

다변량분석 1번째 과제

Assignment 1

콘크리트 압축강도 분석

학과
학번
이름

산업경영공학부
2016170863
추창욱

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

답) 제가 선정한 데이터는 여러 재료들을 혼합하여 콘크리트를 만드는 과정에서 각 재료의 양에 따른 콘크리트의 압축강도를 예상해볼 수 있는 데이터 셋이다. 이 데이터 셋을 고르기 이전에 의료보험비를 예측해보는 데이터 셋, 와인의 품질을 예상해볼 수 있는 데이터 셋 등 여러 선택지가 있었으나, 이전에 공사장에서 일을 해본 경험이 있는데 그때 들어가는 여러 재료들에 따라 콘크리트가 갈라지거나 또는 잘 굳지 않는 등의 현상을 목격한적이 있었기에 늘 궁금했던 분야라 선택하게 되었다. 또한, 아예 지식이 없는 분야보다 가장 명확하게 결과를 얻을 수 있을 것 같아 보이는 데이터를 분석함으로써 재미도 생길 것 같아 콘크리트 압축 강도에 대한 데이터셋을 고르게 되었다.

데이터셋 다운로드 링크: <https://www.kaggle.com/maajdl/yeh-concret-data>

Q2) 해당 데이터셋의 종속변수와 설명 변수는 어떤 것들이 있는가? 분석 전에 아래 세가지 질문에 대해서 스스로 생각해보고 답변을 하시오.

답) 세가지 질문에 대한 답을 하기 전 가장 먼저 어떤 데이터로 이루어져 있는지를 알 필요가 있다고 생각한다. 현재 이 데이터셋에는 총 charges라는 Output variable이 있고, input variable로는 우선 총 6가지의 변수가 있다.

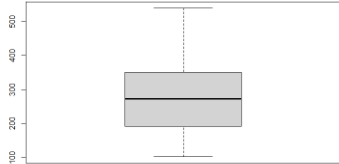
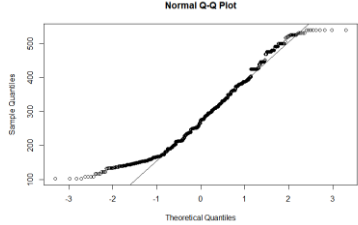
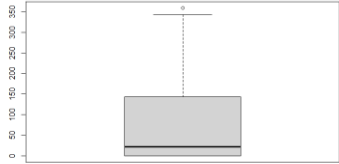
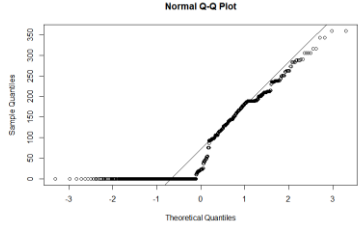
Input Variable	
Cement	kg in a m3 mixture
Blast Furnace Slag	kg in a m3 mixture
Fly Ash	kg in a m3 mixture
Water	kg in a m3 mixture
Superplasticizer	kg in a m3 mixture
Coarse Aggregate	kg in a m3 mixture

Fine Aggregate	kg in a m3 mixture
Age	Day (1~365)
Output Variable	
Concrete compressive strength	MPa

1. 데이터는 종속변수와 설명 변수들 사이에는 선형관계가 있다고 가정할 수 있을 것 같다. 첫번째 이유는 혼합물을 만드는 과정이기 때문이다. 여러 재료들의 양과 배합을 조절해가며 만드는 것은 어떤 공식처럼 작용을 할 것이라고 생각한다. 일 차원적인 생각일 수도 있지만 상식적으로 생각해보았을 때, 혼합을 하는 과정에서 각각 들어가는 물질마다 어떤 효과를 발휘하기 때문에 포함이 되기 때문에 영향을 미칠 것 같다. 또한, Age같은 경우에도 시간이 지나며 변하는 물질은 세상에 수도 없이 많다. 그렇기에 시간이 지날수록 계속은 아니더라도 어느정도 콘크리트 강도에 일정한 영향을 미칠 것 같다는 생각이 들어 이 데이터는 선형관계가 있다고 가정할 수 있을 것 같다.
2. 이 데이터에서 가장 높은 상관 관계를 나타내는 변수들은 cement와 slag일 것이라고 생각한다. 그 이유는 조사를 해본 결과 cement는 저희가 평소에 잘 아는 그 시멘트 이고, 고로슬래그 같은 경우는 구성원소는 일반 암석과 같고, 성분은 시멘트와 유사하다는 정보를 찾았기 때문이다. 또한 화학성분이 시멘트와 유사한 점이 많다는 점에서 일으키는 영향력이 시멘트와 상당히 비슷하여 상관관계가 매우 높을 것으로 예상된다.
3. 종속변수를 예측하는데 필요하지 않을 것이라고 예상되는 변수는 순차대로 Fine aggregate와 coarse aggregate라고 생각한다. 왜냐하면 이는 단순 모래라고 생각해도 무방하기 때문이다. 콘크리트의 압축강도를 결정짓는 가장 중요한 요소는 혼합물들 간의 화학 반응이라고 생각하는데 모래는 단순히 질량을 채우기 위한 수단이라는 생각이 들었기 때문이다. 물론 콘크리트를 딱딱하게 만드는 데 영향을 주긴 할 것 같지만, 압축 강도를 높이는 역할을 하지 않는 단순 질량일 것 같다는 것이 나

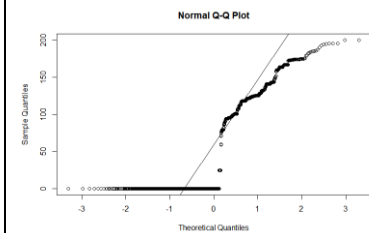
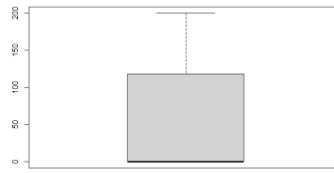
의 의견이다.

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?

