Week 1 : 회귀분석의 기초

2023-2학기 회귀분석팀 클린업에 참가해주신 여러분 안녕하세요.

회귀분석이 통계학에서 참 기초적인 과목이면서도 선뜻 이해하기가 어려운, 통계학의 첫 고비라고도 생각합니다(일단 저는 그랬어요). 3주라는 짧은 시간 동안 중요한 내용들을 컴팩트하면서도 쉽게 이해하실 수 있도록 열심히 해보겠습니다 ㅎㅎ 잘 부탁드려요~!~!

< 목차 >

- 로버스트 회귀의 비용

---

0. 기본 수식

1) 기초수식

- 표본 평균 (Sample Mean)

$$\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i \qquad \bar{y} = \frac{1}{n}\sum_{i=1}^{n} y_i$$

- 표본 분산 (Sample Variance)

$$S_x^2 = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2 \qquad S_y^2 = \frac{1}{n-1}\sum_{i=1}^{n}(y_i - \bar{y})^2$$

- 표본 표준편차 (Sample Standard Deviation)

$$S_x = \sqrt{S_x^2} = \sqrt{\frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2}$$

- 편차제곱합(변동) : S_{xx}, S_{xy}

$$S_{xx} = \sum_{i=1}^{n}(x_i - \bar{x})^2 \qquad S_{xy} = \sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})$$

2) 공분산 (Covariance)

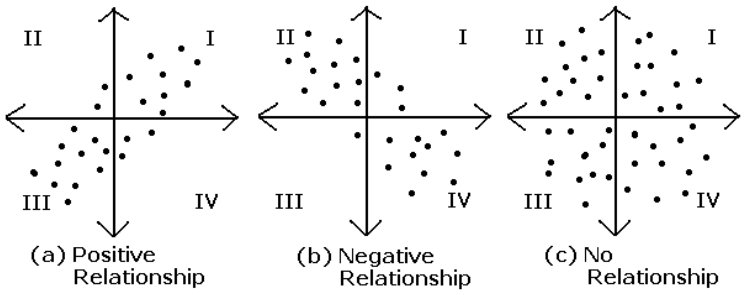
두 확률변수의 선형 관계를 나타내는 값

두 변수가 가지는 **선형관계의 방향성(양, 음)**만 나타낼 뿐, 얼마나 선형성을 갖는지 즉 '강도'는 표현하지 못한다. $(-\infty, \infty)$, 음의 무한대와 양의 무한대 사이의 값을 가진다.

- Sample Covariance 공식

$$Cov(X, Y) = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})$$

- 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 선형 상관관계가 존재하지 않는다.

- 단점

  공분산은 확률변수의 측정단위에 영향을 많이 받기 때문에(not scale free), 상관성의 형태(양, 음, 상관관계 없음)에 대해서는 나타낼 수 있지만, 그 크기가 상관성의 정도를 직접 나타낼 수는 없다.

  ex) A와 B의 공분산과 C와 D의 공분산의 크기가 다르더라도, 각각이 선형관계를 나타내는 정도는 같을 수 있다.

  측정단위가 달라지면 공분산의 값도 달라지기 때문이다. 따라서 단순히 공분산이 더 크다고 해서 선형관계가 강하다고 말할 수 없다.

  → 이를 보완하기 위해 **상관계수(Correlation)**를 사용할 수 있다.

3) 상관계수 (Correlation Coefficient)

확률변수의 절대적 크기에 영향을 받지 않도록 단위화를 진행해 준 **'표준화된 공분산'**이다.

- Sample correlation 공식

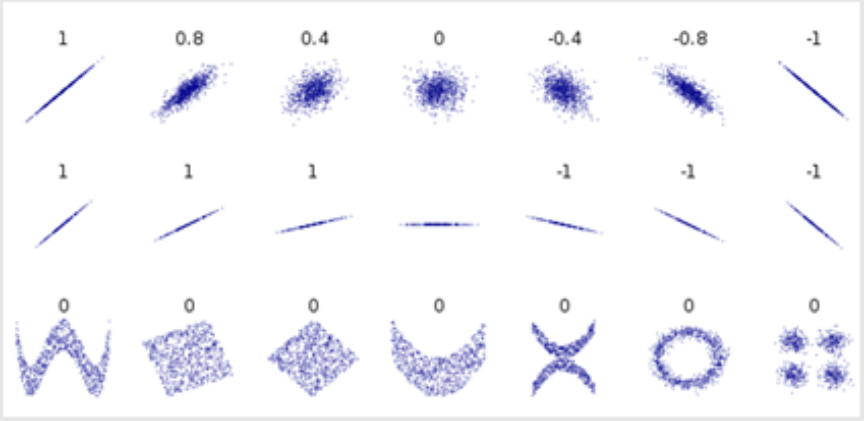
$$r_{xy} = \frac{Cov(X,Y)}{s_x \cdot s_y} = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2}\sqrt{\sum_{i=1}^{n}(y_i - \bar{y})^2}}$$

- 상관계수의 해석

  상관계수는 두 확률변수의 선형 상관관계의 여부와 선형적인 상관성의 크기까지 파악할 수 있는 지표이다. -1부터 1까지의 값을 가지며, 확률변수 X, Y가 독립일때 상관계수는

0이 된다. 일반적으로 0.7이상이면 강한 상관관계를 지닌다고 판단한다.

상관계수가 0(zero-correlation)이면 두 변수는 아무런 선형관계를 가지지 않는다는 의미이지만, 선형관계가 없을 뿐, 비선형관계는 존재할 수 있다.



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

---

1. 회귀분석이란?

