

2. 선형회기분석

선형 회기 분석을 하려면, 아래의 가정을 따라야 한다.

① 독립변수(X)는 이름 그대로 독립적인 형태여야 한다. 변수들끼리 상관관계가 있다면 결과는 왜곡될 수밖에 없다. 예를 들면, 특정 변수와 다른 변수가 선형 관계라면, 비슷한 변수가 두번 학습하게 되므로 정확한 결과를 기대하기 힘들다. 따라서 상관관계가 있는 변수는 제거해 줘야 한다.

② ^{2번씩 확인} ~~변수~~들은 정규분포를 가진다는 가정이 전제되어야 한다. 정규분포를 가지지 않고 한쪽으로 치우친다면 당연히 정확한 결과를 기대하기가 힘들다. 따라서 정규성을 가지지 않는 변수는 정규성을 가질 수 있도록 로그나 지수의 방법으로 치환하여 범위를 늘리거나 줄이는 방식으로 정규분포를 만드는 것도 한 가지 방법이라고 할 수 있다.

③ 종속변수(Y)와 독립변수(X)는 선형 상관관계를 가지는 것을 가정한다. 선형이 아니라면 다른 회귀분석 방법을 사용하거나 기존의 변수를 변경시켜 줘야 한다.

선형 회귀분석을 통해 좋은 모델을 만들기 위해서는 분석 데이터가 아래 4가지 기본가정을 만족해야 한다.

아래 4가지 기본가정을 만족하지 않으면 제대로 된 선형 회귀모델이 생성될 수 없다.

(1) 선형성 : 각 독립변수와 종속변수의 선형성.

(2) 독립성 : 독립변수들 간의 독립성 (독립변수들 간에 상관관계가 없는 것)

(3) 등분산성 : 잔차의 등분산성

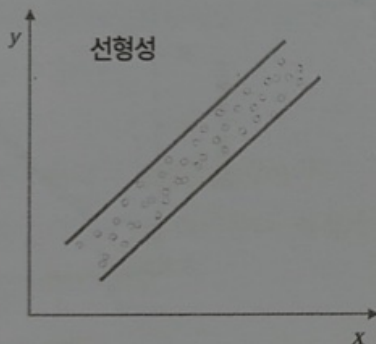
(4) 정규성 : 잔차의 정규성

→ 잔차의 분산 : 잔차의 평균으로부터
잔차들이 떨어져있는 정도

① 선형성

라. 그래프를 활용한 선형회귀분석의 가정 검토

1) 선형성

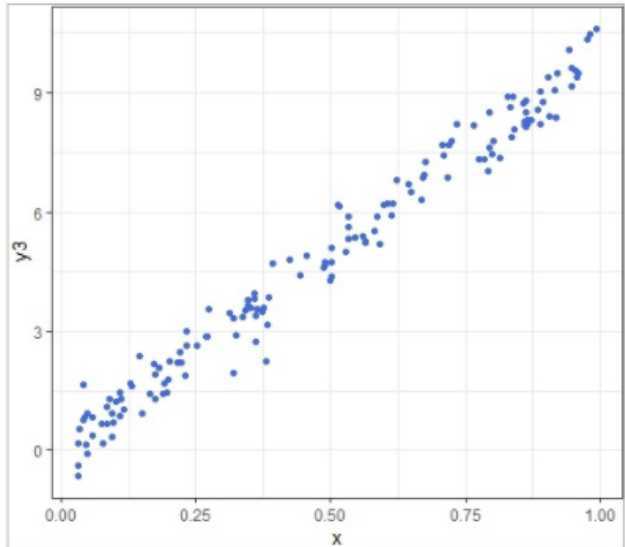
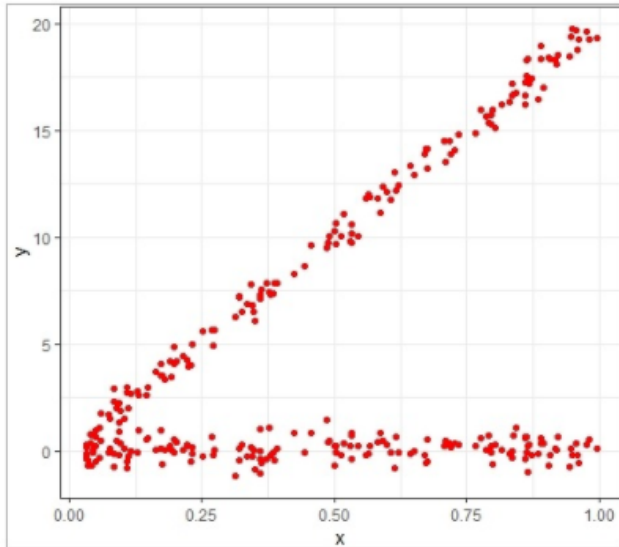


