# Week 1 : 회귀분석의 기초

안녕하세요 여러분 :)

2022-1학기 회귀분석팀장 조수미입니다!!

클린업 1주차가 벌써 시작이네요. 화잇팅!!!

그럼 이제 회귀의 매력에 빠져볼까요?

### < 목차 >

**0. 기본 수식**

* 평균, 분산, 공분산, 상관계수, 행렬 표기

**1. 회귀분석이란?**

* 회귀분석과 회귀식
* 회귀 모델링의 과정

**2. 단순선형회귀**

* 다중선형회귀란?
* 모수의 추정 (LSE)
* 적합도와 유의성 검정

**3. 다중선형회귀**

* 단순선형회귀란?
* 모수의 추정 (LSE)
* 적합도와 유의성 검정

**4. 데이터 진단**

* 이상치, 지렛값, 영향점

**5. 로버스트 회귀**

* 로버스트 방법과 비용

## 0. 기본 수식

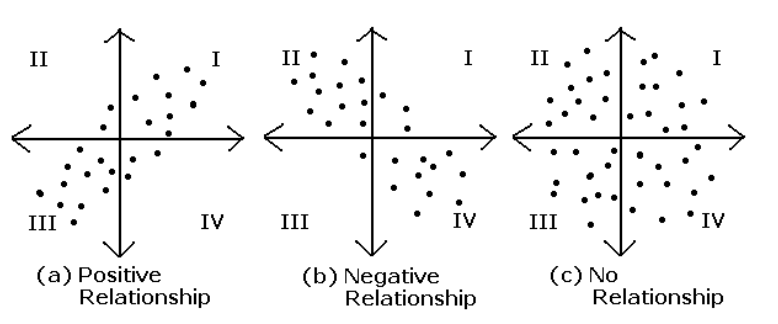
### 1) 기초수식

* 표본 평균 (Sample Mean)
* 표본 분산 (Sample Variance)
* 표본 표준편차 (Sample Standard Deviation)
* 변동 :

### 2) 공분산 (Covariance)

두 개의 확률변수의 상관 정도를 나타내는 지표. 두 변수의 선형관계의 음 또는 양의 방향성만 나타낼 뿐, 얼마나 선형성을 갖는지 즉 ‘강도’는 표현하지 못한다.

* 공식
* 성질



(a) 공분산>0 (b) 공분산<0 (c) 공분산=0

Cov(X, Y) > 0 : X가 증가 할 때 Y도 증가한다.

Cov(X, Y) < 0  : X가 증가 할 때 Y는 감소한다.

Cov(X, Y) = 0  : X, Y 두 변수 간에 아무런 선형 상관관계가 없다.

* 단점
* 확률변수 X, Y의 측정 단위의 크기에 영향을 받기 때문에 상관성의 형태에 대해서는 나타낼 수 있지만, 정확한 정도를 표현하는 데에는 한계가 있다.
* ex) 10점 만점인 두 과목의 상관성이 100점 만점인 두 과목보다 더 크더라도, 공분산은 100점 만점인 두 과목의 공분산이 더 큰 값이 나올 수 있다.
* → 단순히 공분산이 더 크다고 해서 선형관계가 강한 것은 아니다!
* 이를 보완하기 위해 \*\*상관계수(Correlation)\*\*가 존재한다.

### 3) 상관계수 (Correlation Coefficient)

확률변수의 절대적 크기에 영향을 받지 않도록 단위화 시킨 공분산의 단점을 보완할 수 있는 ‘표준화된 공분산’이다.

* 공식
* 상관계수의 해석
* 상관계수는 두 확률 변수의 선형 상관관계를 나타내줄 뿐만 아니라 선형적인 상관성의 크기까지 파악할 수 있는 지표이다. -1부터 1의 값을 갖는다. 확률변수 X, Y가 독립이라면 상관계수는 0이고, 상관계수가 0이면 두 변수는 아무런 **선형관계가 없다.** 단, 선형관계 이외의 비선형 관계가 있을 수 있다.
* 텍스트이(가) 표시된 사진

  자동 생성된 설명

1이면 완전한 상향 직선, -1이면 완전한 하향 직선의 형태를 띈다.

## 1. 회귀분석이란?

### 1) 회귀분석

* 독립변수와 종속변수 간의 관계를 설명하고 모델링하는 통계적 기법
* 두 변수 이상의 변수들 간의 상관관계를 파악하고, 이를 통해 특정 변수의 값을 다른 변수들을 이용하여 설명하고 예측하는 방법
* ex. 환자 혈압 수치와 콜레스트롤 수치와의 관계, 대기오염과 사망률과의 관계
* 지도학습의 한 종류

단순회귀분석 : 한 개의 종속변수와 한 개의 독립변수 사이의 관계를 분석하는 경우

다중회귀분석 : 한 개의 종속변수와 여러 개의 독립변수 사이의 관계를 분석하는 경우

* 회귀분석의 목적
  1. 변수들 간의 관계에 대한 표현
  2. 독립변수에 따른 종속변수의 변화
  3. 미래 관측값에 대한 예측

### 2) 회귀식

우리가 관심있는 종속변수 Y와 독립변수 X의 관계를 함수식(f)으로 표현한 것으로 회귀식의 구성 요소는 크게 세 가지로 구분할 수 있다.

* **독립변수**(independent variable) : 종속변수를 설명하기 위한 변수로 설명 변수(Explanatory variable) 혹은 예측 변수(Predictor)라고도 함
* **종속변수**(dependent variable) : 독립변수에 의해서 설명되는 변수로 반응변수(Response variable)로도 불림
* **오차항**(error term) : 변수를 측정할 때 발생할 수 있는 오차로 설명할 수 없는 무작위성을 가짐

### 3) 회귀 모델링 과정 (예시)

1. 문제 정의
   * 학점을 가장 잘 표현할 수 있는 변수들은 무엇이 있을까?
2. 적절한 변수 선택
   * '통학 거리', '공부 시간', 'SNS 이용 시간', '수강 학점' 등이 영향을 주지 않을까?
3. 데이터 수집 및 전처리
   * 학생 개개인의 학점, 통학 거리, 공부 시간, SNS 이용 시간, 수강학점 데이터 수집
4. 모델 설정과 적합
   * 적절한 회귀분석 모델 선정
   * 선형 vs 비선형 / 단순회귀 vs 다중회귀 / 모수 vs 비모수/ 일변량 vs 다변량 등 고려
5. 모형 평가
   * 설정한 모델이 회귀 가정을 만족하는가? 만족하지 않는다면 수정 → 2주차에서 배워요!☺️
   * 변수들은 유의한가?
6. 모형 해석
   * “현재 상태에서 학점을 3학점을 덜 듣고 SNS 이용시간을 1시간 줄이면서, 현재 주거지에서 통학을 할 때 학점은 평균적으로 0.5 정도 오를 것이다.”