1) 회귀분석의 정의

- 독립변수와 종속변수 간의 관계를 설명하고 모델링하는 통계적 기법
- 변수들 간의 상관관계를 파악하고, 이를 통해 특정 변수의 값을 다른 변수들을 이용하여 설명하고 예측하는 방법

  ex) 암 발병률과 사망률과의 관계, 범죄율과 주택 가격과의 관계
- 회귀분석의 목적
  - 변수들 간의 관계에 대한 표현
  - 독립변수에 따른 종속변수의 변화 파악
  - 미래 관측값에 대한 예측

2) 회귀식

독립변수(Predictor, Feature) X와 종속변수(Response) Y의 관계를 함수식으로 표현한 것

$$Y = f(X_1, X_2, \cdots, X_p) + \epsilon$$

- X_k 독립변수(independent variable) : 종속변수를 설명하기 위한 변수로 설명변수(explanatory variable) 혹은 예측변수(predictor variable)라고 함
- Y 종속변수(dependent variable) : 독립변수에 의해서 설명되는 변수로 반응변수(response variable)라고도 불림
- ϵ 오차항(error term) : 변수를 측정할 때 발생할 수 있는 오차로 설명할 수 없는 무작위성을 지님

3) 상관분석과의 차이

상관분석은 두 변수의 역할이 서로 대등할 때 사용된다. 예를 들어, 키와 몸무게라는 변수에 관심이 있다고 할 때 키로 몸무게를 설명할 수도 있고, 몸무게로 키를 설명할 수도 있을 것이다. 이처럼 연구자의 의도에 따라 target 변수가 바뀔 수 있을 때 두 변수는 대등하다고 한다.

반면, 보석의 가격과 크기라는 두 변수에 관심이 있다고 해보자. 이 때 가격을 target으로 삼고 보석의 크기를 가격을 설명하기 위한 변수로 두는 것이 일반적이고, 크기를 target으로 삼는 경우는 드물 것이다. 이처럼 두 변수가 대등하지 않고 관계에 분명한 방향이 있을 때 회귀분석을 사용하게 된다.

- 상관분석의 한계
    - 상관관계는 두 변수의 관계만 표현할 수 있다. (X의 값이 크기 **'때문에'** Y의 값이 크다고 할 수는 없는 것)
    - 두 변수의 선형적 상관성 정도만 표현할 수 있고, 구체적인 예측과 설명은 불가능

  ⇒ 독립변수가 한 단계 변할 때마다 종속변수가 어떻게 변화할지를 안다면 더 유의미하게 관계를 파악할 수 있다… 회귀분석을 사용하는 이유

📌 **회귀분석과 인과관계**

아까 설명한 이유로, 많은 책에서는 회귀분석을 인과관계를 설명하기 위한 분석 기법이라고 표현한다. 그러나 그것이 회귀분석 자체가 인과관계를 설명한다는 것은 아니다.

회귀분석은 변수간의 상관관계를 기반으로 한 분석으로
독립변수를 통해 종속변수를 예측하는 것을 목표로 한다. 독립변수와 종속변수를 가정해서 분석하긴 하지만, 그 결과 자체가 인과관계를 의미하지는 않는다. 다른 말로 하면 인과관계를 파악하기 위해 회귀분석이 사용되는 것이지, 단순히 모델의 결과가 유의미하다는 사실만으로는 인과관계가 있다고 할 수 없는 것.

※ **참고** (인과관계가 성립하기 위한 요건)

① x가 y보다 시간적으로 먼저이다. (또는 논리적으로)

② x가 있으면 y가 있고, x가 없으면, y도 없다.

③ x와 y사이에 지금보다 더 정확한 영향을 끼치는 원인이 없다.

미묘한 차이를 아시겠나요?.. ㅎㅎ

4) 회귀 모델링 과정

1. **문제 정의**

- 희나의 학점을 가장 잘 표현할 수 있는 변수들은 무엇이 있을까?

2. **적절한 변수 선택**

- $X_1, \dots, X_p$: 공부 시간, 통학 거리, 아침밥 식사 여부

3. **데이터 수집 및 전처리**

- 희나의 학점, 공부 시간, 집에서 학교까지의 거리, 아침밥 식사 여부

4. **모델 설정과 적합**

- 적절한 회귀분석 모델 선택 및 적합
  선형 vs 비선형, 단순회귀 vs 다중회귀, 모수 vs 비모수, 일변량 vs 다변량 등 고려가능

5. **적합성 및 유의성 검정**

- 회귀모델이 얼마나 데이터를 잘 설명하는지, 회귀계수는 통계적으로 유의한지 검정

6. **모형 평가**

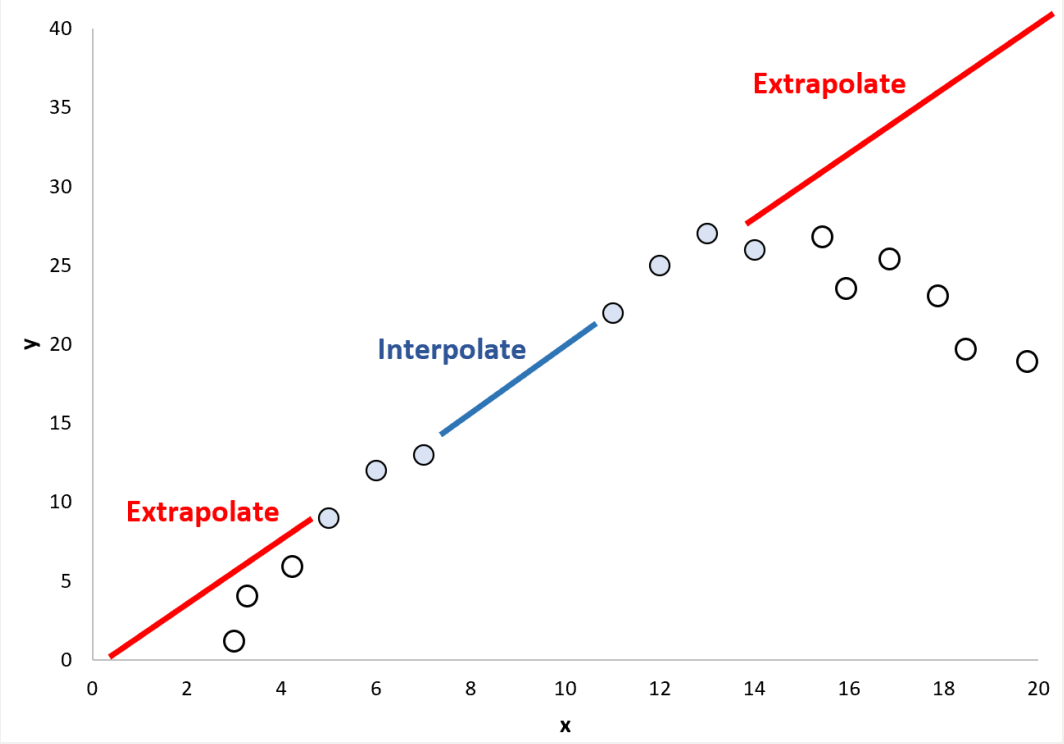
- 설정한 모델이 회귀 가정을 만족하는가? -> 만족하지 않으면 처방 (회귀팀 2주차 예정)

