2021-1 YDMS 5주차 과제

〈Subject : Regression〉

김하은 2019251034

✓preview : 제 6장. 다중 선형 회귀분석 예습

회귀분석이란 변수가 다른 변수에 영향을 받을 때, 그 변수들 간의 함수관계를 규명하기 위해 이용되는 통계적 방법이다. 회귀분석의 목적은 '변수들 간의 관계를 함수로 표현하여 한 변수의 값으로부터 다른 변수의 값을 추정하고 예측'하는 것이다. 한편 회귀분석은 독립변수의 개수 혹은 회귀계수의 형태에 따라 나뉘는데, 이는 다음과 같다.

<기준 ①. 모형에 포함된 독립변수 개수>

▸ 단순 회귀분석 : 종속변수의 변동을 1개의 독립변수의 관계식으로 설명

▸ 다중 회귀분석 : 종속변수의 변동을 2개 이상의 독립변수의 관계식으로 설명

<기준②. 회귀계수의 형태> ; ★ 회귀계수의 형태가 선형인지 아닌지.

▸ 선형 회귀분석 : $y_i = \beta_0 + \beta_1 x_i + \beta_2 x_i^2 + \cdots + \beta_k x_i^k + \epsilon_i$

▸ 비선형 회귀분석 : $y_i = \beta_0 + \frac{\beta_1 x}{\beta_2 + x} + \epsilon_i$

전통적인 통계학에서 추론을 목적으로 회귀분석을 진행한 것과 달리 데이터 마이닝에서의 회귀분석은 예측 및 결과 해석을 목적으로 진행하게 된다. 이때 예측 및 결과 해석을 위해 가장 많이 사용되는 모델이 바로 '다중 선형 회귀모델'이다. 위의 설명을 토대로 생각해보면, 다중 선형 회귀모델은 종속변수의 변동을 2개 이상의 독립변수의 관계식으로 설명할 때 회귀계수의 형태가 선형인 회귀모델임을 알 수 있다. 따라서 다중 선형 회귀모델에서 예측변수와 종속변수 사이의 관계는 다음과 같다.

▸ 다중 선형 회귀모델 : $Y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_p x_p + \epsilon$

모델에 주어진 데이터는 회귀계수를 추정하고 잡음(ϵ)을 정량화하는데 사용되며, 모델의 성능을 평가할 때도 활용된다. 회귀모델링은 회귀계수를 추정하는 것뿐만 아니라 어떤 예측변수를 선택하여 어떤 형태로 진행할 지를 선택하는 것이다. 이때 모델링은 인과관계가 뚜렷하여 평균관계를 포착하는 설명모델링과 인과관계가 확실하지 않아 입력과 출력 간의 연관성을 정량화하는 기술모델링이 있다. 예측분석의 경우 새로운 개별 관측치에 대한 예측에 중점이 맞춰져 있기 때문에 기술모델링에 더 가깝다. 따라서 예측모델에서는 회귀계수나 평균 레코드를 다루지 않고, 생성된 모델을 활용하여 새로운 레코드에 대한 예측에 더 중점을 둔다. 따라서 좋은 설명모델은 데이터를 잘 적합시키는 모델이지만, 좋은 예측모델은 새로운 관측치에 대해 더 정확하게 예측하는 모델이다. 예측모델은 예측의 정확도를 높이기 위해 데이터를 학습세트와 검증세트로 나누어 진행한다. 이때 학습세트는 모델을 추정하는 데에 사용되고, 검증세트는 모델의 예측성능을 평가하는 데에 사용된다. 이러한 이유로 예측모델은 과대적합 방지를 위해 모델링에 사용한 데이터를 비교적 덜 적합한다.

한편 데이터마이닝 회귀분석에서 예측변수(x)로 선택할 수 있는 변수의 수가 많은 경우 문제가 발생한다. 우선 예측변수가 많을수록 데이터에 결측값이 존재할 가능성이 커진다. 무엇보다 예측변수가 많은 경우 '다중공산성'으로 인해 회귀계수의 추정치들이 불안정할 수 있다. (회귀계수는 간결한 모델에서 더욱 안정적이다.) 또한 종속변수와 상관관계가 없는 변수를 사용하면 예측의 분산이 커져 과대적합이 우

려되고, 반대로 종속변수와 상관관계가 있는 변수를 누락하면 예측의 오차나 편향이 커져 과소적합이 우려된다. 이러한 이유로 예측변수는 종속변수와 상관관계가 큰 변수들로만 선택하여 개수를 줄여 나가야한다.

한편 예측변수의 수를 줄이기 위해선 먼저 그 분야의 지식을 활용하는 것이 있다. 여러 예측변수들이 무엇을 측정하고 있는지, 왜 이 변수들이 종속변수 예측에 적절한지를 아는 것이 중요하다. 이때 요약통계량과 그래프, 빈도와 상관관계 테이블, 예측변수 중심의 요약통계량과 산점도, 결측값의 개수 등이 유용하다.

다음으로는 계산력과 통계적 유의성을 이용하는 것이 있다. 이는 크게 전역탐색과 부분집합 선택 알고리즘으로 나누어진다. 먼저 전역탐색은 모든 예측변수들의 부분집합을 평가하여 최적의 예측변수 부분집합을 선정하는 방법이다. 반면 부분집합 선택 알고리즘은 회귀모델이 가능한 모든 공간에서 부분적이며 반복적인 탐색을 통해 최적의 예측변수 부분집합을 선정하는 방법이다. 전역탐색과 같이 일정 기준이 있는 것이 아니기 때문에 최적의 부분집합을 선정하였다고 보장하긴 힘들다. 그렇지만 무엇보다 예측변수가 많을 때 사용하기 적합한 방법이다. 예측변수의 개수가 많지 않은 경우에는 전역탐색이 더 낫다.

우선 전역탐색에서 쓰이는 방법들은 다음과 같다.

▸ 전역탐색 : 수정 결정계수(R^2_{adj})

$$R^2 = \frac{SSR}{SST} = \frac{\sum(\hat{y_i} - \bar{y})^2}{\sum(y_i - \bar{y})^2} \quad \Rightarrow \quad R^2_{adj} = 1 - \frac{n-1}{n-p-1}(1-R^2)$$

전역탐색에서 가장 자주 쓰이는 방법은 수정 결정계수(R^2_{adj})를 이용한 적합 여부 판단이다. 위의 식을 보면 알 수 있듯이, SST와 SSR 모두 제곱합의 형태이기 때문에 변수가 추가되면 결정계수는 자연스레 증가한다. 변수의 개수가 많아져서 결정계수 값이 커진 것이지 모델의 성능이 좋아져서 결정계수의 값이 커진 것이 아니므로 결정계수로는 모델의 성능을 평가할 수 없다. 따라서 이 경우에는 수정 결정계수(R^2_{adj})가 자주 쓰인다. 기존의 결정계수와 마찬가지로 수정 결정계수도 값이 높을수록 적합성이 좋다는 것을 의미한다. 수정 결정계수를 사용함으로써 예측변수 개수의 증가로 인해 발생했던 결정계수가 증가하는 효과를 배제시킬 수 있다.