# 2. 단순선형회귀

### 1) 단순선형회귀 (Simple Linear Regression)

단순선형회귀는 하나의 종속변수(Y)와 하나의 독립변수(X)만을 가지며 **두 변수의 관계를 가장 잘 표현하는 직선을 추정하는 것이 목적**이다.

지도, 다른, 묶음, 보트이(가) 표시된 사진

자동 생성된 설명

* **회귀 모델**
* 우리는 변수 간의 실제 관계를 표현하는 함수식을 알 수 없다. 그렇기 때문에 "변수의 관계는 아마도 이럴 것이다" 하는 가정 하에 아래와 같은 직선 함수 식을 가정하게 된다. 여기서의 가정은 선형성 가정이다. → 2주차에서 배워요!☺️
* 설명

,

- : 종속변수 y의 i번째 관측값

- : 독립변수 x의 i번째 관측값

- : i번째 관측값에 의한 랜덤 오차

이 때, 정규분포 가정

-: 회귀계수 또는 우리가 추정해야 할 모수

더 좋은 모델을 만들기 위해서는 회귀계수를 잘 추정해야 한다

→ 이러한 방법을 모수적(parametric) 방법이라고 한다.

* 해석
* x가 한 단위 증가할 때, y는 평균적으로 만큼 증가함
* **왜 직선인가?**
* 변수의 영향력을 간단하게 모형화 할 수 있기 때문이다. 위의 그림처럼 2차원 평면 위의 직선은X와 Y의 일대일대응 관계를 통해 독립변수의 변화에 따른 종속변수의 변화 또한 직관적으로 이해할 수 있기 때문에 가장 단순한 형태의 직선을 사용한다.

텍스트, 시계이(가) 표시된 사진

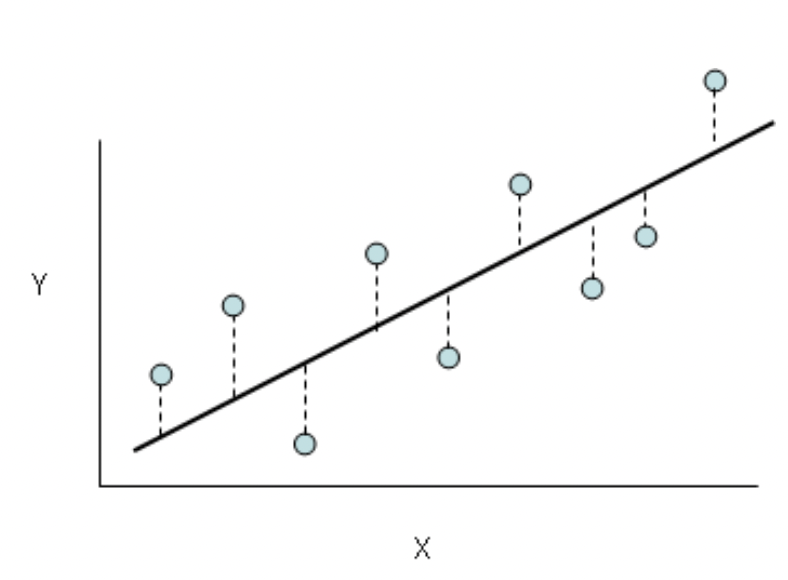
자동 생성된 설명

* 가장 오른쪽 그림처럼 고차함수로 추정을 할 경우 모델의 복잡도가 높아져 과적합(overfitting)의 원인이 될 수 있다. 과적합 문제는 모델의 분산을 높이고, 검증 데이터의 예측 성능을 저하시킨다. →자세한 내용은 데이터마이닝팀 클린업 1주차 참고!
* 현실의 많은 데이터는 선형적으로 생성되지 않기 때문에 단순한 선형 회귀식은 예측 성능이 떨어진다. → local regression, smoothing spline, GAM(Generalized Additive Model)등의 비선형 모델 사용

### 2) 모수의 추정 - 최소제곱법 (Least Square Estimation Method)

앞서 설명했듯이 우리는 실제 모델을 알 수 없기 때문에 모델을 하나 가정했다. 그 형태는

이다. 분석하고자 하는 회귀모형을 결정한 후에는 모형에 포함된 모수 을 추정해야 한다. 이 **모수들을 잘 추정하는 것은 곧 좋은 회귀 모델을 만든다는 의미**이다. 그렇다면 어떤 추정이 좋은 추정일까? 직관적으로 우리가 만들어낼 회귀 직선과 관측치 사이의 오차가 작으면 작을수록 좋은 추정이다. 이러한 아이디어를 이용한 방법이 바로 최소제곱법!

* 최소제곱법(LSE)
* 위 직선에서 각 점으로부터 구하고자 하는 최적 직선까지의 수직거리가 오차이며, 이러한 오차제곱합이 최소가 되게 하는 직선을 찾는 것이다!
* 
* 즉, 점선의 제곱합을 최소화하는 방법

선형대수학 관점에서 최소제곱법을 이해해본다면, 회귀추정량 은 반응벡터 Y를 X가 생성하는 공간에 투영(projection)하여 구해짐을 나타낸다.→ 자세한 내용은 선형대수학팀 클린업 참고!

* **최소제곱법을 이용한 모수 추정**

위의 식에서 함수 가 바로 오차제곱합이고, 이 오차제곱합을 최소화시키는 을 찾는 것이 목적이다.

* 위 함수  
  는 **아래로 볼록한 이차함수**이기 때문에 항상 최소값을 가지며, 각각의 모수를 편미분하여 ‘미분값=0’ 을 만족시키는 값이 우리가 구할 추정량(estimator)이 된다. 이를 식으로 표현하면 아래와 같다.
  + 텍스트, 손목시계이(가) 표시된 사진

    자동 생성된 설명
* 이러한 추정법을 통해 얻은 추정치 을 최소제곱추정치(Least Square Estimator)라고 한다. 이를 정리하여 아래와 같이 공식화 할 수 있다.
* **왜 오차제곱합을 최소화할까?**
  + 미분이 편리하기 때문이다. ‘미분값=0’을 통해 바로 추정할 수 있다.
* → 오차의 절대값을 사용할 수도 있지만, 오차의 절대값을 목적함수로 사용하게 된다면 미분이 불가능한 점이 존재하여 수치적 방식을 사용하게 되고 계산이 오래걸린다.
* 
* 오차제곱합 오차의 절대값
  + 오차가 클수록 더 큰 패널티를 부여할 수 있기 때문이다. 오차제곱합 그래프의 경우 중심에서 멀어질수록 기울기가 상승하고 있죠? 따라서 인정할 수 있는 정도의 오차의 영향력은 줄이고, 인정할 수 없는 오차의 영향력은 높이는 방법.
* **BLUE(Best Linear Unbiased Estimator)**
* LSE를 이용한 추정은 아무런 조건(가정) 없이 사용 가능하다는 장점이 있지만, 다음 세 가지 조건이 갖춰질 경우 매우 유용한 성질을 갖는다.
  1. 오차들의 평균은 0
  2. 오차들의 분산은 로 동일(등분산)
  3. 오차 간의 자기상관성 없음 (uncorrelated)

위 세 가지 조건이 만족된다면 가우스-마코프 정리에 의해 LSE는 Best(분산이 제일 작은) Linear(선형) Unbiased Estimator(불편추정량)가 된다! 즉, 선형불편추정량 중에서 분산이 항상 작다는 것이고, 분산이 작다는 것은 추정량이 안정적이라는 의미이므로 좋은 추정량이라고 할 수 있다.

