<예측 애널리틱스 과제>

2014170852 조영관

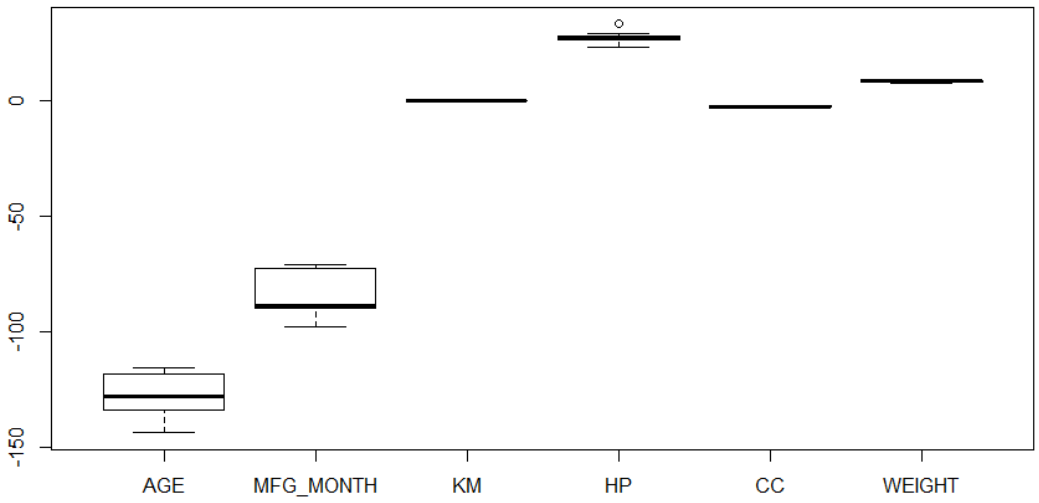
<training set을 변형시켜가며 LASSO 파라미터의 변동을 확인하기>

사용한 DATA는 'TOYOTA COROLLA' DATA를 사용하였습니다.

TARGET VALUE값 Y는 PRICE(가격)이며 나머지 입력변수 X에는 AGE, MFG_MONTH, KM, HP, CC, WEIGHT 가 있습니다. 이 데이터에서 TRAINING SET을 다양하게 추출하여 LASSO 파라미터의 경건함을 확인해보겠습니다.

	AGE	MFG_MONTH	KM	HP	CC	WEIGHT
1	-133.726	-89.967	-0.0082	26.83	-2.105	7.740
2	-121.022	-72.049	-0.0197	23.37	-2.366	8.679
3	-135.768	-73.504	-0.0106	33.14	-2.946	8.982
4	-116.792	-89.198	-0.0194	26.71	-2.435	8.130
5	-128.017	-71.756	-0.0100	25.46	-2.627	8.310
6	-144.926	-91.661	-0.0080	29.17	-2.713	9.249
7	-116.825	-97.917	-0.0125	26.98	-2.825	8.500

수치 결과값은 다음과 같습니다. Training set을 7번 변경시켜서 돌려본 각 변수의 베타 값 결과입니다. 이 것을 box plot으로 표현해보겠습니다.



위의 box plot을 보면 베타 값의 분산이 크지 않고 training set의 변화에 큰 차이 없이 robust 함을 알 수 있습니다.

<입력변수 간 상관관계가 크고 작은 정도에 따른 LASSO PARAMETER의 ROBUST 유무 확인>

먼저 입력변수 간 상관관계가 큰 경우를 확인해 보겠습니다.

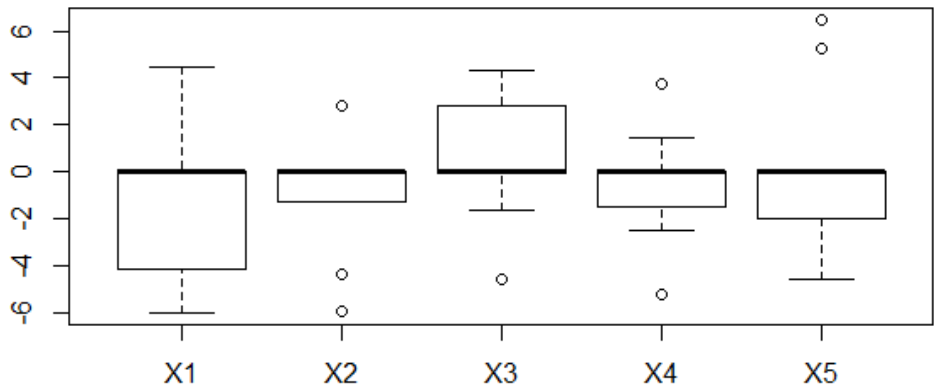
상관관계가 큰 경우의 DATA는 그대로 TOYOTA COROLLA의 데이터를 이용하되, 이 중 입력변수는 KM만 남기고, TARGET VALUE인 Y값을 남기고 나머지는 소거하였습니다.