7. **모형 해석**

- 희나가 현재보다 주 당 2시간 더 공부하고, 자취방에서 통학을 하고, 아침밥을 꼬박꼬박 챙겨 먹는다면 → 학점이 0.3만큼 오를 것이다!

⚠️ **외삽(extrapolation)**

모형의 적합 및 해석에서, X의 관측값의 범위를 벗어나는 영역에도 적합된 회귀직선을 적용하는 것(외삽)에는 주의해야한다. 관측하지 않은 영역에서는 두 변수의 관계가 어떤지 알 수 없기 때문이다.



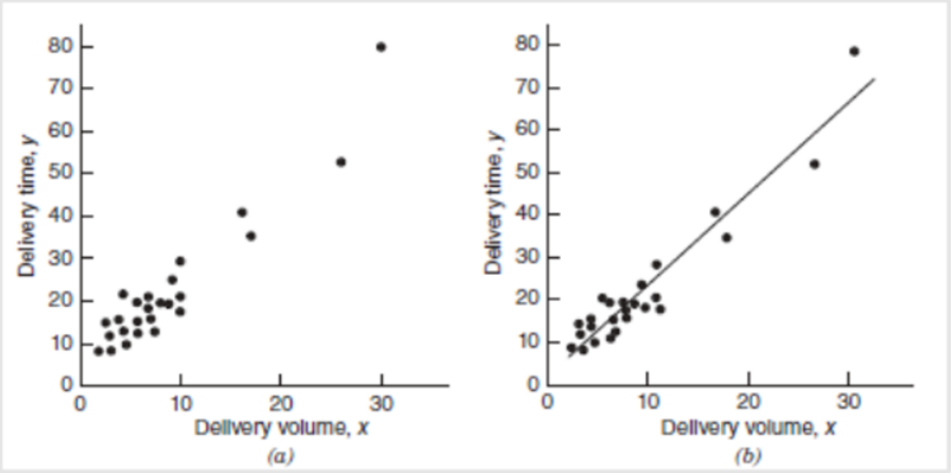
관측된 범위 밖에서 두 변수가 비선형 관계에 있는 경우

2. 단순선형회귀분석

1) 단순선형회귀(Simple Linear Regression)

하나의 독립변수 X와 하나의 종속변수 Y의 관계를 가장 잘 표현할 수 있는 **직선**을 찾는다.

주어진 데이터를 가장 잘 설명할 수 있는 직선을 찾아 수식화하는 것이 단순선형회귀이다.



→ '변수의 관계는 **선형적**이다'는 가정 하에 직선 함수식을 가정함 (회귀팀 2주차 예정)

- 단순선형회귀식

모집단의 관점(Population Regression Model) : $y = \beta_0 + \beta_1 x + \varepsilon$

관측치의 관점(Sample Regression Model) : $y_i = \beta_0 + \beta_1 x_i + \epsilon_i \ \ (i = 1, \ldots, n)$

- 회귀모델 설명

$y_i = \beta_0 + \beta_1 x_i + \epsilon_i \qquad i = 1, ..., n \qquad \epsilon_i \overset{iid}{\sim} N(0, \sigma^2)$

y_i : 종속변수 y의 i번째 관측값

x_i : 독립변수 x의 i번째 관측값

ϵ_i : i번째 관측값에 의한 랜덤오차 (이 때, 정규분포 가정 $\epsilon_i \overset{iid}{\sim} N(0, \sigma^2)$ 만족)

β_0, β_1 : 회귀계수 또는 우리가 추정해야 할 모수 (parametric 방법)

→ 더 좋은 모델을 만들기 위해서는 회귀계수를 잘 추정해야 한다.

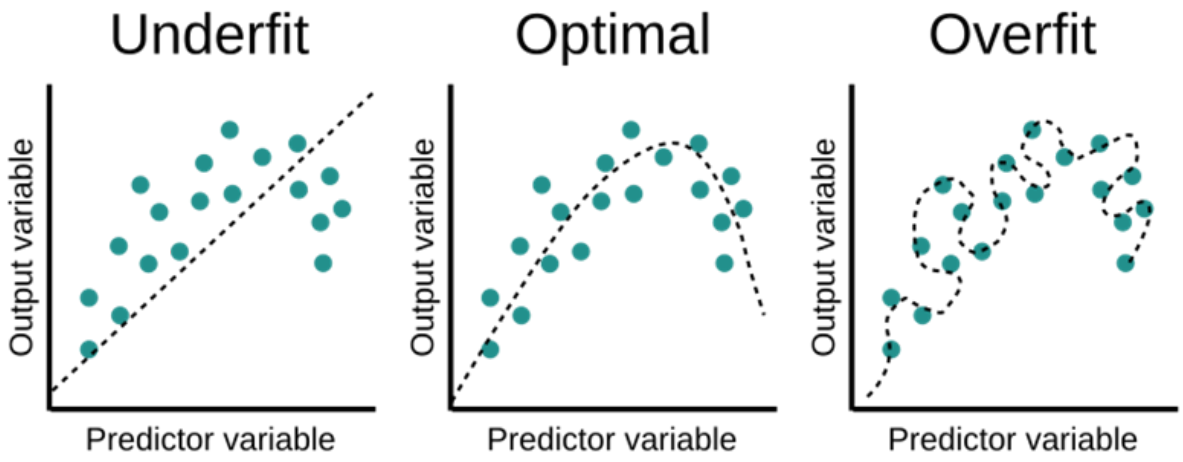
- 단순선형회귀 모델의 해석

x가 한 단위 증가할 때, y**는** β_1**만큼 증가**한다. (Y의 평균값의 변화량이 β_1!)