단변량 통계량 계산 및 분석		
<pre> > mean(insurance\$cement) [1] 281.1679 > sd(insurance\$cement) [1] 104.5064 > skewness(insurance\$cement) [1] 0.5087389 > kurtosis(insurance\$cement) [1] 2.476052 </pre>		
<p>1. Cement</p> <p>평균은 281을 가지며, 표준편차 104정도로 퍼져 있음을 확인할 수 있다. 왜도는 자료의 대칭성을 알아보는 척도로 보통 0에 가까우면 좌우대칭이 가장 잘 맞는 분포라고 하지만 이 변수의 경우 0.5정도를 가지고 대칭성이 어긋난다. 즉 왼쪽으로 기울어진 형태라고 할 수 있다. 첨도는 정규분포대비 봉오리의 높이를 알아보는 척도인데 보통 3정도가 기준이라고 명시하며 이경우 2.47정도의 값을 가지며 일반 정규분포에 비해 조금 낮은 높이를 가진다고 볼 수 있다. 이는 박스 플롯으로도 확인이 가능한데, 더 직관적으로 보기 위해 qqplot을 그려본 결과, 좀 어긋나느 부분이 있기는 하나 정규성을 크게 벗어나지는 않다고 판단된다.</p>		
<pre> > mean(insurance\$slag) [1] 73.89583 > sd(insurance\$slag) [1] 86.27934 > skewness(insurance\$slag) [1] 0.7995503 > kurtosis(insurance\$slag) [1] 2.488468 </pre>		
<p>2. Slag</p> <p>이 데이터는 평균은 약 74, 표준편차는 86정도로 퍼져 있다. 왜도는 0.79로 꽤나 높은 값을 가져 왼쪽으로 기울어진 형태일 것이며, 첨도는 2.48로 생각보다는 양호한 높이를 가진 것 같다. 박스 플롯을 보면 이 현상이 정확히 보이며 qqplot도 마찬가지이다. 특히나 왜도 값이 너무</p>		

크기 때문에 정규성을 만족하지 않는다고 생각할 수 있다.

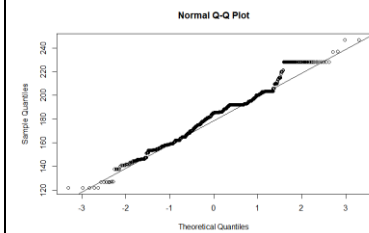
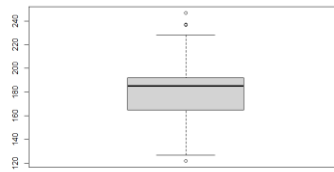
```
> mean(insurance$flyash)
[1] 54.18835
> sd(insurance$flyash)
[1] 63.997
> skewness(insurance$flyash)
[1] 0.536571
> kurtosis(insurance$flyash)
[1] 1.671875
```



3. Flyash

평균은 54인데 표준편차는 대략 64로 상당히 많이 퍼져 있을 거 라는 것을 알 수 있다. 왜도는 0.53으로 이 데이터 또한 왼쪽으로 치우침이 있을 것이며, 첨도는 1.64로 일반 정규분포보다 높이가 낮을 것이다. 박스 플롯을 보면 왼쪽으로 치우침이 보이고 평균에서 값들이 꽤나 퍼진 듯한 모습을 보인다. 또한 결정적으로 qqplot을 보면 정규성을 만족하지 않는다는 것을 확인 할 수 있다.

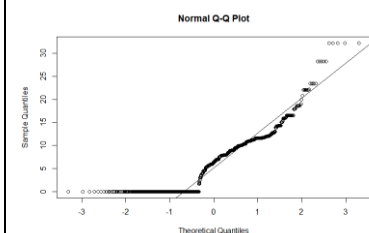
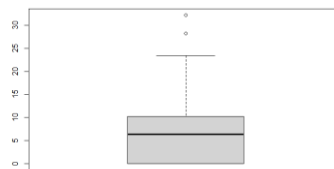
```
> mean(insurance$water)
[1] 181.5673
> sd(insurance$water)
[1] 21.35422
> skewness(insurance$water)
[1] 0.07451966
> kurtosis(insurance$water)
[1] 3.11567
```



4. Water

평균은 181이며 표준편차는 21로 많이 꽤나 괜찮은 퍼짐 정도를 가진 듯 하며 왜도는 0에 가깝고, 첨도 또한 3에 가까운 것을 보아 정규성을 어느정도 된다는 것을 알 수 있다. 박스 플롯과 qqplot에서도 이상치를 제외하고는 어느정도 정규성을 만족한다는 것을 알 수 있다.

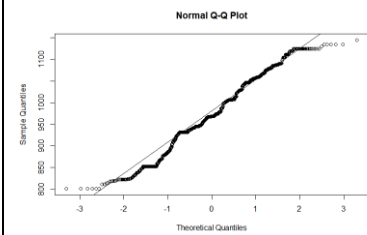
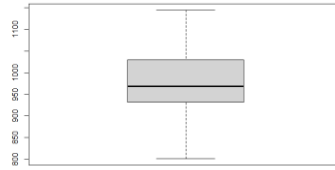
```
> mean(insurance$superplasticizer)
[1] 6.20466
> sd(insurance$superplasticizer)
[1] 5.973841
> skewness(insurance$superplasticizer)
[1] 0.9058809
> kurtosis(insurance$superplasticizer)
[1] 4.398608
```



5. Superplasticizer

이 데이터는 결론부터 말하면 정규성을 띄지 않는다고 판단된다. 왜도와 첨도의 값이 너무 어긋나 있고, qqplot과 박스플롯을 통해서도 정규성을 띄지 않는다는 것이 바로 보인다.

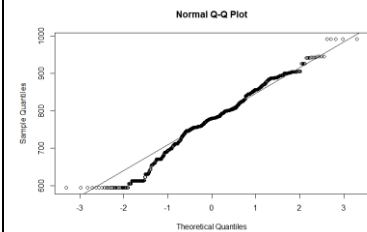
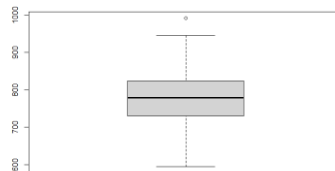
```
> mean(insurance$coarseaggregate)
[1] 972.9189
> sd(insurance$coarseaggregate)
[1] 77.75395
> skewness(insurance$coarseaggregate)
[1] -0.04016115
> kurtosis(insurance$coarseaggregate)
[1] 2.398068
```



6. Coarse aggregate

평균은 972 표준편차는 77로 생각보다 작게 나왔다. 왜도는 -0.04로 오른쪽에 살짝 치우친 형태이며, 첨도는 2.4정도로 높이는 낮다고 볼 수 있다. 첨도는 좀 낮은 편에 속하는 거 같으나 왜도는 큰 차이가 나지 않는다는 점에서 박스플롯과 qqplot을 본 결과 꽤나 정규성을 가진다고 말할 수 있을 것 같다는 판단을 내렸다.

```
> mean(insurance$fineaggregate)
[1] 773.5805
> sd(insurance$fineaggregate)
[1] 80.17598
> skewness(insurance$fineaggregate)
[1] -0.252641
> kurtosis(insurance$fineaggregate)
[1] 2.892499
```



7. Fine aggregate

첨도는 -0.25로 꽤나 오른쪽으로 치우쳤을 거라 생각되며 첨도는 나름 3에 근사한듯한 값을 가진다고 판단된다. 정규성을 조금은 만족한다고 볼 수 있을 것 같다.

