Multivariate Data Analysis

Assignment #1

고려대학교 공과대학

산업경영공학부

2017170857 이우준

[Q1] 본인이 스스로 Multivariate Linear Regression 모델을 적용하는데 적합한 데이터셋을 선정하고 이유를 설명하시오. 데이터셋 탐색은 아래에서 제시된 Data Repository를 포함하여 여러 Repository를 검색해서 결정하시오. 보고서에는 데이터를 다운로드할 수 있는 링크를 반드시 제공하시오.

[A] UCI Repository에 올라와 있는 Seoul Bike Sharing Demand Data Set을 이용하였습니다. 자전거를 빌린 횟수라는 명확한 종속변수와 이를 뒷받침하는 설명변수의 종류가 총 12가지가 있어 해당 데이터 셋을 선정하였습니다. 또한 정성적으로 볼 때 강우량이 많으면 자전거를 빌린 횟수가 적어진다는 선형관계가 생각되어 선정하였습니다.

https://archive.ics.uci.edu/ml/datasets/Seoul+Bike+Sharing+Demand

[Q2] 해당 데이터셋의 종속변수와 설명변수는 어떤 것들이 있는가? 분석 전에 아래 세 가지 질문에 대해서 스스로 생각해보고 답변을 하시오.

1. 이 데이터는 종속변수와 설명변수들 사이에 실제로 "선형 관계관계"가 있다고 가정할 수 있겠는가? 가정할 수 있음/없음 판단에 대한 본인의 생각을 서술하시오.

   [A] 가정할 수 있습니다. 1번 질문에서 답한바와 같이 정성적으로 생각하였을 때 강우량, 혹은 강설량과 자전거를 탈 수 있는 여부는 명확히 상관관계가 있고, 고로 자전거를 빌린 횟수는 강우량 강설량과 음의 선형관계를 가질 것이라고 생각합니다.

2. 이 데이터에서 제공된 설명변수들 중에서 높은 상관관계가 있을 것으로 예상되는 변수들은 어떤 것들이 있는가? 왜 그렇게 생각하는가?

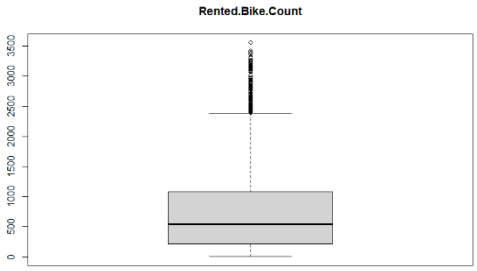
   [A] Holiday 변수와, Temperature 변수가 높은 상관관계가 있을 것이라고 생각합니다. 공휴일날 유동인구가 늘어, 다른 날 보다 자전거를 빌리는 인원이 많을 것이라고 생각됩니다. 또 Temperature 변수는 날씨가 영하로 내려가면 자전거를 타기 좋은 날씨가 아니어서 자전거를 빌리는 인원이 줄을 것이라고 생각되어 두 변수가 높은 상관관계가 있을 것이라고 생각합니다.

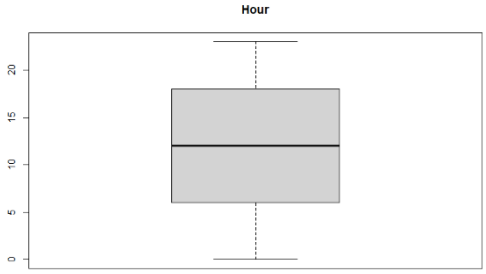
3. 제공된 설명변수들 중에서 종속변수를 예측하는데 필요하지 않을 것으로 예상되는 변수들은 어떤 것들이 있는가? 왜 그렇게 생각하는가?