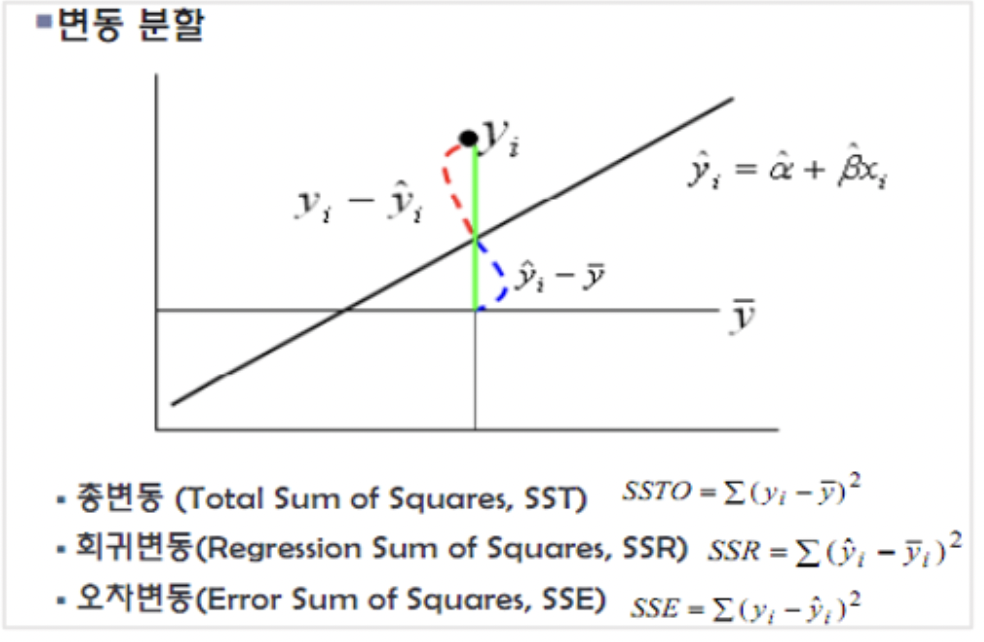
* **최소제곱추정량(LSE) VS 최대가능도추정량(MLE) 비교**
  + MLE : 확률적인 방법에 근거해, 우리의 데이터가 나올 ‘**가능도**'를 최대로 하는 모수를 선택하는 방법
  + ML 방법은 언제나 관측치가 iid라는 가정이 필요
  + 오차의 정규분포 가정이 있다면 MLE와 LSE는 완전히 동일한 추정량 가진다.

### 3) 적합성 검정 (Goodness of fit)

앞서 우리는 단순회귀모형이 무엇인지 배우고, 최소제곱법을 이용하여 단순회귀모형의 모수)를 추정하는 방법을 배웠다. 이렇게 추정한 모수를 우리는 회귀 계수(regression coefficient)라고 부르며, 추정된 회귀 계수를 바탕으로 회귀 직선을 만들 수 있다.

회귀모형을 추정한 후에는 이 회귀직선이 데이터에 얼마나 잘 들어맞는지, 모형에 대한 적합성에 대한 평가를 해야 한다. 이러한 과정을 적합성 검정이라고 하며, 적합성 검정을 위해 잔차라는 개념이 등장한다.

* **잔차(Residual)**
* 추정한 회귀 계수)를 이용해 회귀 직선을 만들었을 때, 오차의 추정량. 결과적으로 오차와 잔차는 다를 바 없지만, 모집단(오차)과 표본(잔차)의 차이이다. 모집단으로부터의 실제 오차 는 처럼 실제 값을 알 수 없기 때문에 추정된 직선을 통해 추정하는 것이다.
* **공식**
* **적합성 검정**
  + SST(Total Sum of Squares, 총 변동) :
  + SSR(Regression Sum of Square, 회귀선이 설명하는 변동) :
  + SSE(Residual Sum of Square, 잔차제곱합, 회귀선이 설명하지 못하는 변동) :
  + SST = SSR + SSE



* + 결정계수 () : 총 변동(SST)에서 회귀식이 설명할 수 있는 비율(SSR). 즉, Y가 X에 의해 설명되는 비율로, 1에 가까울수록 좋다.
  + 잔차와 연관지어 본다면, 잔차제곱합(SSE)은 회귀식이 설명할 수 없는 실제값과 추정값 사이의 오차들로, 총 변동 대비 잔차제곱합이 차지하는 비율이 작을수록 좋다.
  + 위의 각 변동 SST, SSE, SSR 의 수식적 의미는 곧 배울 다중선형회귀에서의 유의성 검정을 이해하는 데에 큰 도움을 주므로, 꼭 이해하고 넘어가자!

### 4) 유의성 검정

다음으로는 전체 회귀식이 아닌 개별 모수의 추정량이 통계적으로 유의한지를 알아보아야 한다.

라는 오차의 정규분포 가정 하에 개별 회귀 계수에 대해 다음과 같은 과정으로 통계적 검정을 할 수 있다.

1. 가설 설정 :
2. 추정량의 분포 :
3. 검정 통계량 :
4. 임계값 :
5. 검정(양측) :

* 도 동일한 방법으로 검정하면 된다.
* 귀무가설을 기각하지 못해도, X와 Y사이에 선형적 관계가 없을 뿐, 아무 의미가 없다는 게 아니다!