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

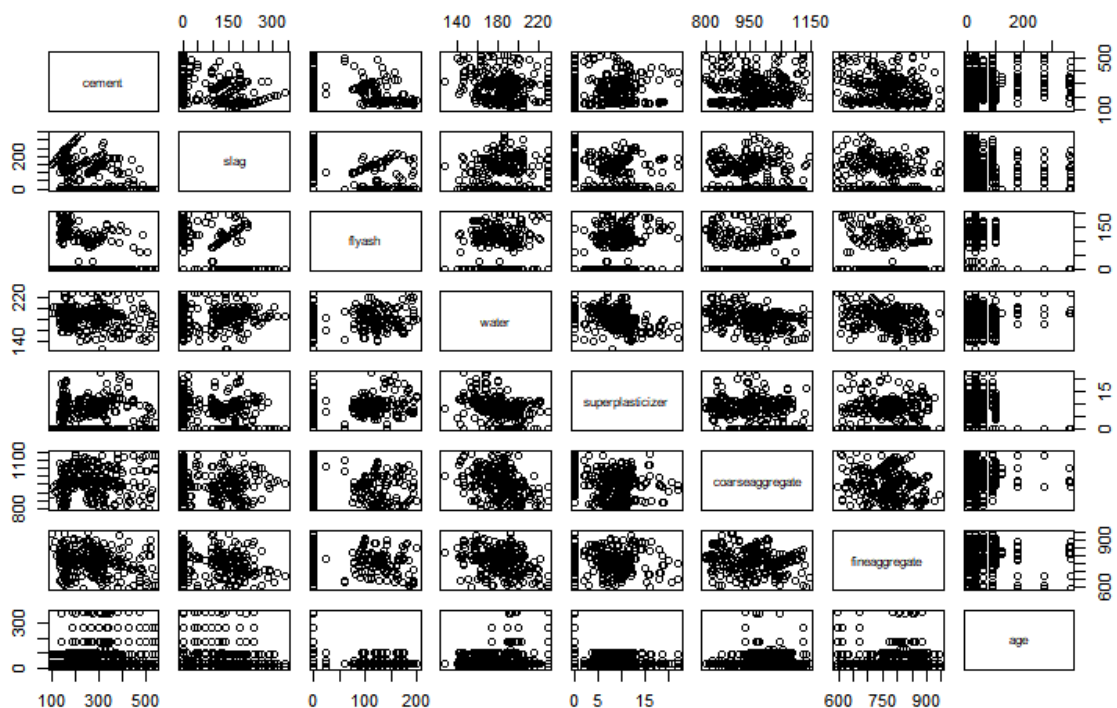
답) Q3에 의하면 박스플롯에 근거하여 이상치를 가지고 있는 변수들은 각각 Slag, Water, Superplasticizer, Fine aggregate라는 것을 알 수 있었다.

```
> which(insurance$fineaggregate>summary(insurance$fineaggregate)[5] + 1.5*IQR(insurance$fineaggregate))
[1] 75 98 121 144 167
> #이상치 데이터 삭제하기
> summary(insurance$slag)
  Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
   0.0    0.0    22.0   73.9  142.9  359.4
> which(insurance$slag>summary(insurance$slag)[5] + 1.5*IQR(insurance$slag))
[1] 554 560
> summary(insurance$water)
  Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
 121.8  164.9  185.0  181.6  192.0  247.0
> which(insurance$water>summary(insurance$water)[5] + 1.5*IQR(insurance$water))
[1] 863 874 937 1020
> which(insurance$water<summary(insurance$water)[2] - 1.5*IQR(insurance$water))
[1] 225 226 227 228 229
> summary(insurance$superplasticizer)
  Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
 0.000  0.000  6.400  6.205  10.200  32.200
> which(insurance$superplasticizer>summary(insurance$superplasticizer)[5] + 1.5*IQR(insurance$superplasticizer))
[1] 77 80 100 103 123 126 146 149 169 172
> summary(insurance$fineaggregate)
  Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
 594.0  731.0  779.5  773.6  824.0  992.6
> which(insurance$fineaggregate>summary(insurance$fineaggregate)[5] + 1.5*IQR(insurance$fineaggregate))
[1] 75 98 121 144 167
```