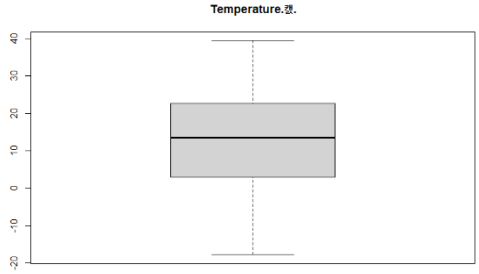
   [A] 날짜 데이터와, Humidity변수가 종속변수를 예측하는데 필요하지 않을 것으로 예상됩니다. 날짜 데이터는 다른 데이터 셋에서의 ID와 같다고 생각합니다, 또 다른 변수들과 달리 Humidity 혼자로는 종속변수를 설명하기에는 비약이 있다고 생각합니다. 불쾌지수와 종속변수와는 선형관계가 있을 수 있어도, 온도와 습도의 관계로 설정되는 불쾌지수와 달리 습도 혼자로는 종속변수를 예측하기가 힘들 것 같습니다.

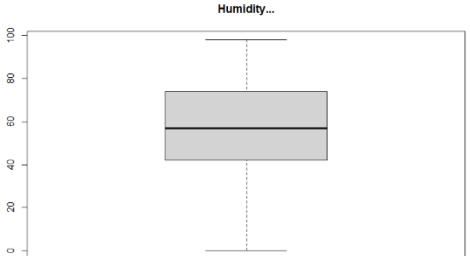
[Q3] 개별 입력 변수들에 대하여 각각 다음과 같은 단변량 통계량을 계산하고 Box plot을 도시하시오: Mean, Standard deviation, Skewness, Kurtosis. 전체 변수 중에서 정규분포를 따른다고 할 수 있는 변수들은 몇 개인가? 정규분포를 따른다고 가정한 근거는 무엇인가?

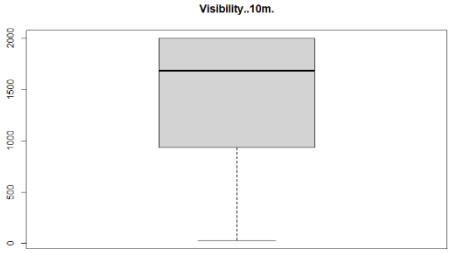
	mean	std	skw	kurt
Rented.Bike.Count	7.291570e+02	642.3511661	1.139296177	0.8182096
Hour	1.150703e+01	6.9208992	-0.001260101	-1.2038689
Temperature.캜.	1.277106e+01	12.1043748	-0.174487963	-0.8902596
Humidity...	5.814719e+01	20.4848387	0.068624651	-0.8129740
Wind.speed..m.s.	1.725883e+00	1.0342812	0.893904614	0.7522147
Visibility..10m.	1.433873e+03	609.0512294	-0.695059770	-0.9697413
Dew.point.temperature.캜.	3.944997e+00	13.2423990	-0.338654700	-0.8195068
Solar.Radiation..MJ.m2.	5.678677e-01	0.8682452	1.509529921	1.1427378
Rainfall.mm.	1.491199e-01	1.1255400	14.611741447	289.7226535
Snowfall..cm.	7.768458e-02	0.4440633	8.289891509	90.4661110
Holiday	4.819846e-02	0.2141980	4.218042351	15.7937471
Winter	2.551683e-01	0.4359816	1.122995852	-0.7389676
Spring	2.551683e-01	0.4359816	1.122995852	-0.7389676
Summer	2.608387e-01	0.4391181	1.089151272	-0.8138456
Autumn	2.288246e-01	0.4201009	1.290849016	-0.3337482