▸ 전역탐색 : AIC와 BIC

AIC와 BIC는 과소적합과 과대적합의 균형을 위한 기준으로, 모델평가 시 적합도뿐 아니라 매개변수의 개수에 대한 penalty도 포함된 지표이다. 더 좋은 모델일수록 두 지표의 값이 작다.

▸ 전역탐색 : 멜로우 C_p ($Mellow's\ C_p$)

$$C_p = \frac{SSE}{\widehat{\sigma^2_{full}}} + 2(p+1) - n$$

멜로우 C_p는 부분집합을 고르기 위해 쓰이는 기준으로, 자료의 표준화된 잔차제곱합 (SSE)의 추정량이다. 이때 모든 예측변수를 사용하는 완전모델에서는 bias가 없다고 가정한다. 이 가정을 바탕으로 부분집합의 모델이 bias가 없다면 평균 멜로우 C_p는 p+1 (예측변수의 개수+1) 과 같다. 이 경우 C_p가 p+1 보다 작으면 좋은 모델로 선정한다.

다음은 부분집합 선택 알고리즘에서 쓰이는 방법이다. 대표적인 반복탐색 알고리즘으로는 전방 선택방법, 후방 소거법, 단계적 선택방법이 있다.

▸ 부분집합 선택 알고리즘 : 전방 선택방법 (forward selection)

전방 선택방법은 예측변수가 전혀 없는 상태에서 유의미한 예측변수를 하나씩 추가하는 방법이다. 이 방법은 추가되는 예측변수의 기여도가 통계적으로 유의하지 않을 때 중단된다.

▸ 부분집합 선택 알고리즘 : 후방 소거법 (backward elimination)

후방 소거법은 모든 예측변수를 사용하는 상태에서 무의미한 예측변수를 하나씩 제거해 나가는 방법이다. 이 방법은 제거되지 않고 남아있는 예측변수들의 기여도가 유의하다고 판단될 때 중단된다. 후방 소거법의 경우 초기 모델을 계산하는 데 시간이 많이 소요되고 불안정하다는 단점이 있다.

▸ 부분집합 선택 알고리즘 : 단계적 선택방법 (stepwise selection)

앞의 두 방법의 단점을 보완한 방법으로 한 번 선택되거나 소거된 예측변수를 다시 고려하여 양방향으로 변수를 선택하는 방법이다. 즉 전진 선택방법에서 시작하였다가 선택된 변수가 3개 이상이 되면 양방향으로 추가와 소거를 번갈아 시행하는 방법이다. 이때 한번 선택되거나 소거된 변수가 고정되지 않기 때문에 상대적으로 많은 경우의 수를 탐색할 수 있어 최적의 결과를 얻을 수 있다. 그러나 그만큼 계산 시간이 많이 소요되기 때문에 학습이 느리다는 단점이 있다.

✓과제 : 회귀모형

3. 다중공선성에 대하여 정리하시오.

다중공선성이란 다중회귀모형에서만 나타나는 경우로, 독립변수들 간에 상관관계가 높을 때 발생하는 문제이다. 다중공선성은 회귀계수와 표준오차를 불안정하게 만들어 검정의 신뢰도를 저하시킨다. 이는 잘못된 변수 해석, 예측 정확도 저하 등을 야기한다. 따라서 다중공선성의 문제를 발생시키지 않기 위해서는 회귀모델 설정 전 미리 다중공선성의 가능성을 확인하는 것이 좋다. 회귀모델에 다중공선성이 있는지 확인하는 방법은 대표적으로 VIF (Variation Inflation Factor, 분산팽창지수) 가 있다.

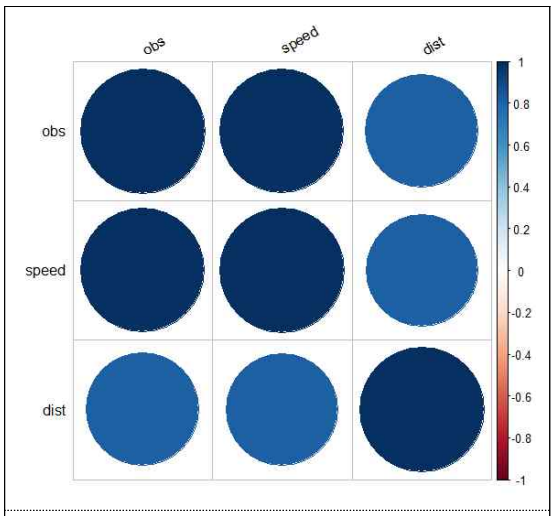
$$VIF_i = \frac{1}{1 - R_i^2}$$

위의 수식에서 R_i^2 은 다중회귀모형에서 i번째 독립변수를 종속변수로 두고, 나머지 독립변수들로 회귀모형을 구성한 보조 회귀식에서 나온 결정계수이다. 이때 각 독립변수 별 VIF를 비교하여 판단하는데, 일반적으로 $VIF > 10$ 이면 다중공선성이 있다고 판단한다. 한편 다중공선성이 발생한 경우, VIF가 높다고 판단된 변수들을 선택하여 제거할 수 있다. 다만 변수를 무조건적으로 제거하는 것이 아니라 VIF가 높더라도 p-value가 유의수준보다 낮은 유의미한 변수라면 제거하지 않는 것이 적절하다.

✓과제 : 회귀모형

1. Cars 데이터 셋을 이용하여 자동차 속도에 따른 제동거리의 관계를 회귀모형을 통해 알아보자.

주어진 Cars 데이터 셋은 obs, speed, distance 의 변수로 이루어진 데이터이다. 이때 ods의 경우 자동차 각각의 데이터를 구별해주는 데이터라 판단하여 speed와 distance 사이의 관계만을 보았다. 따라서 단순 회귀분석으로 두고 진행하였다. 우선 변수 간 상관관계를 보기 위해 상관계수를 구해보았으며 그에 따른 테이블도 그려보았다. 이는 다음과 같다.