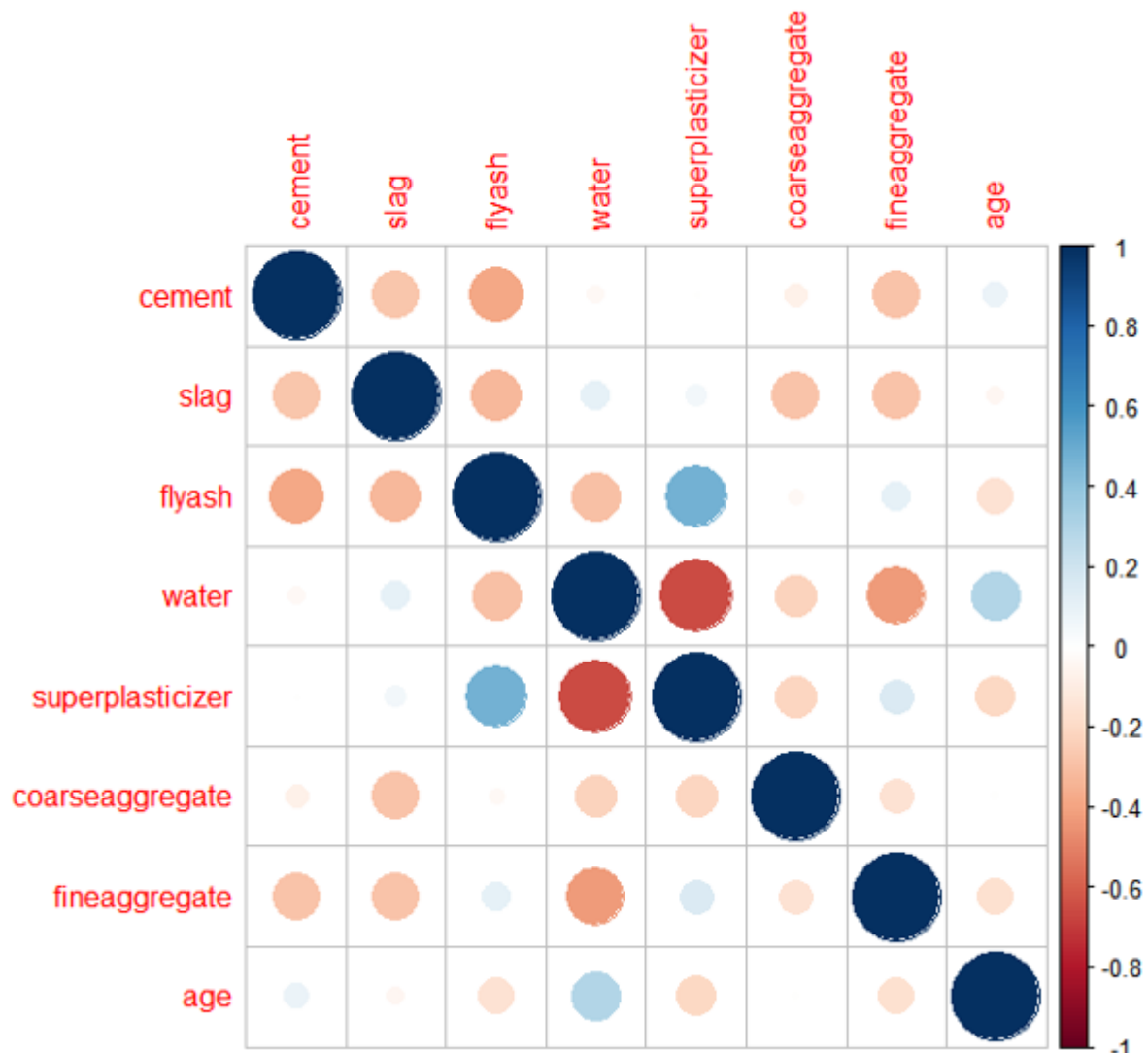
위의 코드를 통해 이상치를 갖는 행이 몇 행들인지를 알아낼 수 있었다. 이상치라는 조건은 통계적으로 상/하위 3표준편차를 벗어나거나 1.5IQR을 벗어나면 이상치로 판단한다고 한다. 즉 극단치 경계를 넘은 값을 가진 행들을 찾아낸 것이다. 이상치를 모두 제거해주고 확인한결과 더 이상 이상치가 나오지 않는것을 알수 있었다.

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?

답) 우선, input 변수들의 상관 관계를 확인하기 위해 모든 변수들간의 가능한 조합의 scatter plot을 표로 구현해보았다.



이런 식의 결과가 나오게 되는데 한눈에 보기에 너무 복잡하고 가시성이 좋지 않아 어떤 변수들간의 상관관계가 높게 나오는지 파악하기가 쉽지 않다. 이를 개선하기 위해 다음과 같이 구현을 해보았다.



	cement	slag	flyash	water	superplasticizer	coarseaggregate	fineaggregate	age
cement	1.000000000	-0.27937070	-0.3814044	-0.03946288	-0.003685037	-0.073829363	-0.2815493	0.086439883
slag	-0.279370696	1.000000000	-0.3275070	0.10382283	0.058051649	-0.288325293	-0.2809450	-0.041653776
flyash	-0.381404422	-0.32750696	1.000000000	-0.29274332	0.479256770	-0.036711198	0.1056053	-0.158122455
water	-0.039462879	0.10382283	-0.2927433	1.000000000	-0.656352066	-0.222398418	-0.4218652	0.297735722
superplasticizer	-0.003685037	0.05805165	0.4792568	-0.65635207	1.000000000	-0.210287997	0.1573840	-0.209219553
coarseaggregate	-0.073829363	-0.28832529	-0.0367112	-0.22239842	-0.210287997	1.000000000	-0.1567492	-0.007905382
fineaggregate	-0.281549289	-0.28094497	0.1056053	-0.42186521	0.157383960	-0.156749179	1.000000000	-0.160562637
age	0.086439883	-0.04165378	-0.1581225	0.29773572	-0.209219553	-0.007905382	-0.1605626	1.000000000

위의 그림과 수치로 보면 관계들을 더 자세히 알 수 있다. 주황색으로 색이 짙어질수록 음의 상관관계가 높아짐을 의미하고 파란색이 짙어질수록 양의 상관관계가 높아짐을 의미한다. 이 데이터의 경우 시멘트부터~ fine aggregate까지는 콘크리트에 들어가는 재료의 양을 의미하므로 음의 상관관계가 높다는 것은 한 재료가 많이 들어가면 상응하는 다른 재료가 적게 들어간다는 의미가 될 것이다. 대표적으로 Water과 Superplaticizer는 아주 강한 음의 상관관계를 보여주는데 물이 많이 들어가면 Superplaticizer는 적게 들어가거나 그 반대의 경우를 생각해 볼 수 있다. 양의 상관관계가 그나마 높은 경우는 Flyash와 Superplaticizer이 될

수 있는데, 둘의 값이 비슷하게 증가하거나 줄어든다는 것을 의미한다. 그렇다면 Water과 Flyash의 관계는 어떨까? 강하진 않지만 당연히 음의 상관관계를 띠는 것을 알 수 있다.