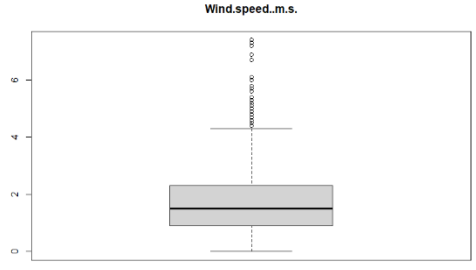


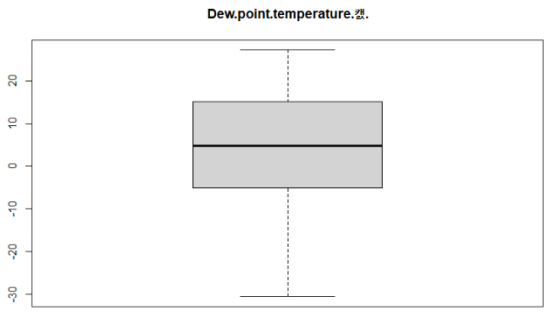


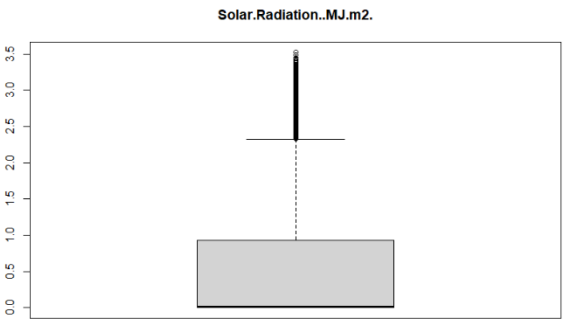


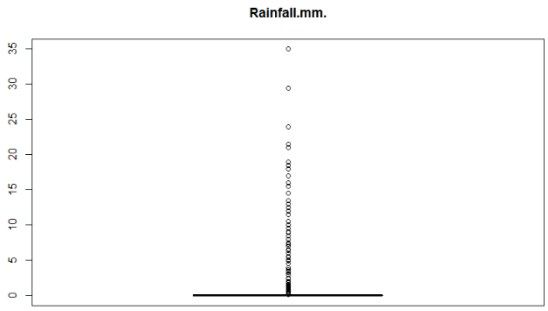


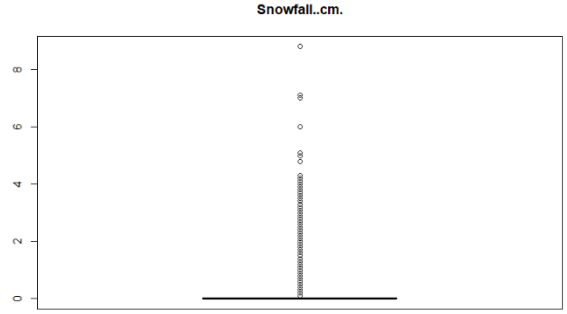


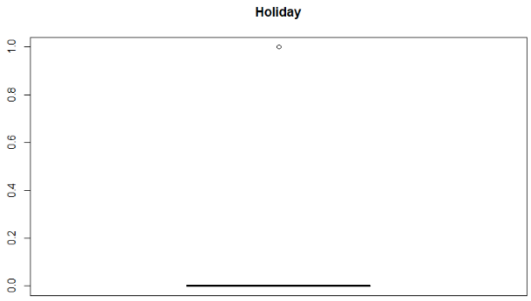

Wind.speed..m.s.
0
2
4
6

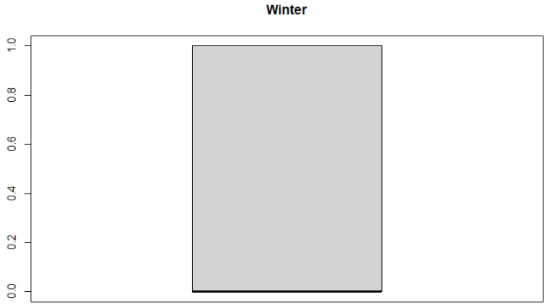

Dew.point.temperature.캘.
20
10
0
-10
-20
-30

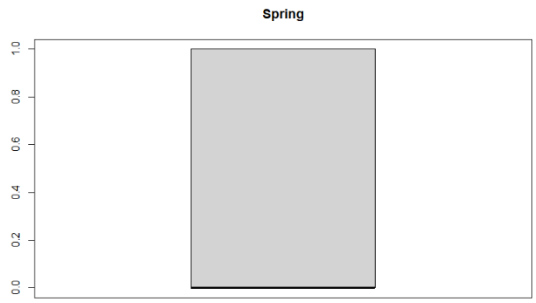