그리고 상관관계를 높이기 위하여 KM와 비슷한 경향성을 띄도록 난수 생성을 활용하여 KM와 큰 상관관계를 갖도록 새로운 입력변수 X2, X3, X4, X5를 생성하였습니다.

그 후 LASSO PARAMETER가 어떻게 변하는지를 확인해보았습니다.

	X1	X2	X3	X4	X5
1	0.0000	-0.0560	0.0000	0.0000	0.0000
2	4.4586	0.0000	3.6242	-1.5036	5.2350
3	-6.0032	0.0000	2.7836	0.0000	0.0000
4	0.0000	-4.3658	0.0000	-5.2350	-2.0144
5	0.0000	0.0000	-0.0520	-2.5220	0.0000
6	0.0000	2.8330	-1.6120	0.0000	-4.6200
7	-4.9530	-1.3060	0.0000	1.4144	6.4684
8	-4.1778	-5.9490	4.3220	0.0000	0.0000
9	0.0000	0.0000	-4.6250	3.7213	-4.1842

수치 결과값은 다음과 같습니다. BOX PLOT에 도시해보겠습니다.



그 결과 위와 같이 나왔습니다.

X2와 X4는 비교적 작은 분산을 보인다고 볼 수도 있지만 X1, X3, X5는 변동성이 큼을 확인할 수 있었습니다. 즉, LASSO PARAMETER는 변수 간 상관관계가 큰 경우에 유의미한 영향을 받고, ROBUST 하지 않음을 알 수 있었습니다.

다음으로, 변수 간 상관관계가 작은 경우를 확인해보았습니다.

상관관계가 작은 데이터 역시 TOYOTA COROLLA 데이터를 그대로 활용하였습니다.

이 중 상관관계가 거의 없는 입력변수들이 어떤 것인지 COR 함수를 이용해 확인해보았습니다.

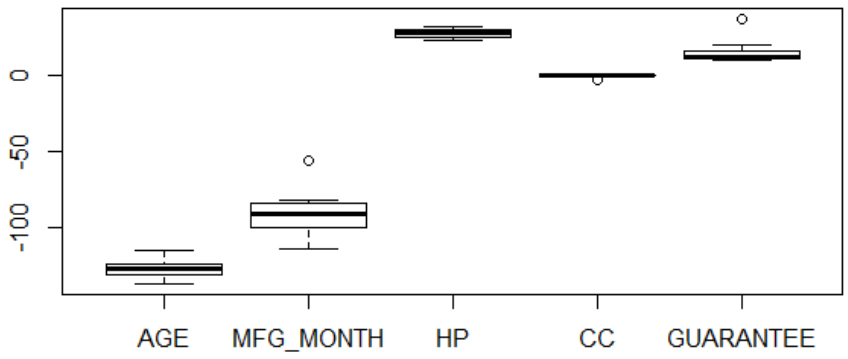
그 결과 AGE, MFG_MONTH, HP, CC, GUARANTEE 이렇게 다섯 가지가 서로 상관관계가 거의 없음(독립)을 확인할 수 있었습니다.

따라서 이 변수들을 이용해 LASSO PARAMETER가 ROBUST 한지를 파악해보았습니다.

그 결과 다음과 같았습니다.

	AGE	MFG_MONTH	HP	CC	GUARANTEE
1	-115.260	-91.330	28.3600	-2.483	37.194
2	-123.536	-56.190	25.4800	0.000	12.979
3	-136.740	-113.830	32.3300	0.075	11.220
4	-133.186	-107.930	25.2040	0.060	20.110
5	-127.528	-81.890	23.3940	0.000	12.107
6	-124.000	-90.527	30.2160	-0.006	10.504
7	-128.880	-86.702	30.5796	-0.420	10.703

수치 결과값은 위와 같습니다. BOX PLOT에 도시해보겠습니다.



위 결과를 살펴보면, 다섯 개의 변수들 모두 변동성이 거의 없음을 확인할 수 있습니다.

즉, 입력변수 간에 독립성이 있을 때 LASSO PARAMETER는 ROBUST함을 알 수 있습니다.