 선택

[종속변수]
런던금속거래소(LME) 전기동 3개월 Official Price

[선정된 독립변수]
오른쪽 [표1]의 총 8개 변수

분석에 사용할 독립 변수를 선정하기 위해 변수 선정 단계를 추가하여 통계적인 방법론에 의거한 변수 선정을 시행하였다. 변수 선정 분석에 사용한 변수는 Korea PDS 애널리스트가 선정한 주요 변수 28개 품목이며, 데이터 기간은 2005.01 ~ 2017.12 사이의 월 평균 값이다.

종속 변수로 런던금속거래소(LME) 전기동 3개월 Official Price 품목의 월 평균가를 사용한다.

1) 변수 선정

일반적으로 적절한 회귀 모델을 찾기 위해 단계적으로 변수를 추가, 삭제하거나 이를 반복하는 방법을 사용한다. 해당 분석에서는 이와 달리 N개의 독립변수가 있을 때, **총 2^n개의 회귀 모델을 만들고 이들 모든 경우에 결과 비교를 통해 변수를 선택 하였다. 결과 비교의 기준은 AIC(Akaike information criterion), BIC와 결정계수(또는 수정 결정계수)이다.** 종속변수를 제외한 27개 변수를 위의 방법으로 변수 선택하여 **총 8개 변수를 선정**하였다.

[표1] 다중회귀의 모든 경우 비교 변수 선택 결과표

변수	변수명
x6	달러/유로 환율
x13	LME+NYMEX+SHFE 전기동 재고량
x21	전기동 SHFE
x22	전기동 CME(NYMEX)
x23	칠레페소/미달러(CLP/USD) 환율
x25	독일 IFO 기업체감지수
x26	Euro 산업생산
x28	위안화/달러

2) 변수 표준화

설명변수 간의 Scale 차이가 크기 때문에 데이터 표준화 작업이 필요하다. 자료 x의 최대값 max(x)와 최소값 min(x)를 이용하여 z를 도출하는 표준화 방식을 이용하였다. 표준편차가 1인 설명변수의 표준화 값 z는 0과 1사이의 값을 가진다.

$$z = \frac{x - \min(x)}{\max(x) - \min(x)}$$

모델링

[분석 기간]

-Train: 2005.01~2011.12까지 학습,
1개월씩 가격 추가해 ~2014.12까지 학습,

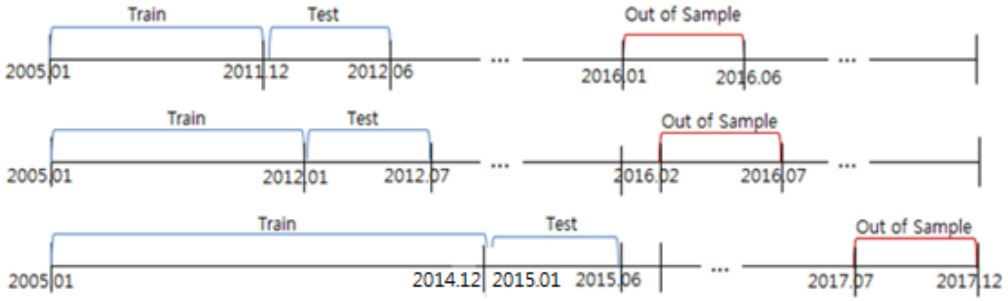
-Test: 2012.01~ 2012.06 6개월씩 예측,
가격 추가하며 최종 2015.12까지 롤링테스트,

-표본 외 예측(Out of Sample Test):
2016.01~2016.06로 6개월씩 예측,
가격 추가하며 최종 2017.12까지 예측 후 롤링테스트.

1) 분석 기간

2005년 1월부터 2011년 12월까지의 데이터로 초기모델을 적용하여 훈련(Train)하였고, 2012년 1월부터 2012년 6월까지 테스트(Test)하였다. 이러한 방식으로 1개월씩 증가되는 가격을 추가하며 2005년 1월부터 최종 2014년 12월까지 훈련하였고, 동시에 2012년 1월부터 2015년 12월까지 6개월씩마다 테스트 하였다.

또한, 표본 외 예측(Out Of Sample Forecasting)을 위해 2016년 1월부터 2012년 6월까지의 6개월간 예측하며 같은 방식으로 최종 2017년 12월까지 2년여간 전망하였다(2005.01 ~2015.06까지의 학습 데이터 사용).



[그림2] 시작 시점을 고정한 Rolling Window Test 방식 (In Sample(파란색), Out of sample(빨간색))

2) 예측 모형 설정

예측(종속)변수는 전기동 3개월(Official) 월 평균 가격이다. 설명변수는 [표1]의 8개 품목이다.

[모형별 모수 선정]

[Alpha 값]
Ridge 회귀는 alpha = 0,
Lasso 회귀는 alpha = 1,
Elastic Net 회귀는
0 < alpha < 1 사이의 파라미터

[Lambda 값]
수축량을 정의하는 파라미터,
예측 오차율을 최소화 하는 모수

보다 효과적인 예측을 위해 **Ridge, Lasso, Elastic Net 회귀 모형**을 사용해 예측을 시도하였다. 이를 비교해보며 각각 모형의 예측력을 확인 해보고 해당 분석 시기에 가장 좋은 모형을 찾아보도록 한다.