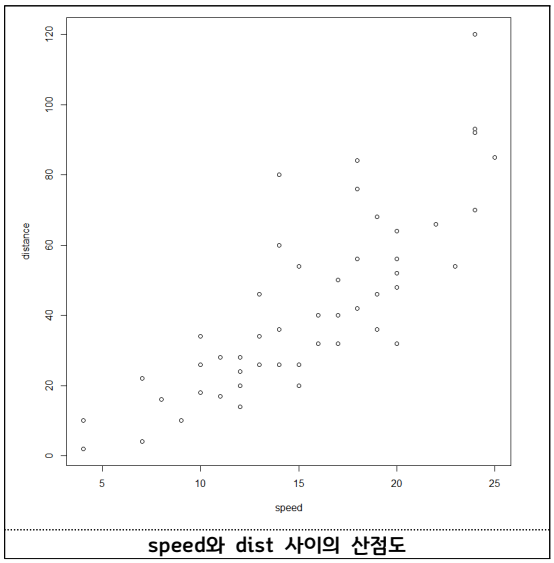


상관관계 테이블

```
> round(cor(cars),2)->a;a
       obs speed dist
obs   1.00  0.99 0.82
speed 0.99  1.00 0.81
dist  0.82  0.81 1.00
```

변수 간 상관관계

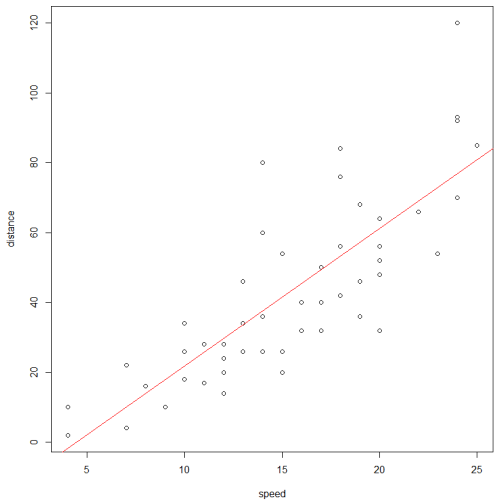
speed와 dist 변수는 매우 높은 양의 상관관계를 보였다. 이를 바탕으로, 두 변수 간 산점도를 그려보았다. 아래의 그림과 같이 강한 양의 상관관계를 띠고 있음을 알 수 있다.



speed와 dist 사이의 산점도

따라서 이를 바탕으로 단순 선형회귀모델을 세워보았다. 단순 선형회귀모델을 돌린 결과 아래와 같이 speed 가 증가할수록 distance 가 증가함을 알 수 있다. (이때 자동차는 속도가 올라갈수록 이동거리는 당연히 길어진다고 생각했기 때문에 당연한 결과라고 판단하였다.)

한편 두 변수 사이의 관계는 $dist_i = 3.9324\, speed_i - 17.5791 + \epsilon_i$ 로 표현할 수 있다.




```
Call:
lm(formula = dist ~ speed, data = cars)

Residuals:
    Min      1Q  Median      3Q     Max
-29.069  -9.525  -2.272   9.215  43.201

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) -17.5791     6.7584  -2.601   0.0123 *
speed         3.9324     0.4155   9.464 1.49e-12 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 15.38 on 48 degrees of freedom
Multiple R-squared:  0.6511,    Adjusted R-squared:  0.6438
F-statistic: 89.57 on 1 and 48 DF,  p-value: 1.49e-12
```

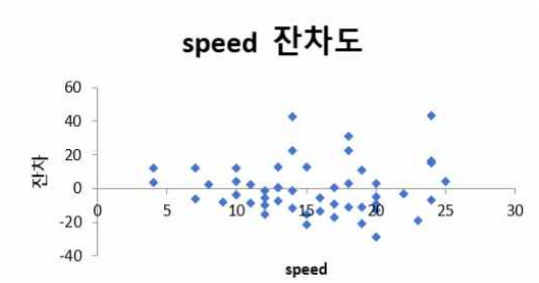
```
> #회귀계수신뢰구간
> confint(car,level=0.95)
                2.5 %    97.5 %
(Intercept) -31.167850 -3.990340
speed         3.096964  4.767853
> #잔차제곱합
> deviance(car)
[1] 11353.52
```

그 외 통계량 값

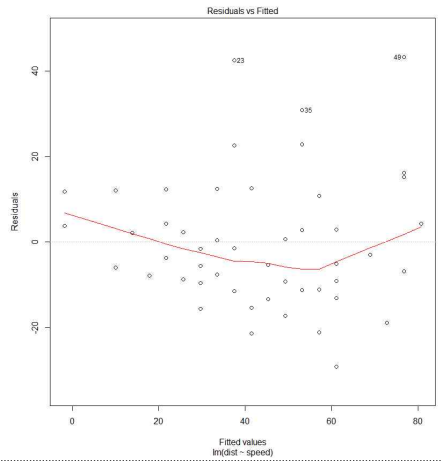
R을 이용하여 생성한 회귀모델의 통계량 값들 (회귀계수, 회귀계수의 신뢰도, p-value, 잔차제곱합, FH, 결정계수 및 조정된 결정계수 값 등) 모두가 엑셀에 명시된 값과 동일하게 나왔음을 확인하였다. 이를 바탕으로 잔차 그래프, 정규 확률도, 생성된 모델의 선형회귀선을 나타낸 그래프를 각각 그려 다시 엑셀과 비교해보았다. 이는 다음의 그래프와 같다. 그래프를 보면 알 수 있듯이 EXCEL과 R의 그래프가 유사하게 나왔음을 알 수 있다. 특히 잔차도, 정규확률도, 선 적합도의 경우 단순 선형회귀모형의 가정을 만족하는 것을 보이는 그래프였는데, 가정이 모두 성립함을 확인할 수 있었다.