- 왜 직선인가?

→ 변수의 영향력을 간단하게 모형화 할 수 있다.

→ 2차원 평면 위의 X와 Y의 일대일대응 관계를 통해 **독립변수의 변화에 따른 종속변수의 변화를 직관적으로 확인**할 수 있다!



- 선형근사를 넘어서 고차근사를 할 경우 모델의 복잡도가 높아지게 되는데, 이 경우 **과적합(overfitting)**의 원인이 된다. 즉, 당장 주어진 training data에 대한 설명성은 높을 수 있지만, 예측의 대상이 되는 test data에 대한 설명성은 떨어진다는 의미 (모델의 분산을 높이고, 검증 데이터의 예측 성능을 저하시킴) (데이터마이닝팀 1주차 클린업 참고)

- 현실의 데이터는 선형적으로 생성되지 않는 경우가 많아, 예측 성능이 떨어지는 경우도 있다. 하지만 선형회귀식의 원리와 가정, 변형을 통해 예측 모델링에 대한 전체 흐름을 이해해야 추후 머신러닝 방법들에 대해 총체적으로 이해할 수 있다.

- 위와 같이 단순 선형 회귀식의 예측 성능 문제 때문에, local regression, smoothing spline, GAM(Generalized Additive Model)과 같은 비선형 모델을 사용하기도 한다.

2) 추정: 최소제곱법(LSE: Least Square Estimation Method)

앞서 모형의 실제 형태를 알 수 없기 때문에 $y_i = \beta_0 + \beta_1 x_i + \epsilon_i$ 형태의 선형모델을 가정했다. 분석하고자 하는 회귀모형을 결정한 후에는 모형에 포함된 모수 $\beta_0,\ \beta_1,\ \sigma^2$를 추정(estimation)해야 한다. 회귀직선만 잘 만들어낸다면 σ^2의 추정은 어렵지 않기 때문에 β 추정에만 집중해도 된다.

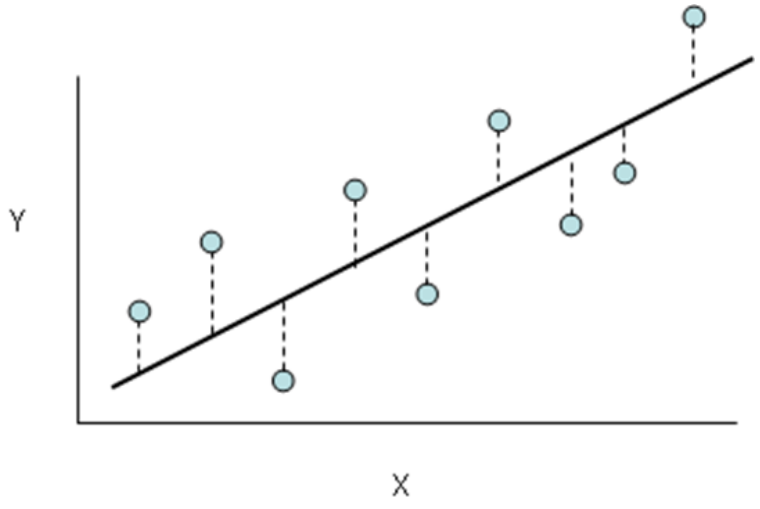
→ **데이터를 가장 잘 표현하는 직선을 찾는 것**이 목표이기 때문!

- 좋은 추정이란?

직관적으로 우리가 만들어낼 회귀직선과 관측치 사이의 오차가 작으면 작을수록 좋은 추정이다.

절대적인 떨어짐(deviation)을 최소화하는 방법을 찾기 위해서 우리는 오차제곱합을 최소화하는 방법을 찾게 되는데, 이를 **최소제곱법(LSE, Least Squared Error)**이라고 한다.

→ 즉, 위 직선에서 각 점으로부터 구하고자 하는 최적 직선까지의 수직거리가 오차이며, 이러한 오차제곱합이 최소가 되는 직선을 찾는 것이다!



점선의 제곱합을 최소화하는 방법!

- 최소제곱법을 이용한 모수 추정

$$\underset{\beta_0,\beta_1}{\operatorname{argmin}}\ J(\beta_0,\beta_1) = \sum_{i=1}^{n}(y_i - \beta_0 - \beta_1 x_i)^2 = \sum_{i=1}^{n}\epsilon_i^2$$

위 식에서 $J(\beta_0, \beta_1)$가 오차제곱합이고, 이 제곱합을 최소화시키는 β_0, β_1를 찾는 것이 목적이다. 이런 제곱꼴의 함수는 아래로 볼록한 Convex 함수이기 때문에, 각각의 모수를 편미분하여 '**미분값 = 0**'을 만족시키는 β_0, β_1값이 우리가 구하고자 하는 추정량(estimator)가 된다.