Solar.Radiation..MJ.m2.
0.0
0.5
1.0
1.5
2.0
2.5
3.0
3.5

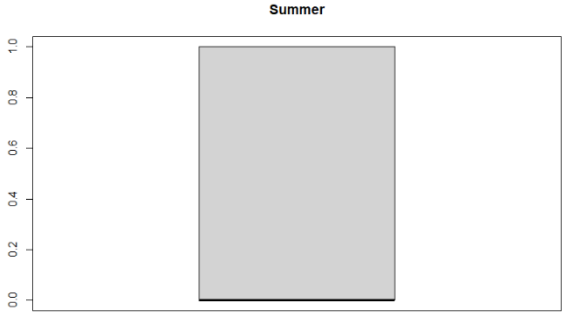

Rainfall.mm.
0
5
10
15
20
25
30
35

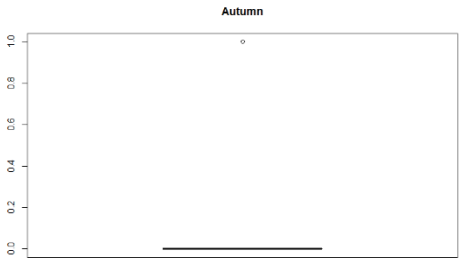

Snowfall..cm.
0
2
4
6
8


Holiday
0.0
0.2
0.4
0.6
0.8
1.0


Winter
0.0
0.2
0.4
0.6
0.8
1.0


Spring
0.0
0.2
0.4
0.6
0.8
1.0


Summer
0.0
0.2
0.4
0.6
0.8
1.0


Autumn
0.0
0.2
0.4
0.6
0.8
1.0

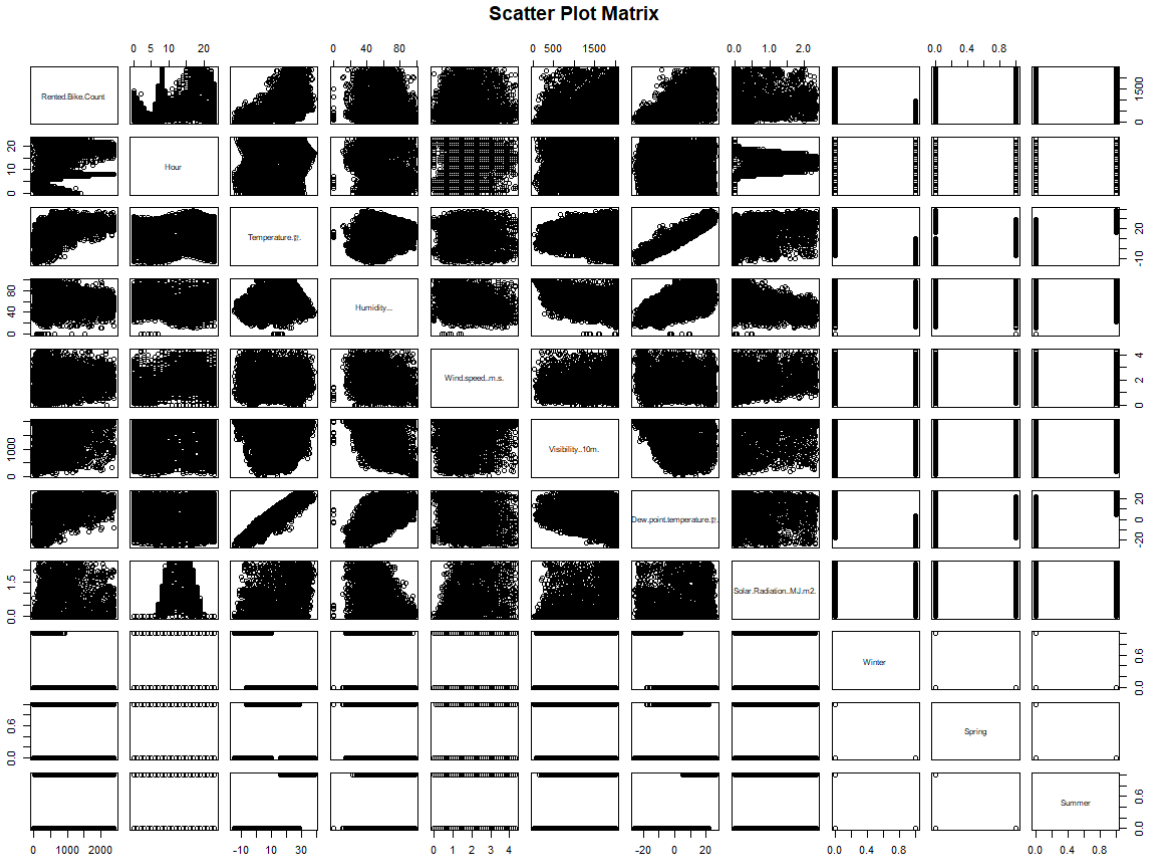