# 3. 다중선형회귀

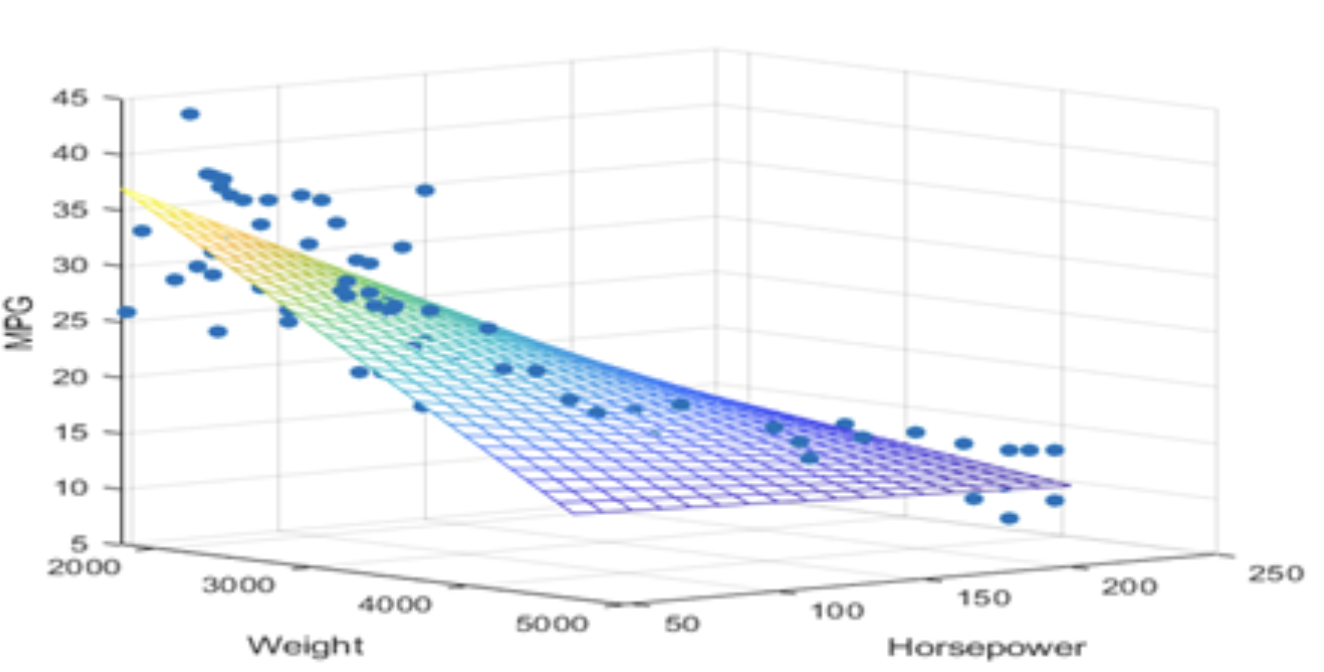
### 1) 다중선형회귀

독립변수 X가 2개 이상인 경우

* 독립변수가 한 개였던 단순선형회귀에서 독립변수 개수만 여러 개로 늘어난 경우이다. 독립변수 한 개와 종속변수 한 개의 단편적인 관계에서 나아가 더욱 복잡한 관계를 설명할 수 있어 자연현상, 사회현상을 파악하기에 용이하다.
* 다중회귀모형은 다차원에 대한 표현을 해야 하므로 벡터와 행렬로 표현하게 된다. 또한 다중회귀모형에서는 독립변수가 여러 개이므로 단순회귀모형과는 달리 독립변수들 간의 관계도 고려되어야 한다.
* 공식

을 가정

* 전체 독립변수의 개수는 p개, 회귀계수는 p+1개 (포함)
* 해석
  + **를 제외한 나머지 X변수들을 고정시킨 상태**에서 가 한 단위 증가할 때 y가 만큼 증가한다. 즉, 이미 다른 변수들이 설명하는 부분 이외에 나머지를 가 설명하는 것이다.
  + 다른 변수들이 고정되어 있다는 것은 다른 변수들이 상수처럼 취급된다는 것! 다른 변수들을 상수화하고 이 상태에서 우리가 관심있는 변수의 기울기를 확인하는 형태이다.
* 다중선형회귀는 데이터를 설명하는 초평면(hyperplane)을 찾는다. 초평면이란 n차원 공간에서의 n-1차원 평면을 의미하며, 2차원 공간에서의 초평면은 1차원 직선이고 3차원 공간에서의 초평면은 2차원 평면임을 생각하면 이해하기 쉽다.



설명변수가 두 개인 경우(위의 사진)라면 평면을 찾고, 세 개라면 공간을 찾음

### 2) 모수의 추정 - 최소제곱법(LSE)

기존 단순선형회귀에서는 추정해야 할 모수가 두 개였기에 각각 편미분 해서 연립방정식을 풀어줬다. 하지만 다중선형회귀에서는 추정해야 할 모수가 $p+1$개이기 때문에 편미분을 통해 구할 경우 매우 복잡해 행렬식을 이용한다. 단순선형회귀와 동일하게 최소제곱법을 이용하며, 정규분포 가정이 있다면 최대가능도추정법도 같은 추정량을 도출한다.

* **최소제곱법을 이용한 모수 추정**
  + 회귀식

* 목적함수를 에 대해 미분하여 미분값 = 0으로 두고 의 추정량 을 구할 수 있다.
  + 추정량
  + 추정된 회귀식
* \*\* 여기서 은 투영 행렬(projection matrix)
* **회귀계수 해석**
* 다중선형회귀의 추정된 회귀 계수는 해석에 매우 유의해야 한다! 회귀 계수 의 의미는 를 제외한 나머지 X변수들을 고정시킨 태에서 를 한 단위 증가시킬 때의 y의 증가량이다. 즉 개별 추정량의 값은 다른 모든 변수들이 고정된 상수값일 때 그 의미를 갖는다.

### 3) 유의성 검정

단순선형회귀와 마찬가지로 최소제곱법에 의해 추정된 p+1개의 회귀 계수 값이 통계적으로 유의한지 통계적 검정이 필요하다.

* **F-test (전체 회귀 계수에 대한 검정)**

회귀분석 문제에 있어서 일반적으로 가장 먼저 시행하는 검정이다. 이 검정이 유의하지 않다면 모델을 다시 세우거나 하는 등의 다른 조치가 필요하다.

1. **가설 설정**

* (모든 회귀 계수는 0이다)
* (적어도 한 개의 회귀 계수는 0이 아니다)

1. **검정통계량**

* MSR: 평균회귀제곱, MSE: 평균오차제곱
* **검정통계량의 의미**
* 단순선형회귀의 적합성 검정에서 SST, SSR, SSE의 의미를 알아보았다. 이를 유의성 검정에도 적용시켜 검정통계량이 갖는 의미를 이해할 수 있다.
* 귀무가설을 기각시키기 위해서는 검정통계량 가 충분히 커야한다. 가 크다는 것은 분모(SSE, 추정된 회귀식이 설명하지 못하는 부분)에 비해 분자(SSR, 추정된 회귀식이 설명하는 부분)보다 꽤나 크다는 의미이다(자유도를 나눠주지만 이는 제외하고 생각하자). 즉, 추정된 회귀식의 설명력이 높다는 것이며, 이는 들이 유의미 할수록 설명력이 높을 것이다라고 직관적으로 생각할 수 있다. 자연스럽게 유의미한 는 큰 SSR을 만들어내고 귀무가설을 기각시킨다. 요약해서 말하자면, 검정통계량 는 회귀식의 전반적인 계수가 얼마나 설명력을 갖는지를 의미한다.