Ridge 회귀의 alpha 값은 0이고, Lasso 회귀는 alpha 값은 1, Elastic Net 회귀는 0 과 1 사이의 파라미터이다. 해당 분석에서는 alpha = 0.7으로 설정하였다. Lambda 값은 계수 축소를 조정하기 위한 하이퍼 파라미터(Hyper Parameters)이다. 이들은 교차 검증을 통해, 예측 오차율을 최소화하는 최적의 모수(Best tuning parameter)를 선정하고 모형을 생성하였다.

교차검증(cross validation)이란 특정데이터를 훈련(Train)전용, 테스트(Test)전용으로 분할하여 훈련용 데이터를 활용해 학습하고, 테스트 데이터로 검증하는 방법이다. 모델의 타당성 검증 방법 중 하나로 사용된다. 해당 분석에서는 10회 간 교차검증 하였고, 오차율이 가장 적게 나타나게 되는 때의 Lambda 값을 각각의 모델에 적용하였다.

3) 모형 예측력 평가 방법

일반적으로 학습(Train), 검증(Test)하는 예측모형에서 사용되는 랜덤 방식이 아닌 연속 방식을 사용하였다. 이는 특정 시점 예측치(t)가 다음 학습에 포함되어 다음 시점 예측값에 영향을 주기 때문에 더 작은 MAPE를 산출하는 모형이 구축되기 때문이다.

연속 방식은 데이터의 시작시점을 고정하고 시계열의 t시점까지의 데이터를 이용하여 모형을 추정한 후, 모형 추정에 사용하지 않은 t+1시점의 가격을 예측하는 방법을 사용하여 비교하였다.

평가방법으로는 시계열 분석에서 일반적으로 사용하는 예측평가 방법인 **평균 절대 백분율 오차 Mean Absolute Percent Error (MAPE)를 이용하여 오차율을 비교** 하였다. 추정식은 아래와 같다.

$$\mathrm{M} = \frac{100}{n}\sum_{t=1}^{n}\left|\frac{A_t - F_t}{A_t}\right|$$

신축망 회귀 모형 전망 결과

모델별 전망치에 대해 MAPE(%) 오차율을 In Sample(Train에 따른 Test결과)과 Out of Sample(표본 외 예측)으로 나누어 검정하였다.

1) 기간에 따른 모형별 오차율 결과 – In Sample

[표2] In Sample(Train에 따른 Test 오차)에 대한 모형별 MAPE(%) 오차율 결과표

MAPE(%)	Ridge	Lasso	Elastic Net	Winters	Arima
1 개월	1.60%	1.32%	1.20%	2.33%	1.34%
1~3 개월	4.04%	3.47%	3.55%	4.60%	3.61%
1~6 개월	8.08%	7.63%	7.35%	7.96%	7.05%

1개월, 1~3개월, 1~6개월의 예측 기간에 따른 모형별 오차율(MAPE)을 비교하였다. 먼저, 모형별 비교 시 Ridge, Lasso, Elastic Net 모형의 1개월 오차율이 1%대로 매우 작았다. 특히 Elastic Net 모형의 1, 1~6개월의 오차율이 가장 적게 나타났다. Lasso와 Elastic Net 모형은 시계열 모형인 Arima, Winters 모형의 성능에 견주어도 비슷한 수준의 예측력을 나타냈다.

2) 모형별 오차율 결과 – Out of Sample

[표3] Out of Sample에 대한모형별 MAPE(%) 오차율 결과표

MAPE(%)	Ridge	Lasso	Elastic Net	Winters	Arima
1 개월	6.11%	5.55%	3.14%	1.97%	1.41%
1~3 개월	10.94%	7.72%	7.22%	7.15%	6.60%
1~6 개월	19.56%	17.36%	17.45%	14.88%	14.35%

표본 외 예측(아웃샘플) 역시 Elastic Net 회귀 모형의 MAPE오차율이 가장 적었다. 자기 자신의 과거 값을 예측에 이용하는 시계열 모형인 Arima와 Winters의 성능이 좋기 때문에 Ridge, Lasso, Elastic Net 모형의 결과를 함께 비교하기에 어려움이 있다. 그럼에도 불구하고 Elastic Net 회귀 모형의 1개월, 1~3개월 오차율이 각각 3%, 7%대를 나타내며 비교적 좋은 예측력을 보였다.