[A] 전체 변수 중 정규분포를 따른다고 할 수 있는 변수는 Rainfall, Snowfall, Holiday 세 변수를 제외한 모든 변수가 정규분포를 따른다고 할 수 있습니다. 왜도와 첨도의 값이 [-2,2] 사이에 존재하면 정규분포를 따른다고 가정 할 수 있는데 이를 만족시키지 못하는 변수는 위 세 변수 밖에 없습니다.

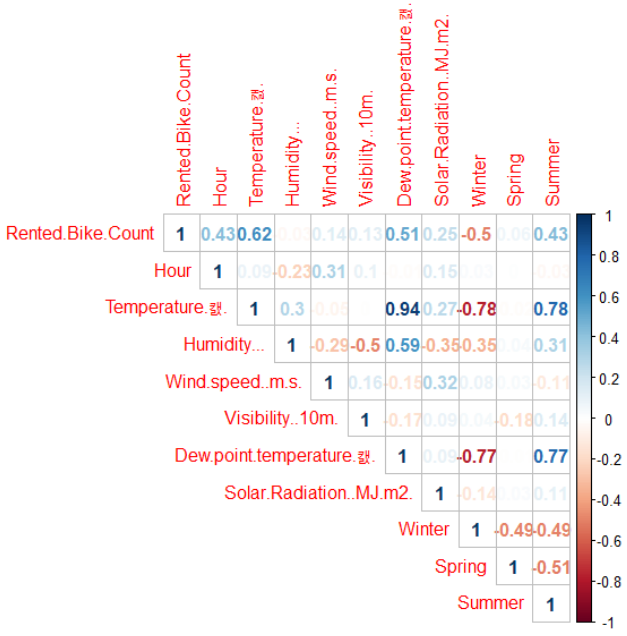
[Q4] [Q3]의 Box plot을 근거로 각 변수들에 대한 이상치(너무 값이 크거나 작은 객체) 조건을 정의하고, 해당하는 객체들을 데이터셋에서 제거해 보시오.

[A] Box plot 을 바탕으로 각 변수들의 이상치 조건을 확인해보면 아래표와 같이 나오는 것을 확인할 수 있습니다.