1. **임계값**

* : 검정통계량이 임계값보다 크다면 귀무가설을 기각할 수 있다.
* 우리가 일반적으로 R에서 summary를 통해 확인하는 F-statistic
* 텍스트이(가) 표시된 사진

  자동 생성된 설명
* **Partial F-test (일반화된 F-test, 일부 회귀 계수에 대한 검정)**
* F-test의 일반화된 경우라고 생각하면 된다. 앞서 배웠던 F-test는 다중회귀모델 내의 모든 회귀 계수에 대해 한번에 유의성을 검정하지만 Partial F-test는 전체 계수 중 일부에 대해서만 유의성을 검정한다. 귀무가설을 베타 계수들 = 상수(주로 0)으로 두고 검정을 한다.
  1. **가설 설정**
     + Full Model (FM) : 기존 모든 변수를 사용한 다중회귀모형
     + Reduced Model (RM) : 일부 회귀 계수를 특정한 값(보통 0)으로 두는 축소모형
     + (RM이 맞다)
     + (FM이 맞다. 적어도 한 개의 회귀계수는 0이 아니다)
  2. **검정통계량**
  + 검정통계량의 의미
  + F-test에서 검정통계량이 갖는 의미와 맥락을 같이 한다. Partial F-test 에서 귀무가설을 기각시키기 위해서는 검정통계량 가 충분히 커야한다. 가 충분히 커지기 위해서는 이 보다 충분히 커져야 한다. 변수가 제거된다면 당연히 는 커진다. 하지만 제거된 변수가 의미 있다면, 는 매우 커질 것이며, 가 귀무가설을 기각시킬 만큼 충분히 커질 것이다.
  + 즉, 기본적으로 > 이지만, >>>>> 이 되어야 한다.
  1. **임계값**
  + 검정통계량이 임계값보다 크다면 귀무가설을 기각할 수 있다.
  + Partial F-test가 좀 더 일반화된 검정이지만, 보편적으로 회귀식 전체에 대한 검정을 사용한다.
* **t-test (개별 회귀 계수의 유의성 검정)**
* 회귀 모델에 포함된 여러개의 회귀 계수가 아닌, 개별 회귀 계수가 유의한지를 판단하기 위해 사용한다.
  1. **가설 설정**
     + 귀무가설 (다른 변수들이 다 적합된 상태에서 는 통계적으로 유의하지 않다.)
     + 대립가설 (다른 변수들이 다 적합된 상태에서 는 통계적으로 유의하다.)
  2. **검정통계량**
  3. **임계값**
  + 해석
* 다시 한 번 강조하자면, t-test는 해당 변수 자체가 유의미한지를 확인하는 것이 아니라, 다른 변수들이 다 적합된 상태에서 를 추가적으로 적합했을 때 이것이 유의미한 회귀식의 설명력 증가를 가져오는지를 확인하는 것이다! 따라서 회귀분석에서 개별 변수에 대한 검정은 한 변수에 대한 Partial F-test와 완전히 동일하다. 그렇기 때문에 t-test로 변수 선택을 하는 것은 매우 위험하다. 다른 변수들이 이미 고정되어 있을 때의 유의성이기 때문에, 다른 회귀식을 가정하면 해당 변수의 유의성이 바뀔 수 있다. → (변수 선택법은 3주차에서 배워요!😊)
* 텍스트이(가) 표시된 사진

  자동 생성된 설명
* 앞에서 언급한 예시를 다시 보자. 위 결과는 중고차 판매 데이터를 가지고 선형 회귀를 적합시킨 R 결과이다.
* 여기서 F-test 값을 확인했을 때, 회귀식에 대한 귀무가설을 기각할 수 있다. 이후 개별 변수에 대한 t-test 결과를 보면, TV와 Radio 변수는 해당 회귀계수가 0 ( '') 이라는 귀무가설을 기각하고, Newpaper 변수는 기각하지 못한다. 이를 해석해보면,
  + 해석
* ‘다른 변수들(Radio, Newspaper)가 적합된 상황에서, TV를 추가적으로 적합하는 것은 회귀식 설명력을 통계적으로 유의미하게 증가시킨다.’
* ‘다른 변수들(TV, Radio)가 고정된 상황에서, Newspaper를 추가적으로 적합하는 것은 회귀식 설명력을 통계적으로 유의미하게 증가시키지 않는다.’
* **F-test vs. t-test**
* F-test와 t-test 중에서는 회귀 모델 전체에 대한 F값을 먼저 봐야 한다. 전체 모델에 대한 F값을 확인함으로써 모델 전체가 통계적으로 유의한지 아닌지 확인할 수 있으며, 더욱 일반적인 검정이기 때문이다.
  1. 전체 회귀식에 대한 검정이 더 엄격하고, 2) F를 기각 못해도 t는 기각하는 경우가 있을 수 있기 때문!

### 4) 적합성 검정 (Goodness of fit)

단순선형회귀와 마찬가지로 선택한 변수들로 적합한 회귀모델이 데이터에 얼마나 잘 들어맞는지 확인하기 위해 적합성 검정이 필요하다. 똑같이 SST, SSR, SSE를 이용하여 적합성을 판단하지만, 조금 다르다.

* 의 문제점: 변수가 늘면 자연적으로 값이 증가

(총 변동은 고정되어 있는데, X변수가 추가되면 회귀식으로 설명되는 변동이 아주 조금이라도 증가할 수 밖에 없기 때문이다.)

단순선형회귀와 달리, 다중선형회귀의 경우 변수가 두 개 이상이 이에 따라 설명력이 높아지고 SSR은 자연히 증가한다. SSR이 증가하면 기존의 결정계수는 증가하게 된다. 설령 의미 없는 변수를 모델에 추가하더라도 말이다. 이러한 문제를 보완하기 위해 수정된 결정계수가 등장한다.

* **수정 결정계수** **(Adjusted R square)**

* + 변수가 추가됨에 따라 증가하는 결정계수에 변수 개수라는 패널티를 부과한 형태라고 생각하면 된다.
  + 변수의 개수가 다른 두 회귀식을 비교할 때 사용
  + 가 더 높은 회귀식이 더 좋은 회귀식
* **수정 결정계수 활용**
  + 수정 결정계수는 (3주차에 배울 예정인) AIC, BIC 처럼 변수의 개수가 다른 두 회귀식을 비교할 때 유용하다. 만약 변수의 개수가 다른 두 회귀식을 결정계수를 통해 적합성을 비교한다면 의미가 없더라도 변수가 더 많은 쪽의 결정계수가 높을 수 있다. 이를 고려하여 수정 결정계수를 사용한다.
  + 다만 결정계수처럼 "전체 변동 중에 회귀식이 설명하는 변동"으로 해석할 수 없다.