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 만족될만한 수준인지 정성적으로 판단해 보시오.

```
call:
lm(formula = csMPa ~ ., data = insurance_trn_data)

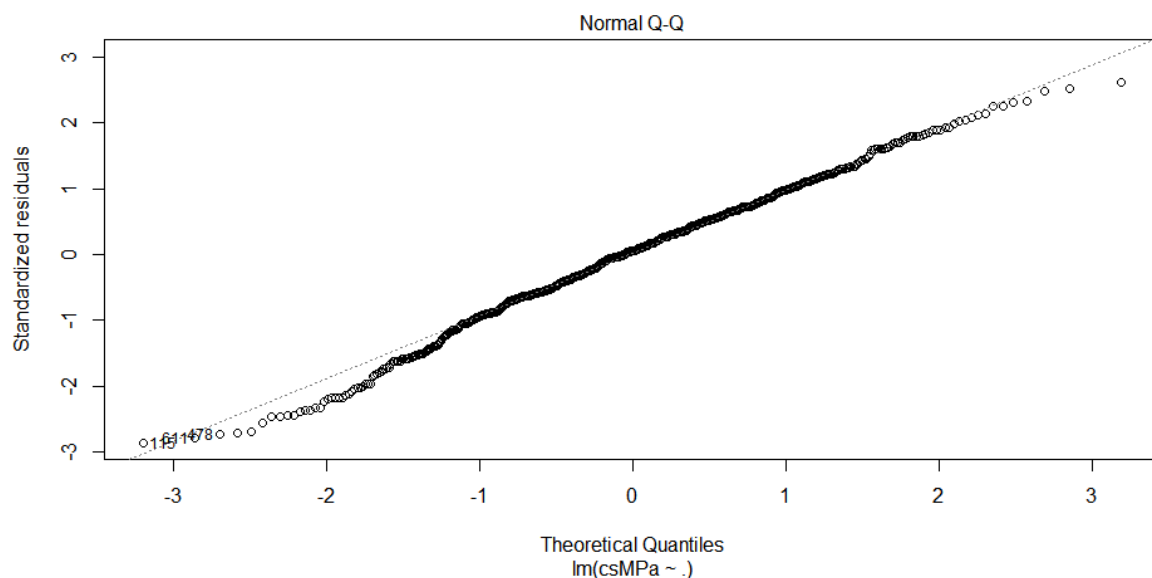
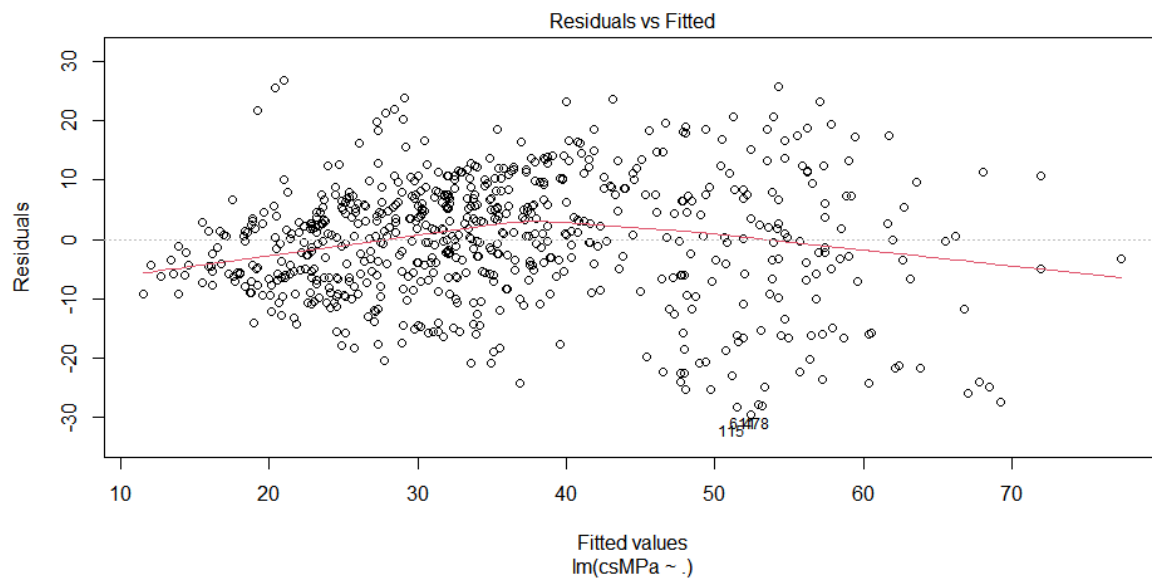
Residuals:
    Min       1Q   Median       3Q      Max
-29.4948  -6.3421   0.5396   6.8605  26.8454

Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept)  43.279597   33.986097   1.273 0.203285
cement       0.100135    0.010517   9.521 < 2e-16 ***
slag         0.084485    0.012638   6.685 4.75e-11 ***
flyash       0.056876    0.015635   3.638 0.000295 ***
water       -0.243363    0.053492  -4.550 6.35e-06 ***
superplasticizer 0.276339    0.130457   2.118 0.034509 *
coarseaggregate -0.006192    0.011599  -0.534 0.593658
fineaggregate -0.001542    0.013577  -0.114 0.909611
age          0.111204    0.006260  17.763 < 2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 10.32 on 694 degrees of freedom
Multiple R-squared:  0.6045,    Adjusted R-squared:    0.6
F-statistic: 132.6 on 8 and 694 DF,  p-value: < 2.2e-16
```

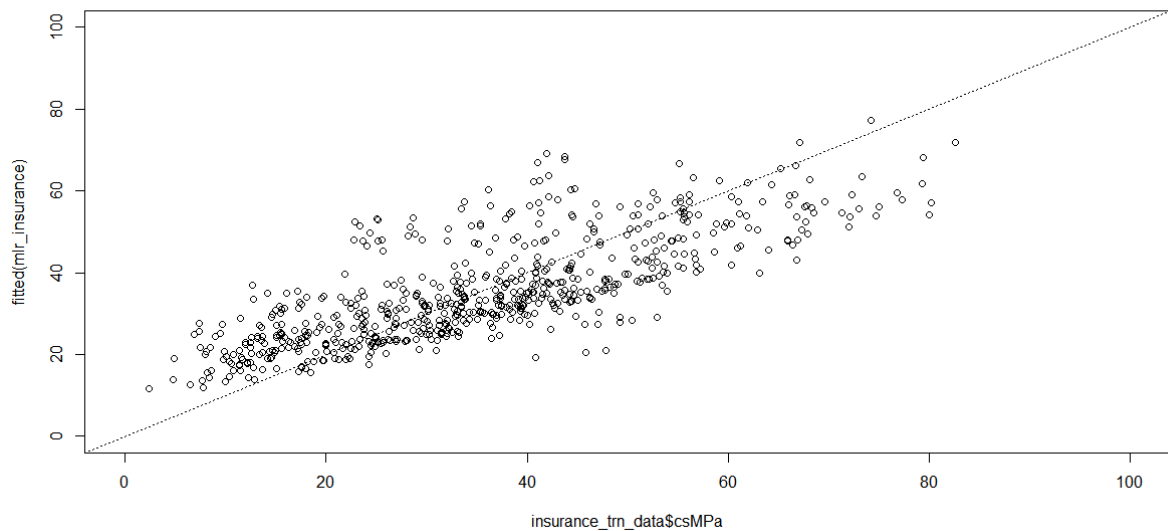
답) 현재 결정계수는 모형의 설명력을 의미한다. adjusted R-squared는 다변량 회귀 분석에서 독립변수가 유의하든 유의하지 않든 독립변수의 수가 많아지면 결정계수가 높아지는데 이를 보완하기 위한 계수이다. 그러한 수정된 결정계수의 값 0.6이 나왔다는 것은 대략 60% 정도의 설명력을 가진 모형이라는 의미가 된다. 선형성이 그닥 높지는 않은 데이터셋이라는 해석을 할 수 있다.