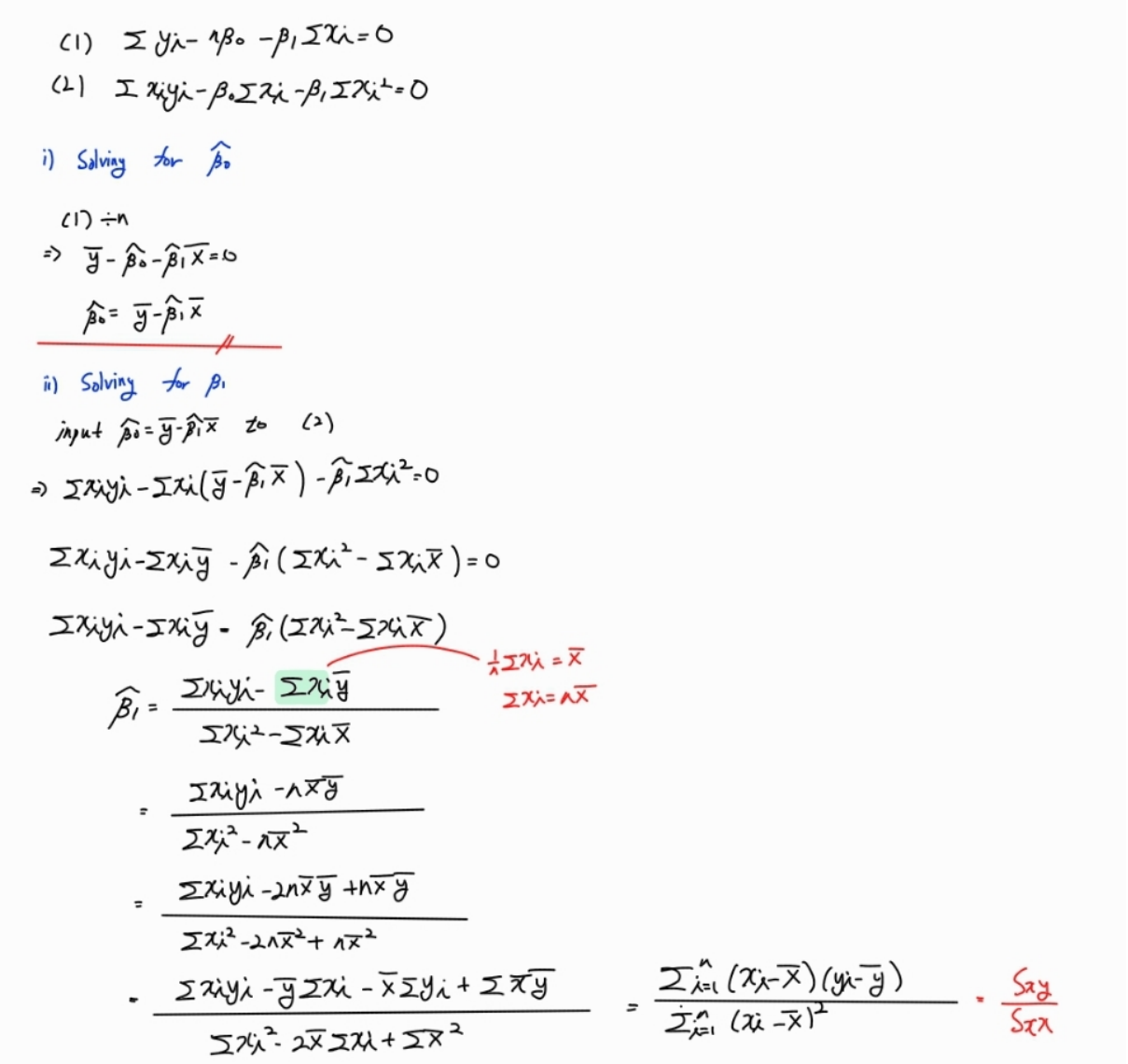
$\rightarrow$ LS estimators, $\hat{\beta}_0$ and $\hat{\beta}_1$ satisfy

$$\frac{\partial J}{\partial \beta_0}|_{\hat{\beta}_0, \hat{\beta}_1} = -2\sum_{i=1}^{n}(y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i) = 0 \ \cdots (1)$$

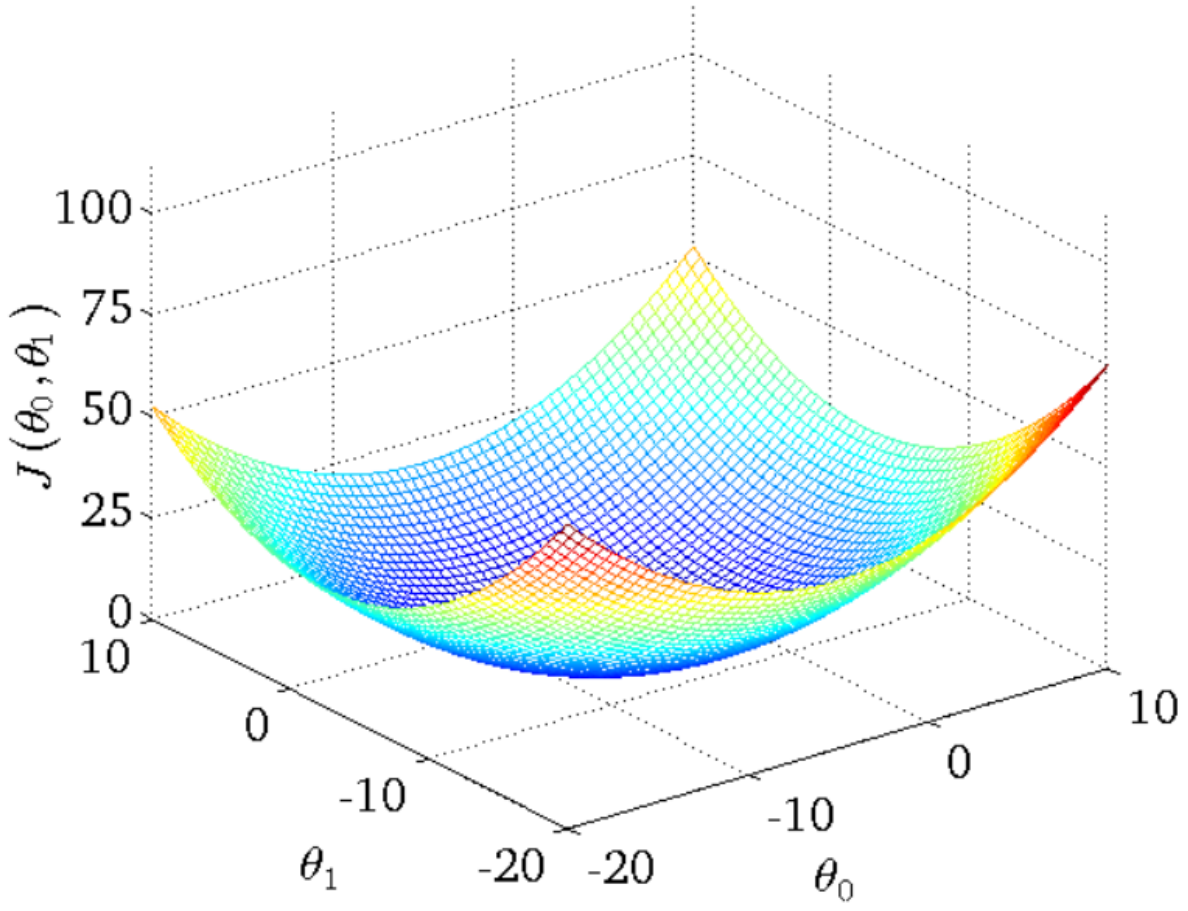
$$\frac{\partial J}{\partial \beta_1}|_{\hat{\beta}_0, \hat{\beta}_1} = -2\sum i = 1^{n}(y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i)x_i = 0 \ \cdots (2)$$