▸단순 선형회귀모형의 가정

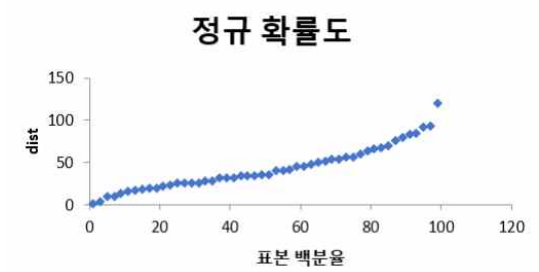
①회귀함수의 선형성 ②모형 오차항의 등분산성 ③모형 오차항 분포의 정규성 ④모형 오차항의 독립성



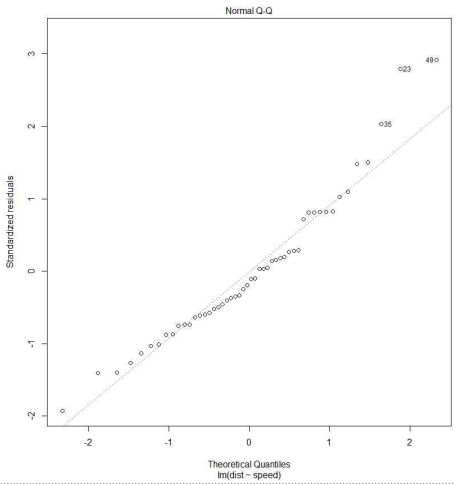
EXCEL



R

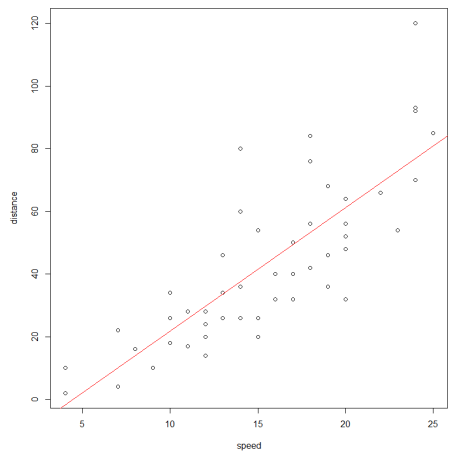


EXCEL



R

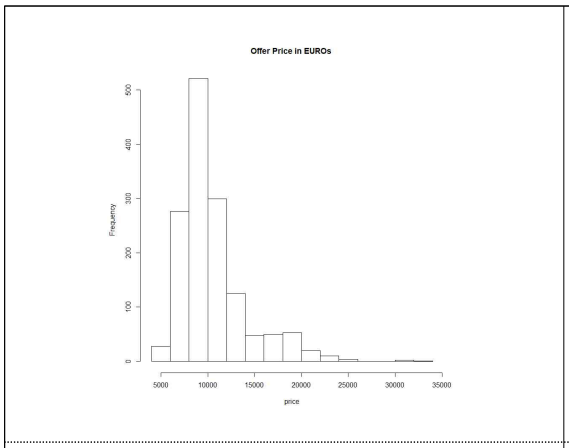
EXCEL



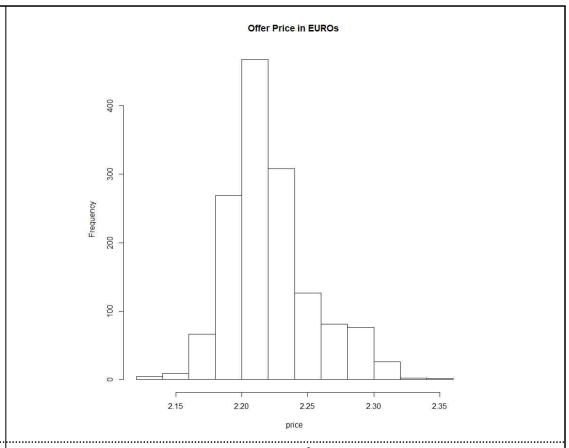
R

2. ToyotaCorolla 데이터 셋은 도요타 코롤라 중고차의 가격 예측에 관한 데이터입니다. 본 데이터에는 중고차 가격에 영향을 미칠 것이라 판단되는 독립 변수들이 있습니다. 중고차 가격에 대하여 다중 선형 회귀분석을 적합시켜 어떠한 변수들이 중고차 가격에 어떻게 영향을 미치는지 알아보시오.

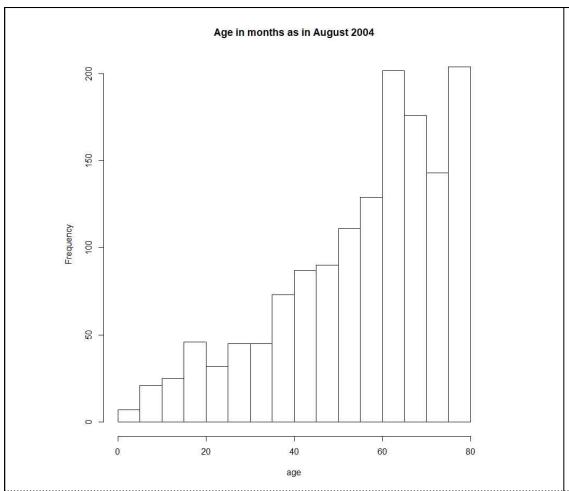
우선 분석하기에 앞서 EDA를 진행하였다. 0과 1로만 구성된 범주형 변수를 제외하고는 모든 변수에 대해 진행하였다. 특히 연속형 변수에서 왜곡 분포가 있는 경우에는 정규화를 시켜 데이터의 균형을 맞추었다. 실제로 정규성을 따르는 데이터일수록 회귀분석 모델이 좋게 나오기 때문에 더욱 변수변환을 진행하게 되었다.