최소자승법의 가정 중에는 잔차들의 분포가 Homoscedasticity를 만족해야 한다는 것과 잔차가 정규분포를 따라야 한다는 가정이 있다. 이를 확인하기 위해 다음과 같은 플롯을 구현해보았다.



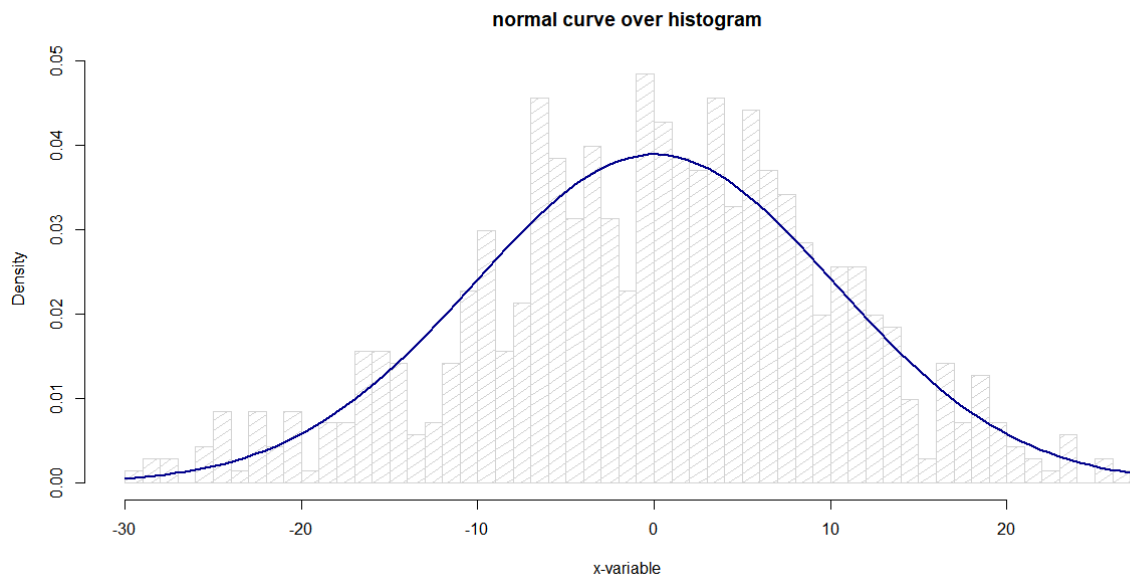
우선 첫번째 플롯을 보면 빨간선이 아래로 이차곡선의 형태를 띠는 것을 알 수 있으므로 Homoscedasticity를 만족하지 않을 가능성이 높다는 것을 알 수 있다. 반면 Q-Qplot을 보면 의외로 잔차들이 선에 근접하여 분포하여있다는 것을 느낄 수 있는데, 엄밀히 보자면 -1구간에서부터 조금씩 나가려는 경향은 있으나 그리 크게 벗어나지는 않는다는 점에서 정규분포를 따른다고 할 수 있을 것

같다.



위의 플롯 같은 경우는 정답과 예측한 데이터를 비교해보는 것인데, 선을 기준으로 점들이 아예 모여있지는 않지만 어느정도 모여있는 듯한 형태를 띄운다는 것을 알 수 있다. 이로써도 이 데이터셋이 완전히 선형성을 가지고 있지는 않지만 어느정도 선형성을 띄는 데이터라는 것을 알 수 있다.

정규분포에 대한 그래프도 히스토그램으로 그려보자면,



와 같이 나오는데 정성적으로 보자면 정규분포와 매우 흡사한 형태를 띄고 있다고 볼수 있다. 실제로 skewness=-0.2429693이 kurtosis=2.978154가 나온 것을 확인할 수 있었다.

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

```
call:
lm(formula = csMPa ~ ., data = insurance_trn_data)

Residuals:
    Min       1Q   Median       3Q      Max
-29.4948  -6.3421   0.5396   6.8605  26.8454

Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept)   43.279597   33.986097   1.273 0.203285
cement         0.100135    0.010517   9.521 < 2e-16 ***
slag           0.084485    0.012638   6.685 4.75e-11 ***
flyash         0.056876    0.015635   3.638 0.000295 ***
water        -0.243363    0.053492  -4.550 6.35e-06 ***
superplasticizer 0.276339    0.130457   2.118 0.034509 *
coarseaggregate -0.006192    0.011599  -0.534 0.593658
fineaggregate  -0.001542    0.013577  -0.114 0.909611
age            0.111204    0.006260  17.763 < 2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 10.32 on 694 degrees of freedom
Multiple R-squared:  0.6045,    Adjusted R-squared:  0.6
F-statistic: 132.6 on 8 and 694 DF,  p-value: < 2.2e-16
```

위에서 구축한 모형의 결과는 다음과 같았다. 유의수준이 0.01이라 함은 유의확률이 0부터 0.01 사이에 있는 변수들이 유의미한 변수들이 될 것이다. 현재 이 데이터셋의 경우, cement, slag, flyash, water, age가 유의미한 변수들이 될 것이다.

	cement	slag	flyash	water	age	csMPa
cement	1.00000000	-0.27934301	-0.38207361	-0.04125704	0.08779895	0.48717248
slag	-0.27934301	1.00000000	-0.32678543	0.10242357	-0.04784041	0.13810148
flyash	-0.38207361	-0.32678543	1.00000000	-0.29031114	-0.15310051	-0.08752292
water	-0.04125704	0.10242357	-0.29031114	1.00000000	0.29121921	-0.29021184
age	0.08779895	-0.04784041	-0.15310051	0.29121921	1.00000000	0.33405965
csMPa	0.48717248	0.13810148	-0.08752292	-0.29021184	0.33405965	1.00000000

위의 표에서 빨간 네모속에 있는 값들만을 보면 된다. 시멘트, slag, age와 같은 변수들이 양의 상관관계를 가지며 그중 시멘트가 가장 큰 양의 상관관계를 가진다. 반면, flyash, water은 음의 상관관계를 가지고 있는 것을 볼 수 있다.