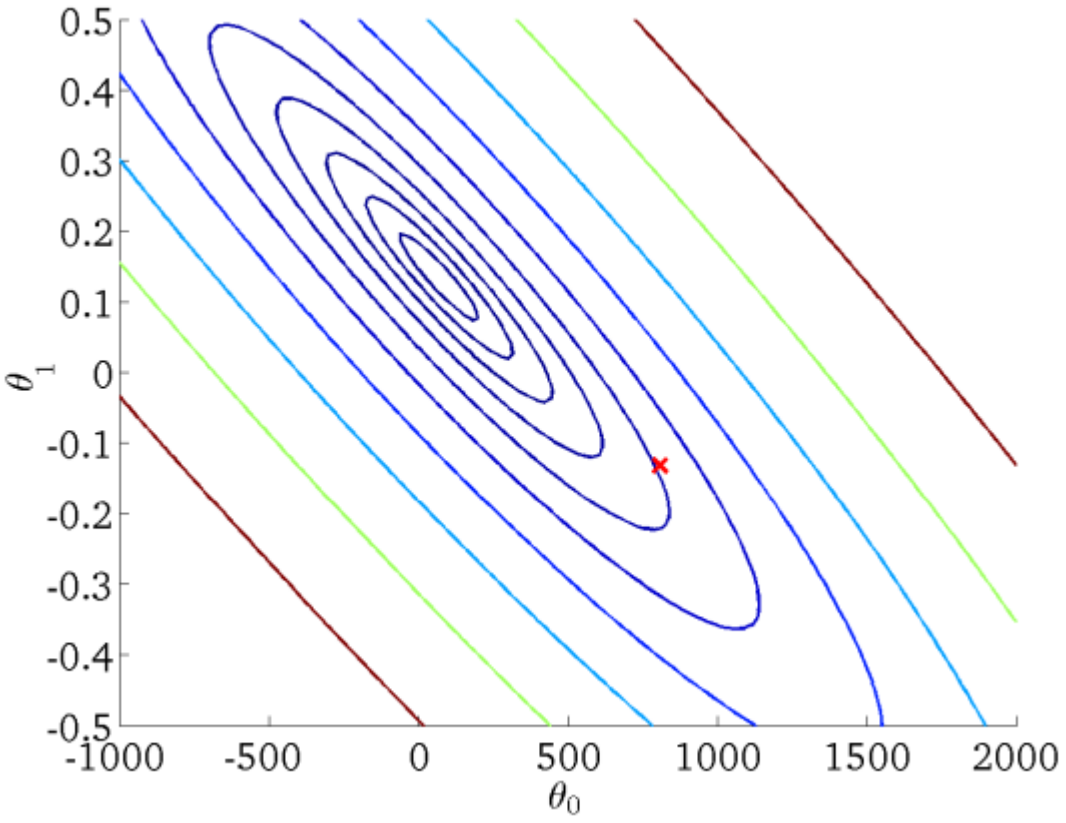
위 식들을 정규방정식(normal equation)이라 하고 이 normal equation을 β_0, β_1에 대해 연립하면 회귀계수의 추정치를 얻을 수 있다.

과정은 다음과 같고,



오차제곱합 $J(\beta_0, \beta_1)$을 기하학적으로 나타내면 다음과 같다.





특히 위의 등고선 플랏(contour plot)은 3주차에 배울 Ridge, LASSO 에서 또 등장하니 익혀두도록 하자.

The solution to the normal equation

$\hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x} \quad , \quad \hat{\beta}_1 = \frac{S_{xy}}{S_{xx}}$

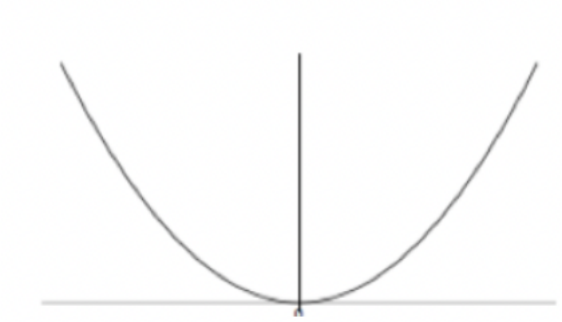
where $\bar{y} = \frac{1}{n}\sum_{i=1}^{n} y_i,\ \bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i,$
$S_{xy} = \sum_{i=1}^{n} y_i(x_i - \bar{x}) \quad \text{and} \quad S_{xx} = \sum_{i=1}^{n}(x_i - \bar{x})^2$

이렇게 최소제곱법을 통해 얻은 **추정치 β_0과 β_1을 최소제곱추정치(LSE : Least Square Estimator)**라고 한다.