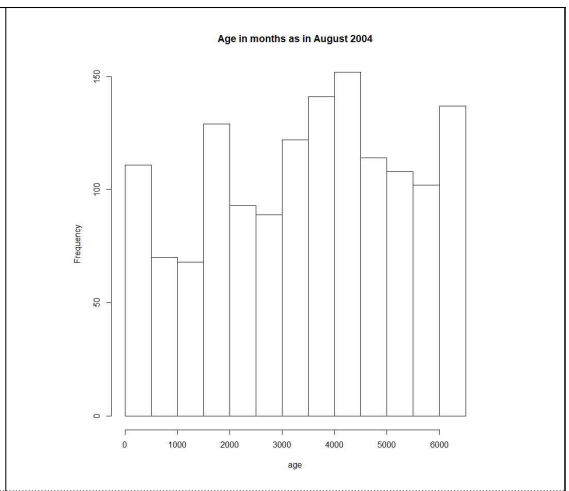
수정 전



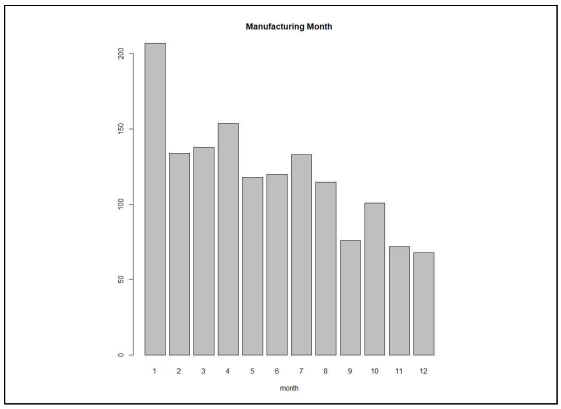
수정 후

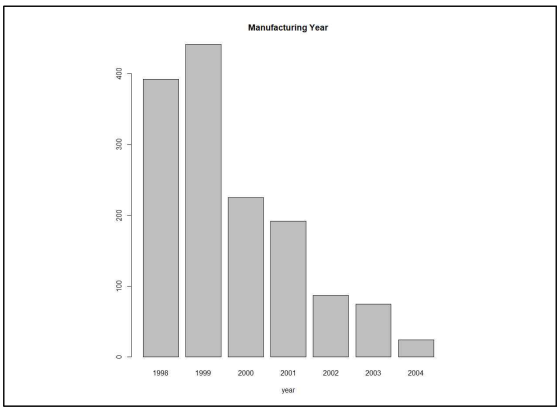


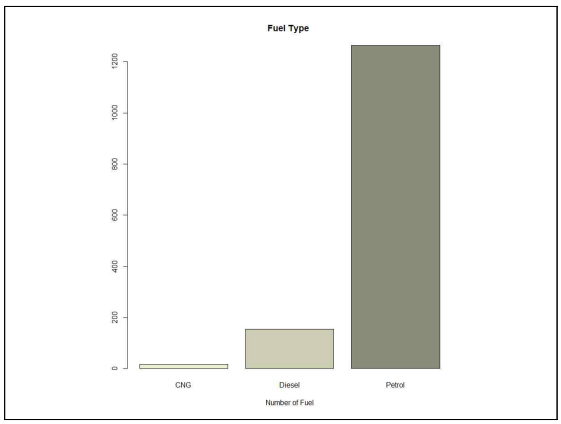
수정 전

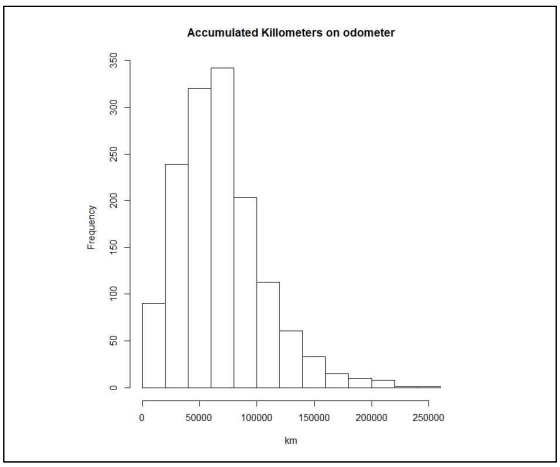


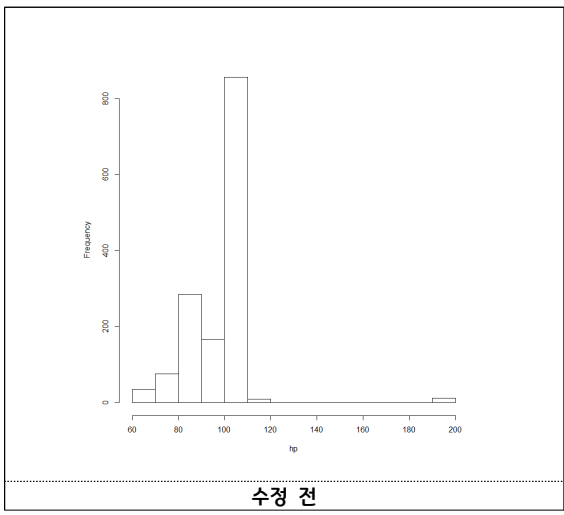
수정 후


Manufacturing Month
0
50
100
150
200
1
2
3
4
5
6
7
8
9
10
11
12
month

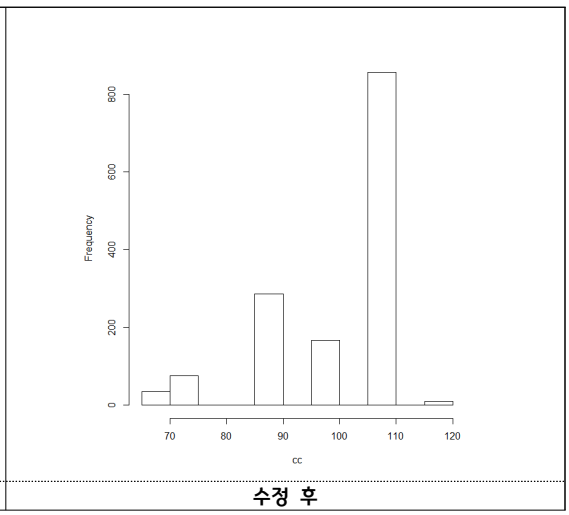

Manufacturing Year
0
100
200
300
400
1998
1999
2000
2001
2002
2003
2004
year


Fuel Type
0
200
400
600
800
1000
1200
CNG
Diesel
Petrol
Number of Fuel

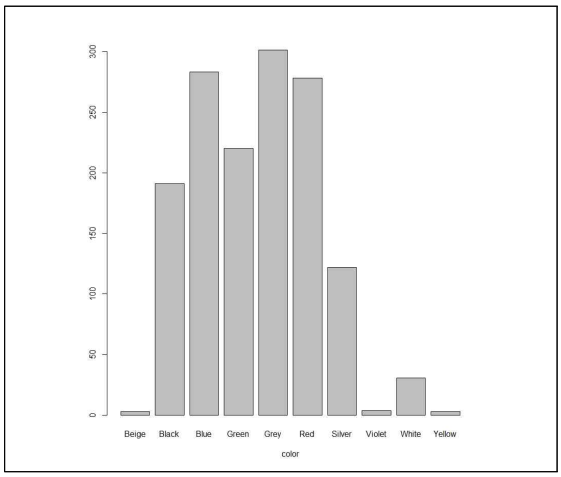

Accumulated Killometers on odometer
Frequency
0
50
100
150
200
250
300
350
0
50000
100000
150000
200000
250000
km