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

	RMSE	MAE	MAPE
insurance	10.20337	8.196187	34.20614

답) MAE 는 8.196187이라는 결과가 나왔는데, 이는 평균적으로 $1m^3$ 의 혼합물 당 압축강도 오차가 MAE만큼 나왔다는 것이며, MAPE는 원래 그 혼합물의 압력강도 기준으로 34% 정도 상대적인 오차가 발생했다는 의미가 된다. RMSE 는 차이의 제곱을 평균해서 루트 씌운 값이 10.20337이 나왔다는 것을 의미한다.

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.

답) 우선 유의미한 변수들 중에서 먼저 데이터를 줄일 것이다. 우선 일차적으로 cement, slag, flyash, water, age 네가지의 입력변수로 줄어들게 된다. 첫번째 스텝으로 이러한 과정을 해야 겠다고 생각을 한 이유는 변수들 중에서 실제로 영향을 미치는 변수들을 이용해야 어떤 변화가 생길것이라고 생각했기 때문이다. 그 다음으로는 상관관계를 볼 것이다. 우선 입력 변수들끼리의 상관관계를 볼 것인데, 여기서 유의해야 할 점은 상관관계가 매우 높게 나오는 변수들은 하나의 변수가 다른 변수의 속성을 가지고 있다고 봐도 무방하기 때문에 하나로 줄여서 봐도 될것이다 라는 것이다. 하지만 현재 sorting된 입력 변수들간의 상관 관계를 보면 하나로 줄일만큼 큰 상관 관계를 가지고 있는 변수가 없다는 것을 알 수 있다. 그럼에도 원래 변수 수의 절반 이하로 입력 변수를 사용해야하므로 종소변수와의 상관관계를 봐야 한다. 이때 유의해서 볼점은 상관관계가 음이든 양이든 관계가 얼마큼 존재하는지이다. Flyash가 비교적 종속변수와의 상관관계가 0에 가까운 값을 가지며, 이는 다른 값들에 비해 미묘한 상관을 준다는 의미가 될 수 있다. 그러므로 최종적으로 Cement, Slag, Water, Age를 선택하기로 한다.

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

답)

```
> summary(mlr_insurance_remake2)

Call:
lm(formula = csMPa ~ ., data = insurance_remake2_trn_data)

Residuals:
    Min       1Q   Median       3Q      Max
-32.947  -7.841   0.499   7.627  28.882

Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept)  63.653731    4.142050   15.37  <2e-16 ***
cement        0.084558    0.004285   19.73  <2e-16 ***
slag          0.068606    0.005124   13.39  <2e-16 ***
water        -0.339254    0.021870  -15.51  <2e-16 ***
age           0.109296    0.006866   15.92  <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

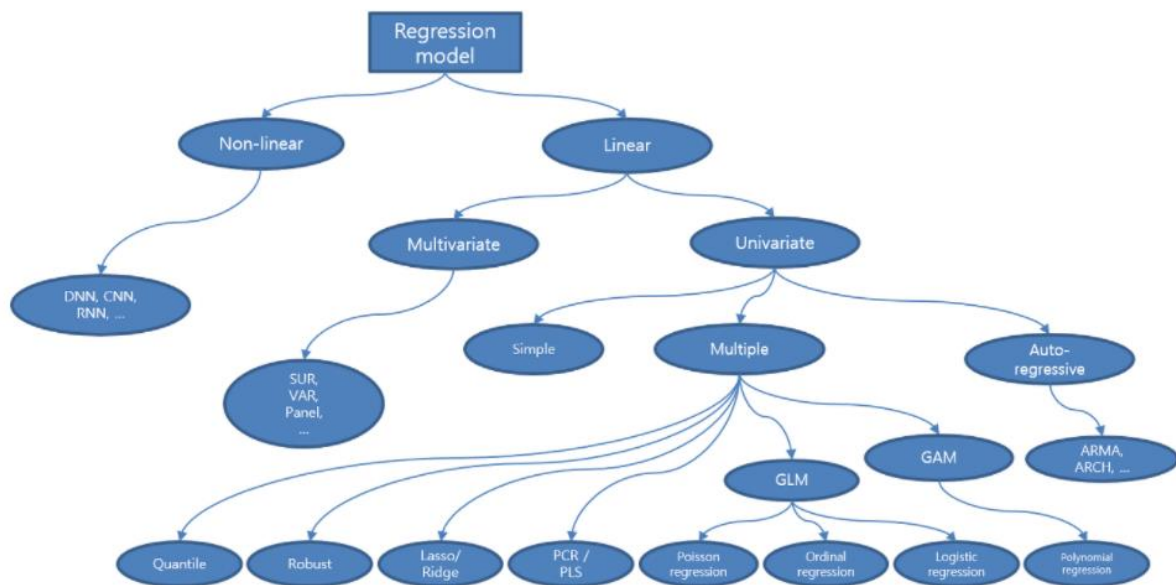
Residual standard error: 11.1 on 696 degrees of freedom
Multiple R-squared:  0.5507,    Adjusted R-squared:  0.5482
F-statistic: 213.3 on 4 and 696 DF,  p-value: < 2.2e-16
```

	RMSE	MAE	MAPE
바꾸기 전	10.20337	8.196187	34.20614
바꾼 후	10.47525	8.410546	36.17270

처음 결과를 접했을 때 엄청 당황스러웠다. 유의미성이 상당히 높다고 판단되는 변수들만 이용하여 모델을 새로 구축하였기 때문에, 높은 확률로 결정계수와 수정된 결정계수의 값이 둘다 증가하고 RMSE, MAE, MAPE의 값도 당연히 설명력이 높아져 낮아질거라고 예상하였기 때문이다. 결정계수의 값들이 바꾸기 전보다 바꾼후 더 줄어든 결정적인 이유는 변수의 개수가 줄었기 때문이라고 예상된다. 독립변수가 많을수록 결정계수의 값은 데이터의 유의미성과 관계없이 설명력이 높게 나오게 된다. 그러다 이러한 변수의 개수를 절반으로 줄인다는 것은 그만큼 설명력이 떨어질수 밖에 없음을 의미하는것이라고 생각된다. 또한 선택된 변수만을 가지고 모델을 구축한 이 선형회귀 모델에서는 결정계수의 값과 RMSE,

MAE, MAPE의 값을 보았을 때 선형성을 띄지 않을 가능성이 높다고 해석이 된다.