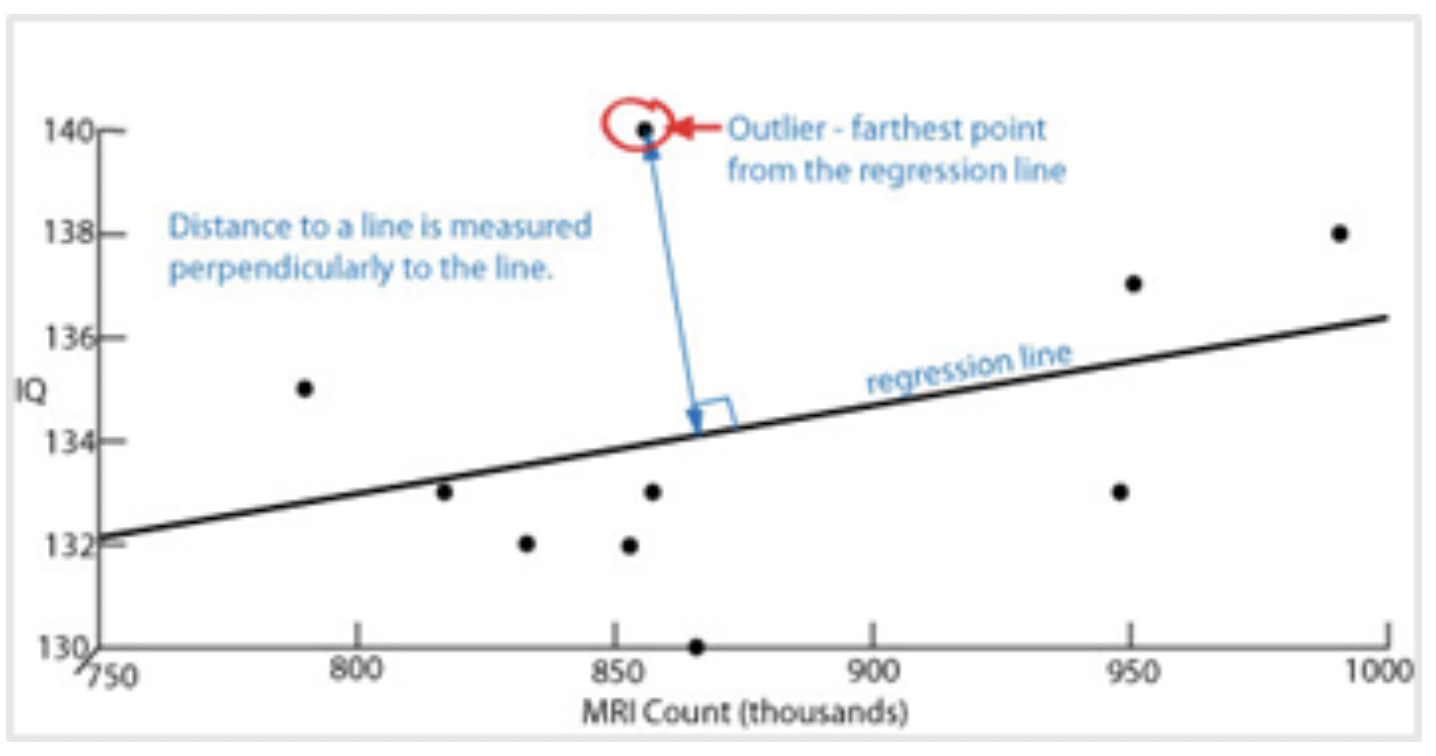
# 4. 데이터 진단

### 1) 표준화 잔차

* **잔차**
* **스튜던트화 잔차(studentized residual)**
* 잔차는 y 값의 단위에 영향을 많이 받기 때문에, 좀 더 일반화된 상황에서 적용할 수 있도록 표준화가 필요하다. 는 모수이므로 알 수 없기 때문에, 추정량 을 넣어주기 때문에 표준화 잔차가 아닌 스튜던트화 잔차라고 불린다.

### 2) 이상치 (Outlier)

표준화 잔차가 매우 큰 값. 즉, (표준화했을 때) $y$ 의 기준에서 절대값이 큰 값!

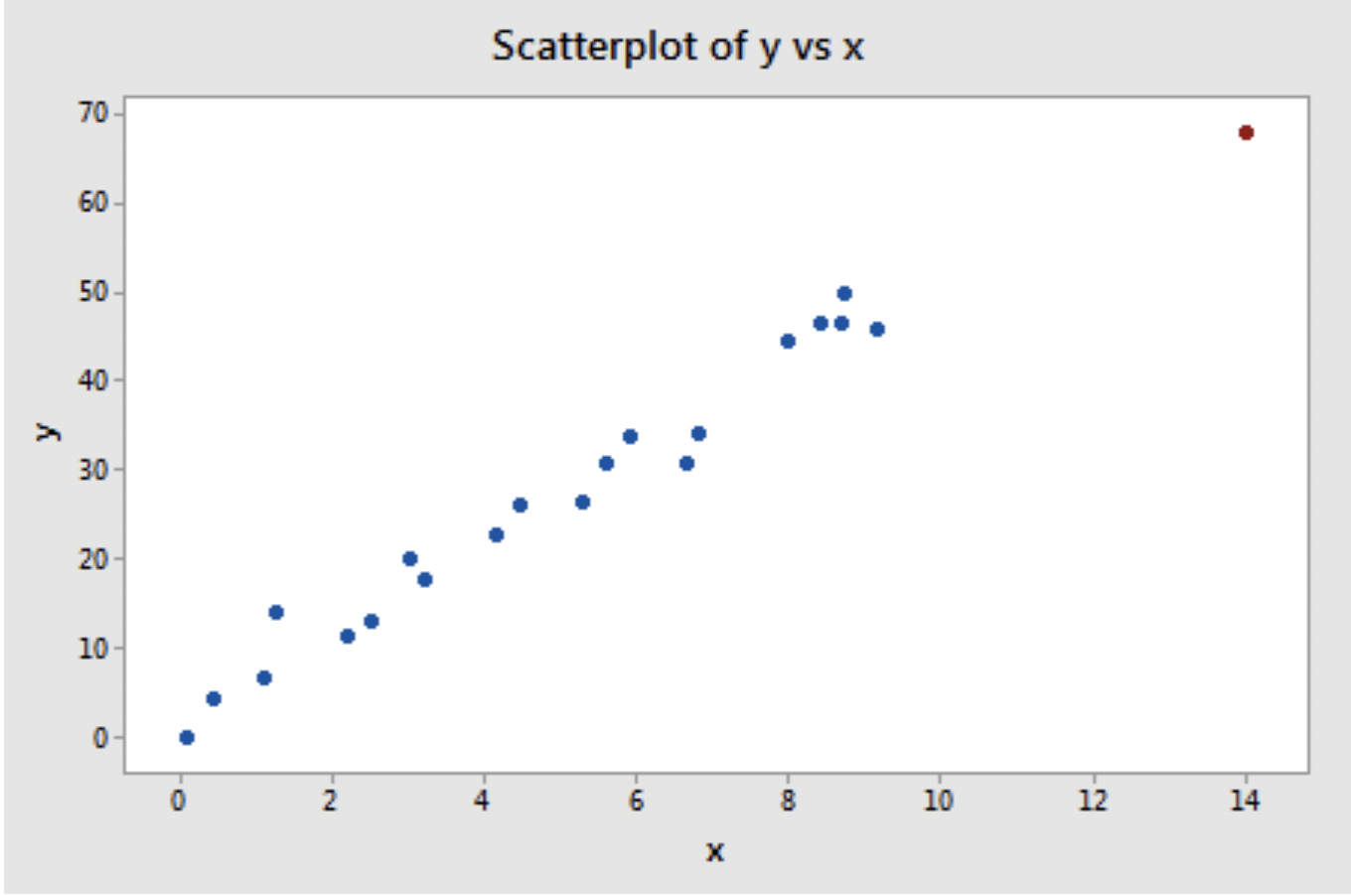


* 보통 이면 이상치라고 판단한다.