	LCL	UCL
Rented.Bike.Count	2.0	2387.00
Hour	0.0	23.00
Temperature.캘.	-17.8	39.40
Humidity...	0.0	98.00
Wind.speed..m.s.	0.0	4.30
Visibility..10m.	27.0	2000.00
Dew.point.temperature.캘.	-30.6	27.20
Solar.Radiation..MJ.m2.	0.0	2.32
Rainfall.mm.	0.0	0.00
Snowfall..cm.	0.0	0.00
Holiday	0.0	0.00
Winter	0.0	1.00
Spring	0.0	1.00
Summer	0.0	1.00
Autumn	0.0	0.00

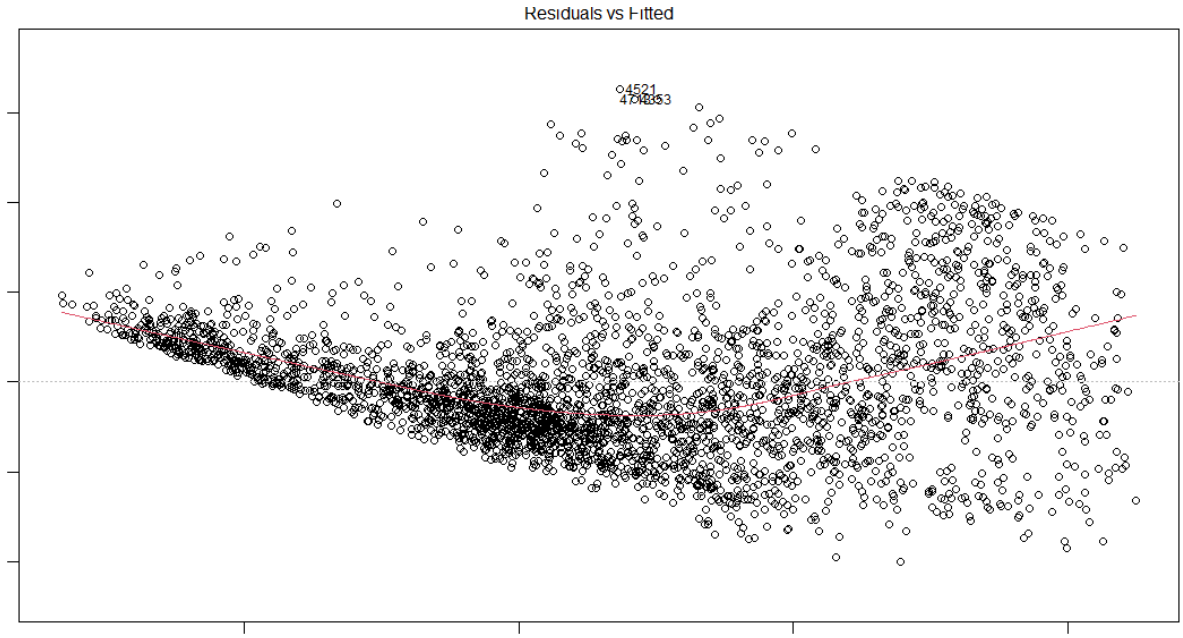
[Q5] 가능한 모든 두 쌍의 입력 변수 조합에 대한 산점도(scatter plot) 및 Correlation plot (hint: "corrplot" 패키지의 corrplot() 함수 사용) 상관관계를 계산해 보시오. 어떤 두 조합의 변수들이 서로 강한 상관관계가 있다고 할 수 있는가?