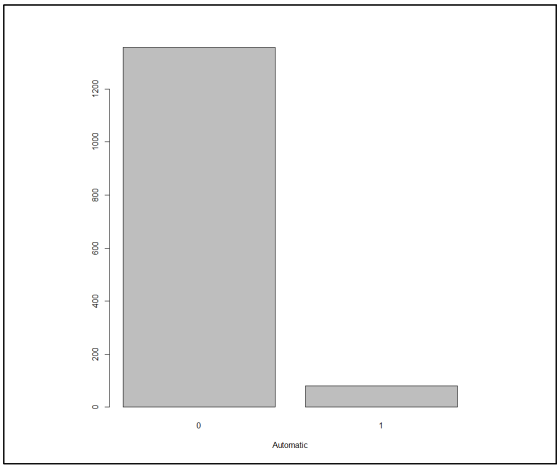

Frequency
0
200
400
600
800
60
80
100
120
140
160
180
200
hp

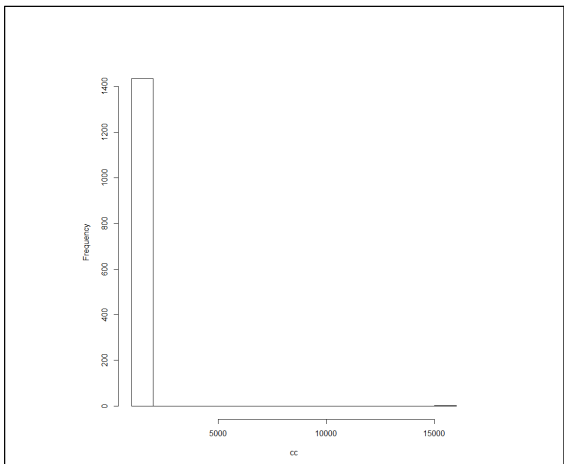
수정 전

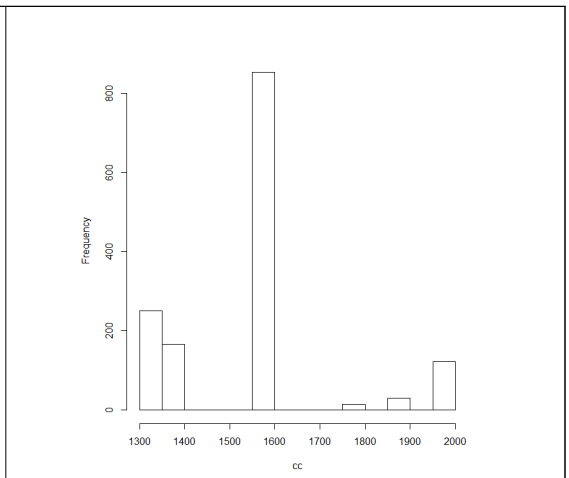

Frequency
0
200
400
600
800
70
80
90
100
110
120
cc

수정 후

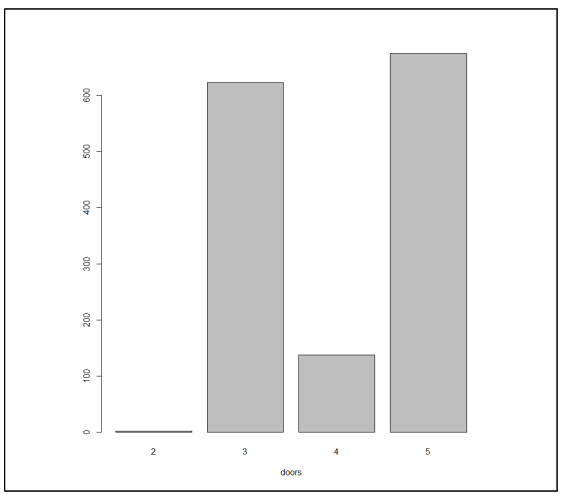

0
50
100
150
200
250
300
Beige
Black
Blue
Green
Grey
Red
Silver
Violet
White
Yellow
color