### 3) 지렛값 (Leverage point)

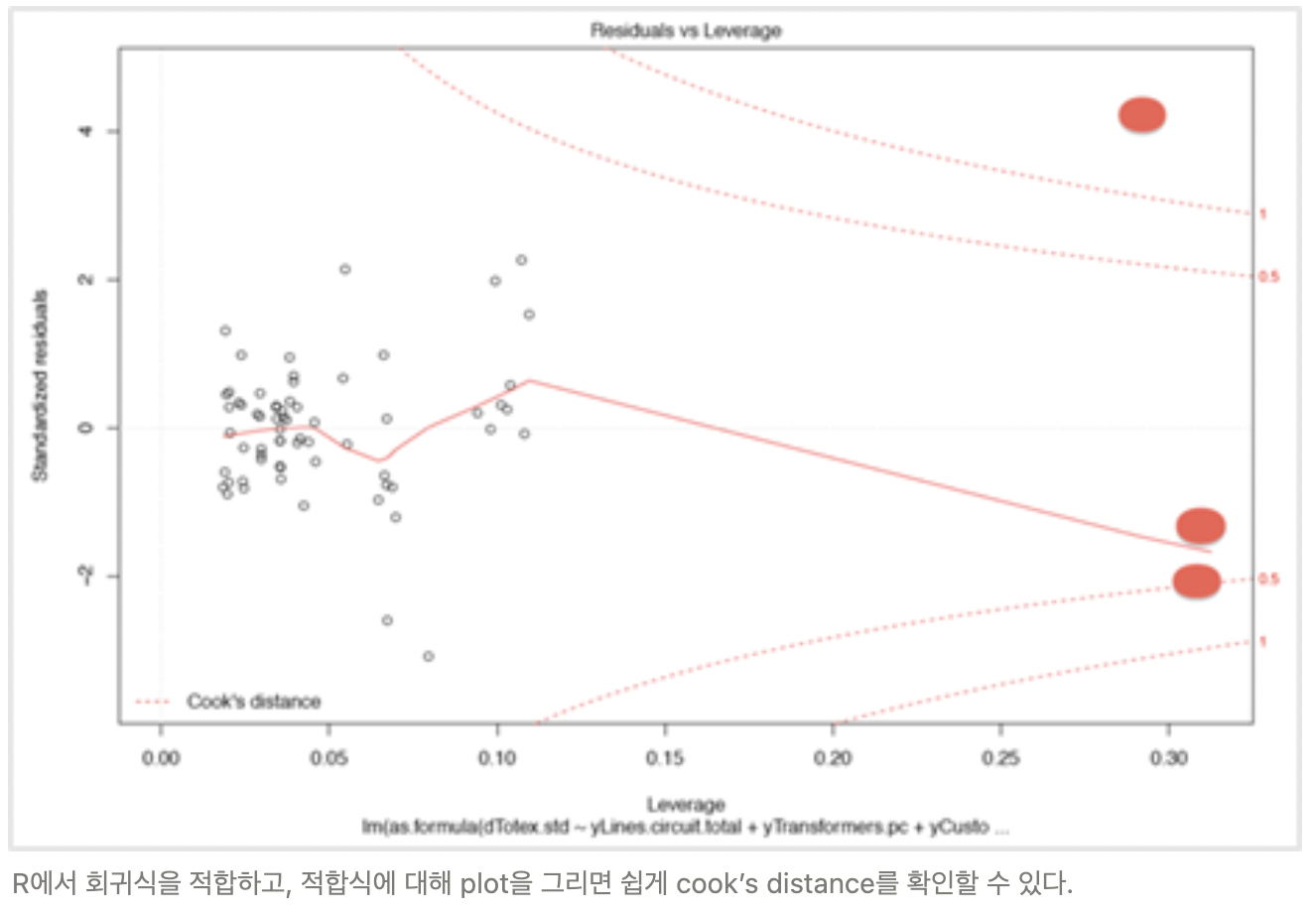
x의 평균 에서 멀리 떨어져 있어 기울기에 큰 영향을 주는 값

(표준화했을 때) x의 기준에서 절댓값이 큰 값!

* 라고 표현 가능하다.
* 식을 보면, 값과 의 차이가 클수록 가 커진다. → x 평균에서 멀수록 레버리지 값이 상승한다.
* 이면 지렛값으로 판단한다.
* 

### 4) 영향점 (Influential point)

한 관측치가 회귀직선의 기울기에 상당한 영향을 주는 점

* Outlier일지라도 X 평균 주위에 위치할 경우 기울기를 변화시키지 못하고, Leverage일지라도 회귀직선의 연장성에 있을 수 있다. 즉, Outlier나 Leverage point라고 해서 회귀직선을 변화시킨다고 말하기는 어렵다. 따라서 Outlier와 Leverage를 동시에 고려하는 지표가 필요하고, 이를 **Cook’s Distance**라고 한다.
* 
* **Cook's Distance**
* 영향점을 확인하는 표준적인 지표이다. 특정 데이터를 지웠을 때 회귀선이 변하는 정도를 나타낸다.
  + Outlier와 Leverage 각각이 커지면 커질수록 가 커진다.
  + 보통 이면 영향점으로 판단한다.
  + 
  + R에서 회귀식을 적합하고, 적합식에 대해 plot을 그리면 쉽게 cook’s distance를 확인할 수 있다.

### 5) 영향점의 처리

영향점의 존재는 추정량을 불안정하게 (분산을 크게) 만들고, 이는 잘못된 모델의 해석과 예측 성능 저하를 일으킬 수 있기 때문에 적절한 처리가 필요하다.

* 영향점 제거
* → 하지만 데이터를 삭제하는 것은 늘 신중하고 조심해야 한다! 의미를 가진 이상치라면?
* 이를 고려하는 이상치에 강건한 (robust) 모델링이 필요하다.