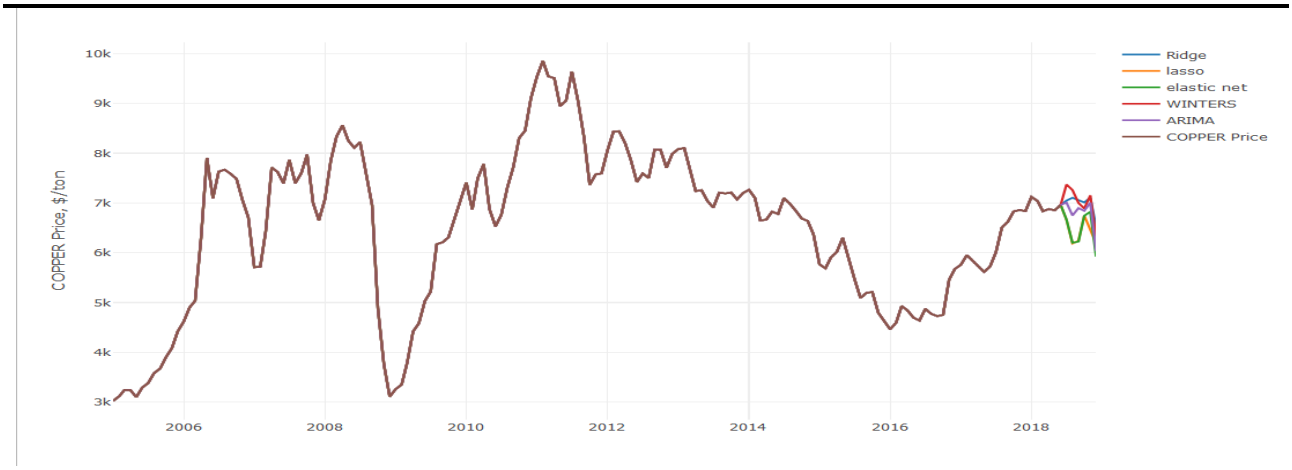
해당 분석에서의 표본 외 예측(Out of sample)은 2005.01 ~2014.12까지 학습시킨 데이터를 통해 2016.01 ~2017.12까지를 2년간 전망하였다. 랜덤 방식이 아닌 연속(순차적) 방식으로 훈련했기 때문에, 공백이 있는 기간만큼 예측에 어려움이 있을 수 있다.

3) 추후 전망

앞서 언급된 정규화 모형(Ridge, LASSO, Elastic Net)의 분석 결과를 통해 예측 성능을 확인하였다. 그러므로 2018년 7월 이후 6개월 동안의 가격을 예측해 보았다.

예측 결과 Elastic, Lasso, Arima 모형은 7월 이후 하락하는 모습을 나타냈고, Winter, Ridge모형은 8월까지 상승하다 하락하는 것으로 나타났다. 전반적인 모형이 18년 하반기에는 하락세의 방향으로 예측되며, 톤당6천 달러 초중반까지 내려갈 것으로 나타났다.

[그림 3] 2018년 이후 전기동 가격 전망 그래프



참고 www.koreapds.com

4) 결론

전기동 가격을 예측하기 위해 회귀 계수에 대해 제약 조건을 추가함으로써 모형이 과도하게 최적화되는 현상을 막는 정규화 선형 회귀를 이용하였다.

모형이 과도하게 최적화되면 모형 계수의 크기도 증가하는 경향이 나타나기 때문이다. 따라서 계수의 크기를 제한하여 정규화 하는 모델을 사용하기도 하며 이에는 Ridge, LASSO, Elastic Net 회귀 모형이 있다.

Ridge, LASSO, Elastic Net 세 가지 회귀 모형에 대해 알아보고, 모형의 결과를 비교하여 어떠한 모형이 가장 잘 예측 하는지를 확인해 보았다.

LASSO와 Elastic Net이 해당 예측 기간 동안 낮은 예측 오차를 나타내며 좋은 예측력을 보였다. 시계열 모형인 Arima, Winters 모형의 성능에 견주어도 비슷한 수준의 예측력을 나타냈다.

※ 별 첨

[표4] 변수 선정 단계에서 선정된 독립변수

KoreaPDS 전기동 주요변수(선정 전)	예측에 사용한 독립변수(선정 후)
28 개	9 개
[전기동](3 개월(o) LME) – 종속변수	[전기동](3 개월(o) LME) – 종속변수
중국 정련 전기동 수입량	달러/유로 환율
전세계 전기동 소비량	LME+NYMEX+SHFE 전기동 재고량
전세계 전기동 생산량	전기동 SHFE
NYMEX 전기동 투기적순매수건수	전기동 CME(NYMEX)
달러/유로 환율	칠레페소/미달러(CLP/USD) 환율
WTI NYMEX 선물	독일 IFO 기업체감지수
중국 경기선행지수	Euro 산업생산
OECD 경기선행지수	위안화/달러
미국 산업생산	
유로 산업생산	
일본 산업생산	
LME+NYMEX+SHFE 전기동 재고량	
미국 신규주택착공건수	
미국 건축허가	
LME 정련 전기동 재고량	
NYMEX 정련 전기동 재고량	
SHFE 정련 전기동 재고량	
칠레 전기동 생산량	
전기동 SHFE	
전기동 CME(NYMEX)	
칠레페소/미달러(CLP/USD)	
칠레 전기동 정광 생산량	
페루 구리 광산 생산량	
독일 IFO 기업체감지수	
Euro 산업생산	
미국 전기동 소비량	
위안화/달러	

작성 : 코리아PDS 박소윤 (soyoon@koreapds.com) 연구원