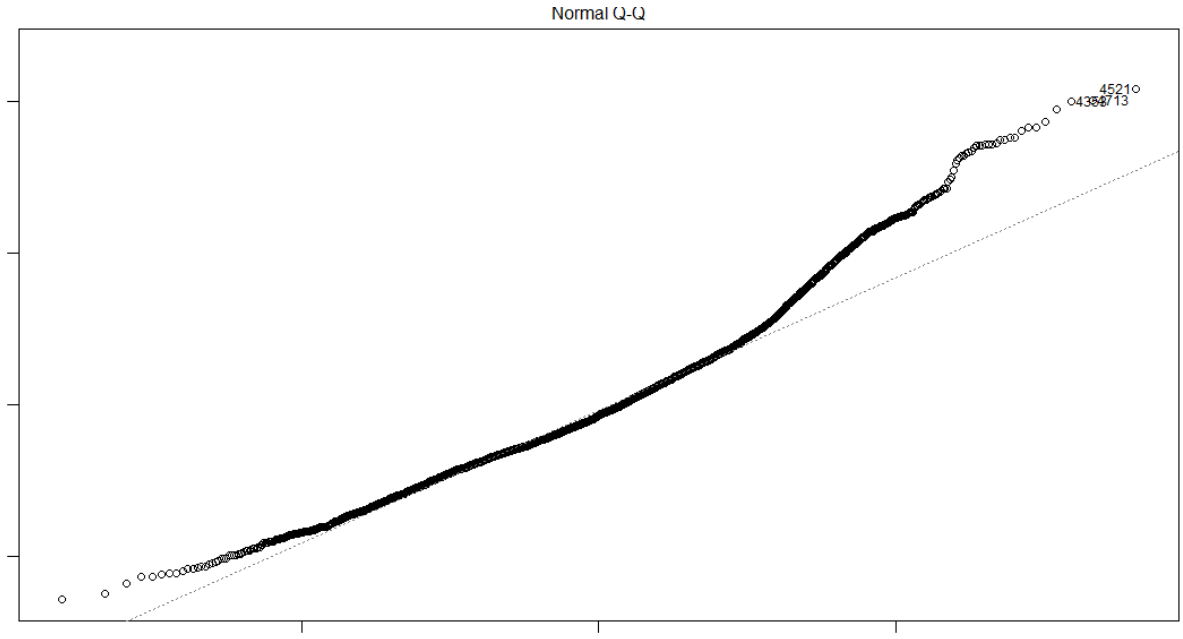




[A] Scatter Plot과 Correlation Plot을 그려보았을 때 Temperature 와 Dew point temperature, Temperature와 Winter, Temperature와 Summer, Dew point temperature와 Winter, Dew point temperature와 Summer가 강한 상관관계가 있다고 할 수 있습니다.

[Q6] 전체 데이터셋을 70%의 학습 데이터와 30%의 테스트 데이터로 분할한 후 모든 변수를 사용하여 MLR 모델을 학습해 보시오. Adjusted R2값을 통해 데이터의 선형성(linearity)을 판단해 보시오. Residual plot과 Q-Q Plot을 도시하고 Ordinary Least Square 방식의 Solution이 만족해야 하는 가정들이 만족될 만한 수준인지 정성적으로 판단해 보시오.





[A] MLR 모델 학습결과 Adjusted R Square값은 0.5571로 약한 선형 상관관계를 띄는 것을 볼 수 있습니다. Residual Plot을 보았을 때 개형이 미세하게 곡선형 그리는 것을 보아 해당 패턴을 설명할 수 있는 고차원 적인 변수가 부족하다고 생각합니다. QQ plot을 보았을 땐 우측으로 진행하였을 때 45도 선에서 벗어나는 것을 보아 정규성이 지켜지지 않았다고 판단할 수 있습니다.

[Q7] 유의수준 0.01에서 모형 구축에 통계적으로 유의미한 변수들은 어떤 것들이 있는가? 해당 변수들은 종속변수와 양/음 중에서 어떤 상관관계를 갖고 있는가?

[A] 총 4가지 변수가 유의수준 0.01 안에서 통계적으로 유의미한 변수입니다. 먼저 Hour, Temperature 이렇게 두가지 변수는 양의 상관관계를 갖고 있습니다, 그와 반대로 Humidity, Winter 변수는 음의 상관관계를 가지고 있습니다.

[Q8] Test 데이터셋에 대하여 MAE, MAPE, RMSE를 계산하고 그에 대한 해석을 해 보시오.

[A] Test 데이터 셋에 대하여 구한 MAE MAPE RMSE는 다음 표와 같습니다.