Extra question) 이 외 해당 데이터셋을 통해 MLR 관점에서 가능한 추가적인 분석을 웹에서 검색해서 수행하고 그 결과를 해석해 보시오.



<자료출처: <https://brunch.co.kr/@gimmesilver/38> >

현재 선형회귀모델을 구축해본 결과, 이 데이터는 독립변수와 종속변수간에 선형성을 그다지 높게 가지지 않는다는 결론을 내릴수 있었다. 현재 나의 데이터셋은 종속변수가 하나인 univariate이며 여러 독립변수가 있기 때문에 multiple이다. 이때 사용할 수 있다고 판단되는 분석기법에는 polynomial regression과 Genalized Additive Model이 있다. 여기서 GAM기법을 이용해보기로 하였다.

```
Call: gam(formula = csMPa ~ s(cement, df = 4) + s(slag, df = 4) + s(water,
  df = 4) + s(age, df = 4), data = insurance_remake2)
Deviance Residuals:
    Min       1Q   Median       3Q      Max
-22.5972  -4.2636  -0.4657   4.5732  27.7763
```

(Dispersion Parameter for gaussian family taken to be 48.9693)

```
Null Deviance: 278146.1 on 1000 degrees of freedom
Residual Deviance: 48185.77 on 983.9999 degrees of freedom
AIC: 6754.653
```

Number of Local Scoring Iterations: NA

Anova for Parametric Effects

	Df	Sum Sq	Mean Sq	F value	Pr(>F)
s(cement, df = 4)	1	71516	71516	1460.43	< 2.2e-16 ***
s(slag, df = 4)	1	25230	25230	515.22	< 2.2e-16 ***
s(water, df = 4)	1	14701	14701	300.21	< 2.2e-16 ***
s(age, df = 4)	1	32908	32908	672.02	< 2.2e-16 ***
Residuals	984	48186	49		

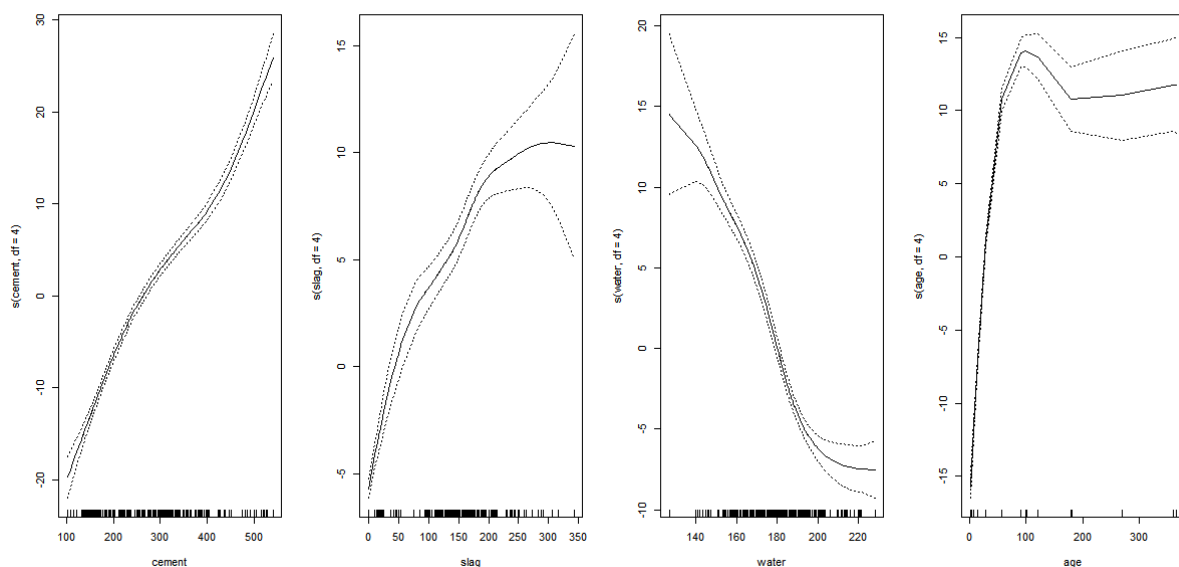
signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Anova for Nonparametric Effects

	Npar	Df	Npar F	Pr(F)
(Intercept)				
s(cement, df = 4)	3	15.18	1.163e-09	***
s(slag, df = 4)	3	14.39	3.500e-09	***
s(water, df = 4)	3	26.93	< 2.2e-16	***
s(age, df = 4)	3	422.21	< 2.2e-16	***

signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

다음과 같은 결과를 얻을 수 있었다. 이를 더 보기 쉬운 그래프의 형태로 나타내어 보았다.



이 때에 사용한 데이터는 선택된 변수들을 모아놓은 데이터셋을 이용하였다. 그래프를 보면 왜 선형 관계가 아니었는지에 대해 알 수 있다. 시멘트와 같은 경우는 더 많이 넣음에 따라 압축강도를 높여주는 나름의 선형관계를 가지고 있다. 반면 물과 같은 경우는 물을 많이 넣으면 넣을수록 압축강도가 내려가는 데, 이 또한 어찌보면 선형 관계를 나타낸다고 할 수 있다. 하지만 slag와 age의 경우를 보면 slag는 어느 일정 수준을 넣으면 압축강도가 되려 떨어지는 포인트가 존재하고 age같은 경우 시간이 지남에 따라 maximum경계선이 존재한다는 것을 깨달을 수 있었다. P-밸류의 값이 전부 낮다는 점에서 이는 충분히 영향을 미치는 변수들이라고 할 수 있으나 이러한 관계를 가지고 있기에 선형성을 띄지 않는다는 점을 알 수 있다.

결론을 말하자면 이 데이터셋은 혼합물에 대한 데이터셋이므로 화학 작용은 일정한 상태 변화를 일으키지 않기 때문에 선형성이 부족하다는 결론을 내릴 수 있을 것 같다.