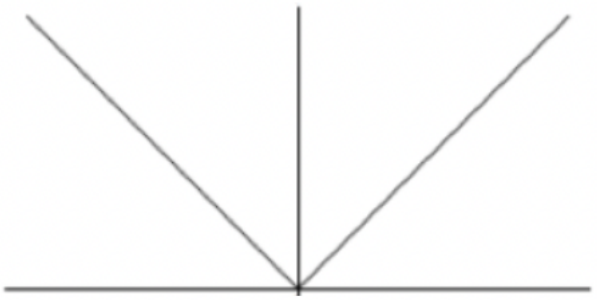
- 왜 오차의 '제곱합'을 최소화할까?

→ 미분이 편리하기 때문

'미분값=0'을 통해 바로 추정할 수 있기 때문에 매우 간단하다.



오차제곱합



오차의 절대값

→ 오차의 절댓값을 사용하는 방법도 있지만, 오차의 절댓값을 목적함수로 사용하게 된다면 미분이 불가능한 점이 존재하여 수치적 방식을 사용하게 되고 **계산이 오래 걸리는 문제**가 발생한다.

이를 수학적으로는 closed-form solution이 존재하지 않는다고 한다. 문제에 대한 해답을 식으로 명확하게 제시할 수 없다는 뜻

→ **오차가 클수록 더 큰 패널티**를 부여할 수 있다. (중심에서 멀어질수록 기울기가 상승하고 있음)

→ 작은 오차의 페널티를 줄이고, 큰 오차의 페널티를 높여 오차가 최소화되는 지점을 찾음

- **BLUE(Best Linear Unbiased Estimator)**

LSE를 이용한 추정은 분포에 대한 가정 없이도 사용 가능하다는 장점이 있다. 하지만, 다음의 특정 조건이 갖추어진다면 더 유용한 성질을 지니게 된다.

1. 오차들의 평균은 0
2. 오차들의 분산은 σ^2으로 동일 (등분산)
3. 오차간에는 자기상관이 없다. (uncorrelated)

이러한 조건들을 모두 만족하면
가우스-마코프 정리에 의해 LSE(최소제곱 추정량)은 BLUE가 된다!

⇒ **Best(분산이 제일 작은) Linear(선형) Unbiased Estimator(불편추정량)**으로 다른 선형불편추정량보다 분산이 늘 작다.

(분산이 작다는 것은 추정량이 안정적이라는 의미이므로 신뢰할 만한 추정량이라고 할 수 있음)

- **최대가능도추정량(MLE)**

회귀모형이 선형인 경우 회귀계수는 LSE를 이용하여 쉽게 구할 수 있으나 비선형인 경우는 LSE를 이용하여 구하기 쉽지 않다. (로지스틱, 시계열 회귀 모형) 이 때 MLE를 사용하여 회귀계수를 추정하게 된다. 최대가능도추정량은 오차항이 갖는 분포에 대한 가정이 필요하다.

> ✅ **최대가능도추정(MLE : Maximum Likelihood Estimator)**
> 확률적인 방법에 근거해서, 어떤 모수가 주어졌을 때 원하는 데이터가 나올 '가능도'를 최대로 하는 모수를 선택하는 방법
>
> 통계적인 직관! (측정값 혹은 관찰값에 대한 최대 가능도(확률밀도함수 PDF의 Y값)를 추정)

※ 오차의 정규분포 가정이 있다면 **LSE와 MLE는 완전히 동일한 추정량**을 산출한다.

3) 적합성(Goodness of fit) 검정

지금까지 단순회귀모형의 정의와 최소제곱법을 이용하여 단순회귀모형의 모수를 추정하는 방법을 배웠다. 이렇게 추정한 모수를 **회귀계수(regression coefficient)**라고 부르며, 추정된 회귀계수를 바탕으로 **회귀직선**을 만들 수 있다.

→ 회귀선을 만들었다면 이 회귀식이 얼마나 우리 데이터를 잘 설명하는지 알아보아야 한다. 모형의 적합성에 대한 평가 과정을 '적합성 검정'이라고 한다.

- 잔차(Residual)

  추정한 회귀계수를 이용해 회귀직선을 만들었을 때, 오차의 추정량

  실제 오차는 회귀계수처럼 **실제값(모수)을 알 수 없기 때문에** 추정된 직선을 통해 추정해야 적합

  $$e_i = y_i - \hat{y}_i = y_i - (\hat{\beta}_0 + \hat{\beta}_1 x_i) \ , \ \sum e_i = 0$$

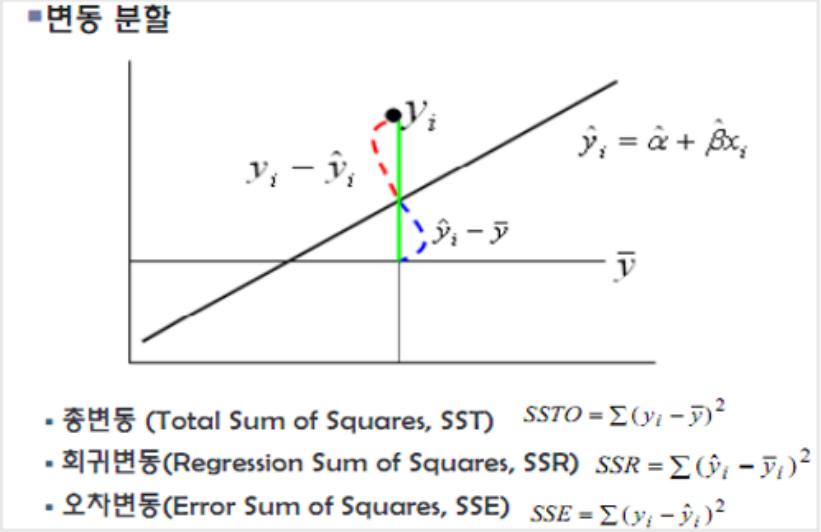
- 적합성 검정

  - SST(Total Sum of Squares, 총 변동) : $\sum (y_i - \bar{y})^2$

- SSR(Regression Sum of Square, 회귀선이 설명하는 변동) : $\sum(\hat{y_i} - \bar{y})^2$

- SSE(Residual Sum of Square, 잔차제곱합, 회귀선이 설명하지 못하는 변동) : $\sum(y_i - \hat{y_i})^2$