- 선형회귀모형에서는 왼쪽의 그래프와 같이 설명 변수(x)와 반응변수(y)가 선형적 관계에 있음이 전제되어야 한다.

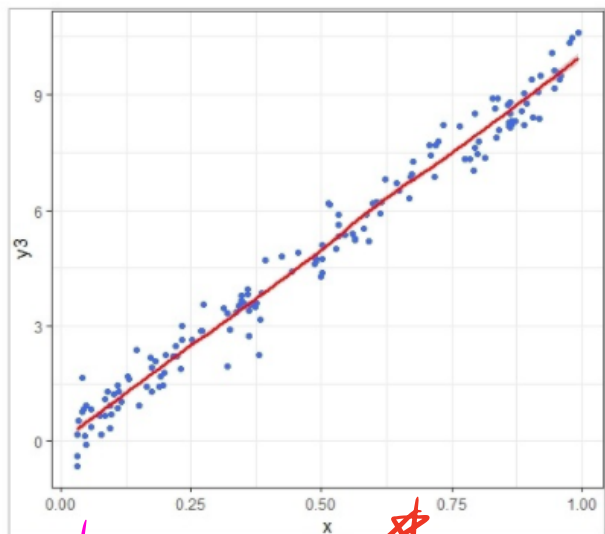
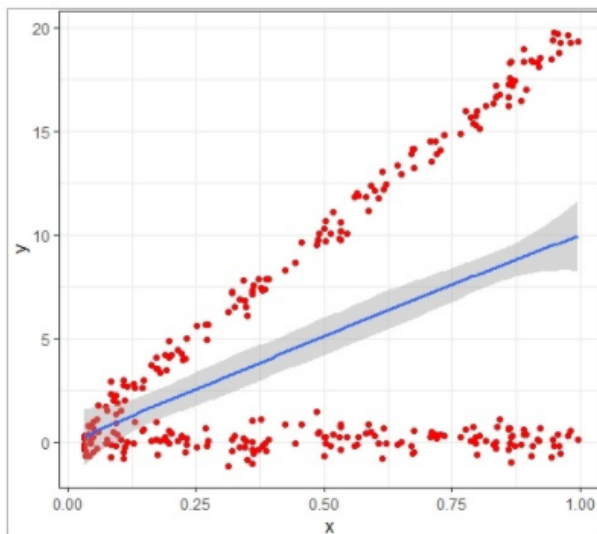
등분산성

등분산성은 회귀분석에서 매우 중요한 가정 중 하나입니다. 여기서 등분산의 주체는 오차입니다. 그렇지만 실제 오차를 정량화 할 수 없으니 오차의 추정치로써 잔차를 사용하게 됩니다. 잔차는 추정된 회귀선과 실제 값의 차이입니다. 즉, 등분산성을 보는 것은 선과 점 사이의 거리가 패턴이 없이 일정한가를 보는 것과 같습니다.

이를 확인하기 위해 조금 극단적인 예를 보겠습니다.

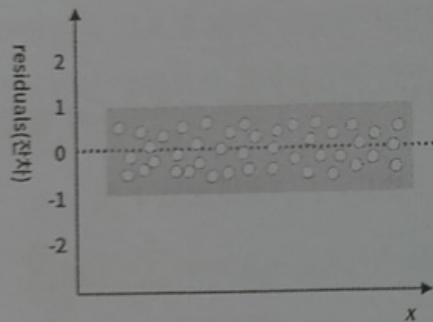


좌측 산점도와 우측 산점도가 있을 때, 두 산점도에 회귀선을 적합시켜보도록 하겠습니다.