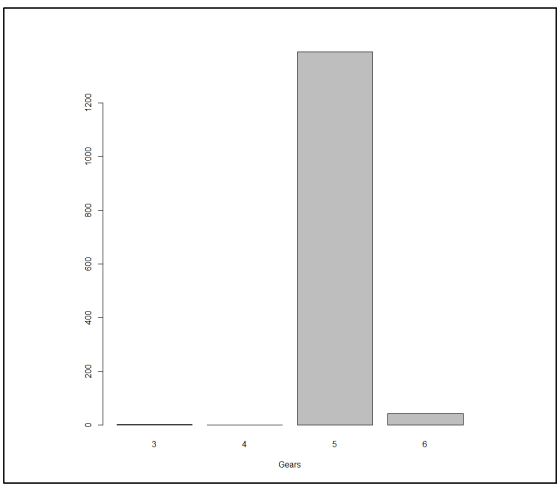

0
200
400
600
800
1000
1200
0
1
Automatic

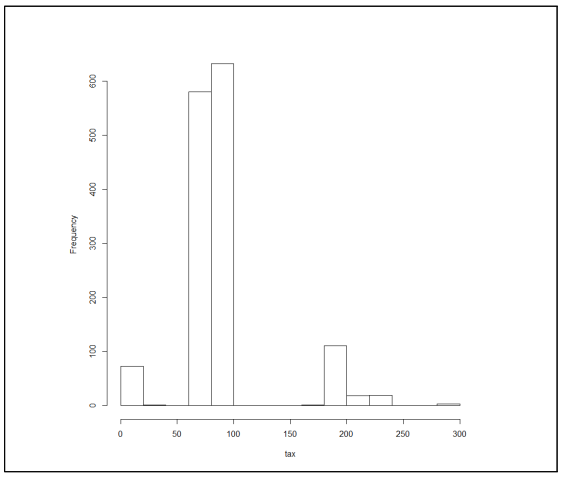

Frequency
0
200
400
600
800
1000
1200
1400
5000
10000
15000
cc

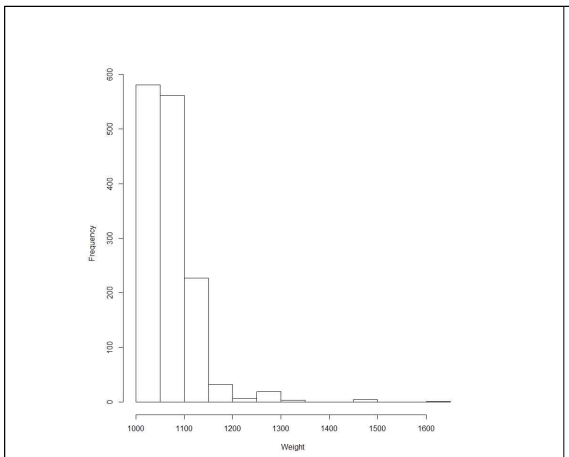

Frequency
0
200
400
600
800
1300
1400
1500
1600
1700
1800
1900
2000
cc

수정 전 **수정 후**

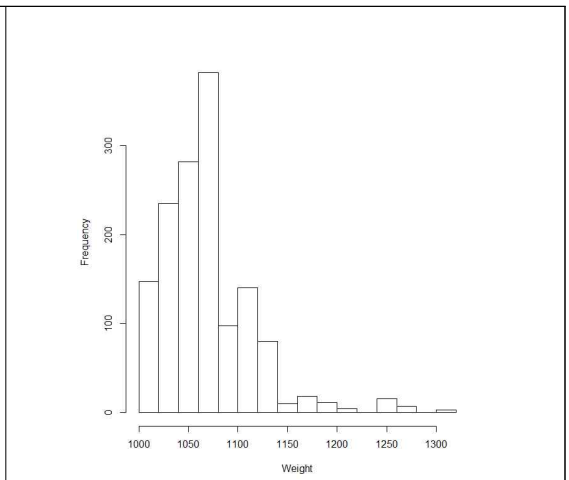

0
100
200
300
400
500
600
2
3
4
5
doors


0
200
400
600
800
1000
1200
3
4
5
6
Gears

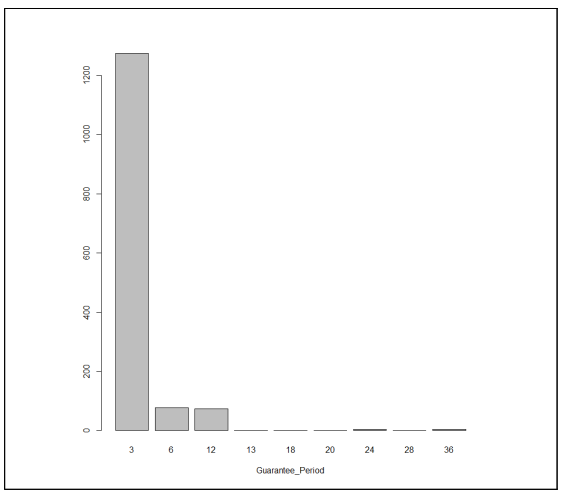

Frequency
0
100
200
300
400
500
600
0
50
100
150
200
250
300
tax


Frequency
0
100
200
300
400
500
600
1000
1100
1200
1300
1400
1500
1600
Weight

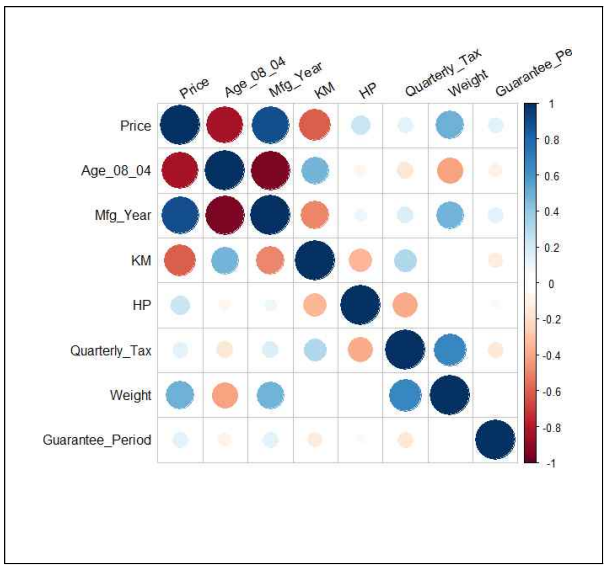
수정 전


Frequency
0
100
200
300
1000
1050
1100
1150
1200
1250
1300
Weight

수정 후


0
200
400
600
800
1000
1200
3
6
12
13
18
20
24
28
36
Guarantee_Period

다음은 연속형 변수 간의 상관관계를 시각화한 테이블이다. 타겟변수인 Price와 높은 상관관계를 보이는 변수는 Age_08_04, Mfg_Year, Weight이다. 이때 Age가 클수록 Price는 작아지는 반면에 Year과 Weight가 클수록 Price가 커짐을 알 수 있다. 자동차 가격임을 생각해보면, 어느 정도 합당한 결과임을 직감적으로 느낄 수 있다.



다음으로 다중공선성을 고려하여 독립변수 간 상관관계가 높은 변수를 삭제하는 과정을 거쳤다. 이때 각 독립변수 별 VIF를 비교하여 판단하였고, $VIF > 10$ 인 경우 다중공선성이 있다고 판단하였다. 단, 변수는 무조건적으로 제거하지 않았고 VIF가 높더라도 p-value가 유의수준보다 낮은 유의미한 변수라면 제거하지 않았다. 또한 다중공선성을 하기 앞서 ID, Model, Cylinders 는 먼저 삭제해주었다. ID, Model 은 회귀모델을 돌리게 될 때 더미변수를 많이 만들며, 무엇보다 범주형 변수라 하기에도 너무 많았기 때문에 제거해주었다. Cylinders는 모든 행이 결측치로 처리되어 오류가 나므로 다중공선성을 진행하기 전에 삭제하였다.

이를 바탕으로 다중공선성을 진행하였는데, GVIF^(1/(2*Df)) 이 2보다 큰 것들 중 가장 큰거를 먼저 삭제하는 형식으로 진행하였다. 예를 들어 변수 a의 GVIF^(1/(2*Df)) 값이 2보다 크면서 가장 크면서 유의확률이 컸다면 삭제해주고, 다시 모델을 돌리고 VIF를 확인하는 방향으로 반복 진행하였다.

```
> vif(model)
                      GVIF Df GVIF^(1/(2*Df))
Age_08_04         9.106169  1        3.017643
Mfg_Month         1.780051  1        1.334186
Mfg_Year         14.688147  1        3.832512
KM                2.160067  1        1.469717
Fuel_Type        81.632181  2        3.005836
HP               19.019798  1        4.361169
Met_Color         1.310464  1        1.144755
Color             1.792252  9        1.032946
Automatic         1.280186  1        1.131453
CC               22.284703  1        4.720668
Doors             1.692673  1        1.301028
Gears             1.250326  1        1.118180
Quarterly_Tax     5.285533  1        2.299029
Weight            8.723951  1        2.953633
Mfr_Guarantee     1.219166  1        1.104159
BOVAG_Guarantee   1.402111  1        1.184108
Guarantee_Period  1.629571  1        1.276546
ABS               2.435498  1        1.560608
Airbag_1          1.601995  1        1.265699
Airbag_2          3.154082  1        1.775974
Airco             1.865697  1        1.365905
Automatic_airco   1.730553  1        1.315505
Boardcomputer     2.700821  1        1.643417
CD_Player         1.569727  1        1.252888
Central_Lock      4.591983  1        2.142891
Powered_Windows   4.656090  1        2.157797
Power_Steering    1.590130  1        1.261003
Radio            61.864983  1        7.865430
Mistlamps         2.153494  1        1.467479
Sport_Model       1.670328  1        1.292412
Backseat_Divider  2.722643  1        1.650043
Metallic_Rim      1.341393  1        1.158185
Radio_cassette   61.614203  1        7.849472
Parking_Assistant 1.051624  1        1.025487
Tow_Bar           1.175213  1        1.084073
```