# 5. 로버스트 회귀

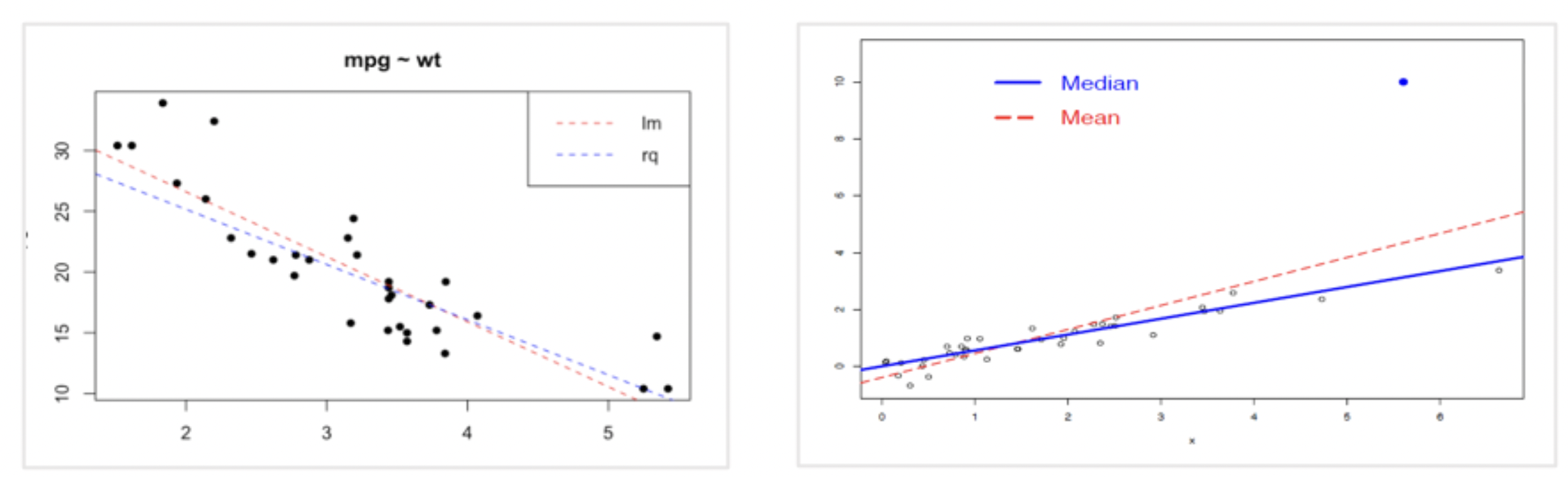
앞서 살펴봤던 영향점, 이상치의 존재를 보완하는 모델이 존재한다. 이러한 모델 강건한(robust) 모델이라고 한다. 이상치의 영향력의 덜 받는 robust한 회귀 모델을 간단하게 알아보자.

### 1) Median Regression

최소제곱회귀는 X에 따른 **평균**적인 Y를 반환한다면, median regression은 X에 따른 Y의 **중앙값**을 반환한다.

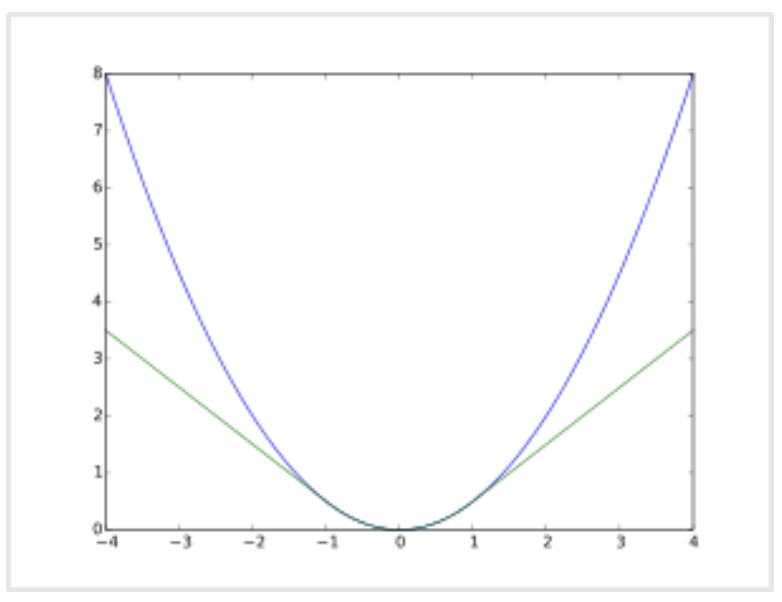
다중선형회귀는 독립변수X의 변화에 따른 종속변수 Y의 조건구 평균을 추정하지만, median regression은 조건부 중간값을 추정한다. 이를 통해서 중심에서 멀리 떨어진 이상치에 덜 민감한 추정량을 가질 수 있다.

(평균, 중앙값, 최빈값에 대해 배울 때, 중앙값은 이상치의 영향을 덜 받는다고 배웠죠?)



* 최소제곱법은 오차의 제곱합을 최소로 하는 추정량을 찾는다. 반면 median regression은 오차의 절대값의 합을 최소화한다. 을 최소화
* 최소제곱법은 이상치에 너무 큰 가중치를 주는 경향이 있는 반면, median regression은 어떠한 경우에도 동일한 가중치를 준다.
* 평균이나 최빈값에 비해 이상치의 영향을 덜 받는 중앙값을 사용한다.
* 분포 가정과 등분산 가정이 없는 모델이다
* R에서는 'quantreg' 패키지의 rq() 함수 사용

### 2) Huber's M-estimation

* 최소제곱회귀는 이상치에 지나치게 큰 페널티를 부여하지만, 동시에 적정수준 안에서는 페널티를 완화시켜준다. Huber’s M-estimation 역시 적정수준의 페널티는 완화시켜주는 형태는 유지하되, 이상치에 대한 지나친 페널티 부여를 없애는 방식이다.
* Huber’s M-estimation의 아이디어는 잔차가 특정 상수값보다 크면, 잔차의 '제곱'이 아닌 1차식으로 바꾸어 이상치에 강건한 회귀계수를 추정하는 것이다.
* **함수**
* 잔차의 절대값이 c이하면 기존 최소제곱추정법의 목적함수와 동일하지만, 상이면 이상치에 대한 패널티를 없앤 일차식의 형태이다.
* 
* 파란색 : 최소제곱회귀 / 초록색 : Huber's M-estimation
* R에서는 'MASS' 패키지의 rlm() 함수 사용

### 3) Least Trimmed Square

통계적 기준에 따라 잔차가 너무 큰 관측치를 제거하고 회귀계수를 추정하는 방식

* 공식

텍스트, 시계이(가) 표시된 사진

자동 생성된 설명

* + 이 때, 는 작은 순서부터 오름차순으로 나열한 잔차이다.
* n개의 obs 중 h개만 사용하여 회귀식을 만드는데, 개의 회귀식 중 가장 잔차제곱합이 작은 회귀식을 사용한다.
* obs가 별로 없는 경우나 영향점이 존재하지 않는 경우 주의해서 사용해야 한다.
* R에서는 'robustbase' 패키지의 ltsReg() 함수 사용