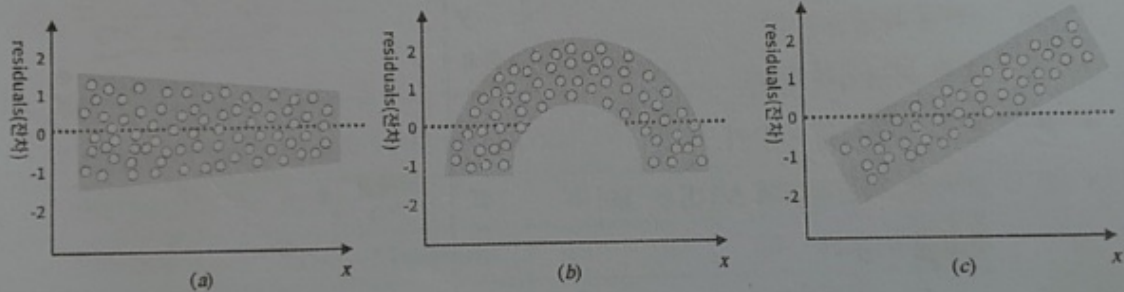
우측의 회귀선은 직관적으로 판단해도 회귀선에 문제가 없습니다. 하지만 좌측 회귀선은 그러지 못합니다. 그 이유는 회귀선과 데이터의 차이(잔차)가 x 가 커지면서 같이 늘어나고 있기 때문입니다. 즉, 등분산성을 만족한다고 할 수 없습니다. 이 때는, 회귀선이 데이터를 잘 설명한다고 보기 어렵습니다. 이처럼 회귀분석에서 등분산성이 위배되면 회귀분석은 데이터를 설명하지 못한다고 판단하기 때문에 좋은 회귀선이라고 할 수 없습니다. 좌측의 회귀선에 대한 잔차의 등분산 진단 그래프로 보면 다음과 같은 플롯을 확인할 수 있습니다.

가) 등분산성을 만족하는 경우



- 설명변수(x)에 대한 잔차의 산점도를 그렸을 때, 왼쪽의 그림과 같이 설명변수(x) 값에 관계없이 잔차들의 변동성(분산)이 일정한 형태를 보이면 선형회귀분석의 가정 중 등분산성을 만족한다고 볼 수 있다.

나) 등분산성을 만족하지 못하는 경우

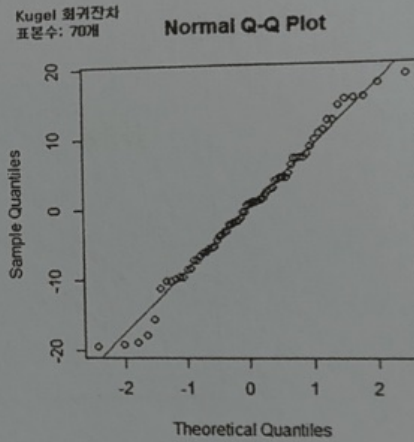


- (a) : 설명변수(x)가 커질수록 잔차의 분산이 줄어드는 이분산의 형태
- (b) : 2차항 설명변수가 필요
- (c) : 새로운 설명변수가 필요

③ 정규성

3) 정규성

- Q-Q Plot을 출력했을 때, 오른쪽의 그림과 같이 잔차가 대각방향의 직선의 형태를 지니고 있으면 잔차는 정규분포를 따른다고 할 수 있다.



단순 선형회귀분석에선
독립변수와 종속변수간의 선형성만 파악해도
회귀분석의 가정 만족 여부를 판단할 수 있다.

5. 가정에 대한 검증

- 단순 선형회귀분석: 입력변수와 출력변수간의 선형성을 점검하기 위해 산점도를 확인
- 다중 선형회귀분석: 회귀분석의 가정인 [선형성, 등분산성, 독립성, 정규성(정상성)]이 모두 만족하는지 확인