진행 전

```
> vif(model)
                      GVIF Df GVIF^(1/(2*Df))
Age_08_04         9.085074  1        3.014146
Mfg_Month         1.777236  1        1.333130
Mfg_Year         14.639874  1        3.826209
KM                2.120878  1        1.456323
Fuel_Type        15.574431  2        1.986566
HP                3.447249  1        1.856677
Met_Color         1.307891  1        1.143631
Color             1.758822  9        1.031866
Automatic         1.250338  1        1.118185
Doors             1.643661  1        1.282053
Gears             1.250069  1        1.118065
Quarterly_Tax     5.216598  1        2.283987
Weight            8.325493  1        2.885393
Mfr_Guarantee     1.217696  1        1.103493
BOVAG_Guarantee   1.388848  1        1.178494
Guarantee_Period  1.609287  1        1.268577
ABS               2.431564  1        1.559347
Airbag_1          1.596581  1        1.263559
Airbag_2          3.122460  1        1.767048
Airco             1.836966  1        1.355347
Automatic_airco   1.723591  1        1.312856
Boardcomputer     2.693544  1        1.641202
CD_Player         1.564664  1        1.250865
Powered_Windows   1.810738  1        1.345637
Power_Steering    1.576330  1        1.255520
Mistlamps         2.091699  1        1.446271
Sport_Model       1.650047  1        1.284541
Backseat_Divider  2.692697  1        1.640944
Metallic_Rim      1.338042  1        1.156738
Radio_cassette    1.210691  1        1.100314
Parking_Assistant 1.049755  1        1.024575
Tow_Bar           1.169800  1        1.081573
```

진행 후

위의 표는 다중공선성을 진행하기 전과 진행하고 난 후의 VIF를 나타낸 것이다. VIF와 유의확률 두가지 모두 고려한 결과 CC, Central_Lock, Radio를 삭제하였다. 이후 남은 변수를 가지고 단계적 선택 방법을 이용하여 예측변수의 개수를 선정하였다. 결과적으로 더미변수가 있음에도 28개 변수로 줄였으며 수정 결정계수는 0.8718이 나왔다.

```
> summary(eunni)

Call:
lm(formula = Price ~ Age_08_04 + Mfg_Year + KM + Fuel_Type +
    Color + Automatic + Gears + Quarterly_Tax + Weight + Mfr_Guarantee +
    BOVAG_Guarantee + Guarantee_Period + Airbag_2 + Airco + Automatic_airco +
    CD_Player + Powered_Windows + Backseat_Divider + Tow_Bar,
    data = h)

Residuals:
      Min        1Q    Median        3Q       Max
-0.069633 -0.005964  0.000053  0.006582  0.060180

Coefficients:
                   Estimate Std. Error t value Pr(>|t|)
(Intercept)      -1.680e+01  1.025e+00 -16.389  < 2e-16 ***
Age_08_04        -2.671e-10  5.168e-11  -5.168 2.71e-07 ***
Mfg_Year          9.415e-03  5.145e-04  18.299  < 2e-16 ***
KM               -1.843e-07  1.113e-08 -16.555  < 2e-16 ***
Fuel_TypeDiesel  -1.440e-03  2.898e-03  -0.497 0.619501
Fuel_TypePetrol   2.248e-02  3.255e-03   6.906 7.59e-12 ***
ColorBlack        6.638e-03  6.422e-03   1.034 0.301507
ColorBlue         5.072e-03  6.412e-03   0.791 0.429092
ColorGreen        3.085e-03  6.418e-03   0.481 0.630788
ColorGrey         5.134e-03  6.415e-03   0.800 0.423616
ColorRed          4.900e-03  6.404e-03   0.765 0.444368
ColorSilver       5.101e-03  6.463e-03   0.789 0.430081
ColorViolet       1.591e-03  8.466e-03   0.188 0.850939
ColorWhite       -4.275e-03  6.696e-03  -0.638 0.523309
ColorYellow       3.865e-03  9.010e-03   0.429 0.667975
Automatic         2.717e-03  1.331e-03   2.041 0.041426 *
Gears             3.592e-03  1.732e-03   2.073 0.038331 *
Quarterly_Tax     1.544e-04  1.604e-05   9.625  < 2e-16 ***
Weight            1.378e-04  1.248e-05  11.038  < 2e-16 ***
Mfr_Guarantee     3.347e-03  6.443e-04   5.195 2.35e-07 ***
BOVAG_Guarantee   5.379e-03  1.108e-03   4.854 1.34e-06 ***
Guarantee_Period  6.418e-04  1.196e-04   5.366 9.40e-08 ***
Airbag_2         -1.295e-03  9.109e-04  -1.422 0.155232
Airco             3.568e-03  7.549e-04   4.726 2.52e-06 ***
Automatic_airco   8.877e-03  1.643e-03   5.404 7.65e-08 ***
CD_Player         2.105e-03  8.488e-04   2.481 0.013237 *
Powered_Windows   2.721e-03  7.269e-04   3.744 0.000189 ***
Backseat_Divider -1.790e-03  1.051e-03  -1.703 0.088865 .
Tow_Bar          -1.893e-03  6.719e-04  -2.817 0.004917 **
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.01098 on 1390 degrees of freedom
Multiple R-squared:  0.8743,    Adjusted R-squared:  0.8718
F-statistic: 345.2 on 28 and 1390 DF,  p-value: < 2.2e-16
```

단계적 선택 방법

다중공선성과 VIF를 모두 진행한 후 나온 데이터를 시각화해보면 다음과 같다.