	RMSE	MAE	MAPE
Seoulbike	407.4276	312.0547	115.6216

MAE는 평균 오차의 절대값의 평균을 뜻합니다. 평균적으로 312대 정도 차이나는 것을 알 수 있습니다. MAPE는 평균오차의 절대값의 비율오차로 다른 모델과 정확도를 비교할 때 쓸 수 있는 수치로 115.6216이 나왔습니다. RMSE는 오차의 제곱의 합의 평균에 루트를 씌운 값입니다. 이를 통해 하루 평균 407도 정도 모델과 실제 값이 차이가 나다는 것을 알 수 있습니다.

[Q9] 만약 원래 변수 수의 절반 이하로 입력 변수를 사용하여 모델을 구축해야 할 경우 어떤 변수들을 선택하겠는가? [Q5]와 [Q7]의 답변을 바탕으로 본인이 선택한 변수들에 대한 근거를 제시하시오.

[A] Hour, Temperature, Humidity, 이 세가지 변수만을 선택하여 모델을 구축하겠습니다. 먼저 이 세 변수는 유의수준 0.01안에서 통계적으로 유의미한 변수입니다. Winter변수는 유의수준 0.01안에서 통계적으로 유의미한 변수이지만 Temperature변수와 상관관계를 보이므로 다중공산성이 일어나는 것을 방지하기 위하여 선택하지 않겠습니다.

[Q10] [Q9]에서 선택한 변수들 만을 사용하여 MLR 모형을 다시 학습하고 Adjusted R2, Test 데이터셋에 대한 MAE, MAPE, RMSE를 산출한 뒤, 두 모형(모든 변수 사용 vs. 선택된 변수만 사용)을 비교해 보시오.

[A] Q9에서 선택한 변수들을 이용해 만든 MLR 모델의 Adjusted R2의 값은 0.5421로 변수를 줄이기 전 모델의 Adjuster R2 Square 값과 비슷한 수치를 보였습니다. 테스트 데이터 셋이 대한 MAE, MAPE, RMSE는 아래의 표와 같으며 RMSE, MAE, MAPE 값 모두 유의미하게 줄어들었습니다. 이는 오차율이 더 줄어든 것이므로 Q9에서 변수를 줄인 모델이 이 전 모델보다 더 간단하며 오차율이 적은 나은 모델이라는 것을 뜻합니다.

	RMSE	MAE	MAPE
Seoulbike	395.0764	300.804	98.82387

[Extra Question] 이 외 해당 데이터셋을 통해 MLR 관점에서 가능한 추가적인 분석을 웹에서 검색해서 수행하고 그 결과를 해석해 보시오.

[A] R에서 지원하는 Package 중 하나인 Leap 패키지를 이용하여 변수 선택법 중 BIC 기준으로 최적의 변수들로 이루어진 다중 선형회귀 모델을 선택해보겠습니다. BIC란 Bayesian Information Criterion 모델로 잔차 제곱합과 효과 수를 증가시키는 함수입니다. 반응 변수와 효과 수에 대한 설명되지 않은 변동은 BIC 값의 증가로 이어집니다. 결과적으로, BIC 값이 낮으면 설명 변수 또는 적합 항목이 적거나, 아니면 둘 다 적다는 것을 의미합니다.[1]

총 가능한 9가지의 모델 중 BIC 값이 -2697.753로 가장 낮은 모델은 Hour, Temperature, Humidity 이 세가지 변수로 이루어진 모델입니다. 이는 Q9에서 선정하고 Q10에서 검증한 모델로써 주어진 데이터로 만들 수 있는 다중 선형 회귀 모델 중 가장 최적의 모델이라는 것을 증명합니다.

[1] SAS Support 참조http://support.sas.com/documentation/cdl_alternate/ko/vaug/68027/HTML/default/n1nvrl0wrzsp5hn1i549sro1kiyy.html