- SST = SSR + SSE



- **결정계수 R^2**

  총 변동(SST)에서 회귀식이 설명할 수 있는 비율(SSR)로 (0,1)의 값을 가지며 1에 가까울수록 회귀모형이 데이터를 잘 설명한다고 볼 수 있다.

  $$R^2 = \frac{SSR}{SST} = 1 - \frac{SSE}{SST}$$

  이를 잔차와 연관지어 생각해보면, 잔차제곱합(SSE)는 회귀식이 설명할 수 없는 실제값과 추정값 사이의 오차로, 총 변동 대비 잔차제곱합이 차지하는 비율이 작을수록 좋다.

  결정계수가 높고 낮음에 절대적인 기준은 없다. 변수의 관계를 고도의 정밀도로 밝힐 수 있는 자연과학이나 공학에서는 0.95 이상의 높은 값을 요구하기도 하지만, 변수 간 관계가 복잡하게 얽혀있고 인위적 조정이 불가능한 사회과학 분야에서는 0.4~0.6의 값도 의미 있는 적합도로 해석하기도 한다.

4) 유의성 검정

다음으로 전체 회귀식의 설명량이 아닌 개별 모수의 추정량이 통계적으로 유의한지를 알아보는 검정이 필요하다. $\epsilon_i \sim N(0, \sigma^2)$ 라는 오차의 정규분포 가정 하에 개별 회귀계수에 대해 다음과 같은통계적 검정이 가능하다.

1. 가설 설정 : $H_0 : \beta_1 = 0 \quad vs \quad H_1 : \beta_1 \neq 0$
2. 추정량의 분포 : $\hat{\beta}_1 \sim N(\beta_1,\ \frac{\sigma^2}{S_{xx}})$
3. 검정 통계량 : $t_0 = \frac{\hat{\beta}_1}{se(\hat{\beta}_1)} \sim t_{(n-2)}$
4. 임계값 : $t_{(1-\alpha/2,n-2)}$
5. 검정(양측) : $\text{If} \quad |t_0| > t_{(1-\alpha/2,n-2)}, \quad \text{reject } H_0 \ \text{ at } \ \alpha \ \text{ level}$

β_0 에 대해서도 동일한 방법으로 검정을 진행할 수 있다.

※ 귀무가설(H_0)을 기각하지 못하면 개별 회귀계수는 0이 아니라고 통계적으로 단언할 수는 없다.

이는 'X, Y 에는 **선형적 관계가 없다.**' 라고 할 수 있다. (선형 관계가 없다는 것이지 X, Y 사이에 아무 의미가 없다는 것은 아니다. 비선형 관계일 수도 있으므로..)

---

3. 다중선형회귀

1) 다중선형회귀

독립변수 X가 2개 이상인 경우의 회귀분석

- 독립변수가 한 개였던 단순선형회귀에서 독립변수 개수만 여러 개로 늘어난 경우이다. 단순선형회귀에 비해 더욱 복잡한 관계를 설명할 수 있어 자연현상, 사회현상을 파악하기에 유리하다.
- 다중회귀모형은 다차원에 대한 표현을 해야 하므로 **벡터와 행렬**로 표현하게 된다. 또한, 독립변수가 여러 개이기 때문에 단순회귀모형과는 달리 **독립변수들 간의 관계**도 고려되어야 한다.

- 공식

  $y = \beta_0 + \beta_1 x_1 + \cdots + \beta_p x_p + \epsilon \quad , \quad \epsilon \overset{iid}{\sim} N(0, \sigma^2)$을 가정

  전체 독립변수의 개수는 p개, 회귀계수는 $p+1$개 (β_0 포함)

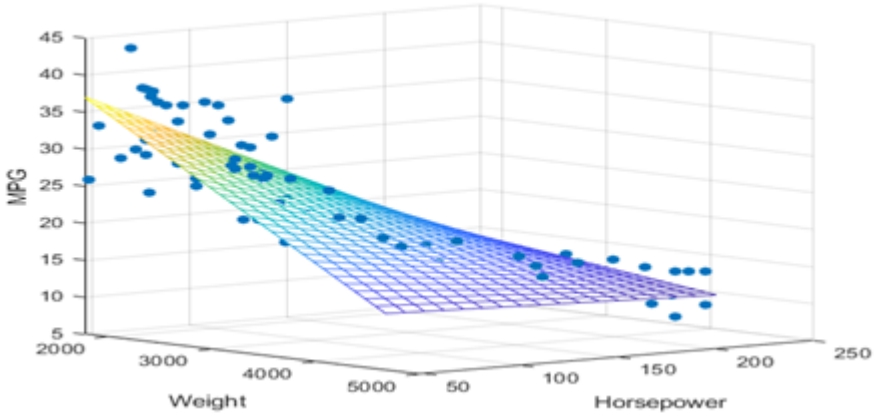
- 해석

  x_j**를 제외한 나머지** X**변수들을 고정시킨 상태**에서 x_j가 한 단위 증가할 때 y가 β_j만큼 증가

  → 이미 다른 변수들이 설명하는 부분 이외에 나머지를 x_j가 설명하는 것이다.

다른 변수들이 고정되어 있다는 것은, 다른 변수들이 상수처럼 취급된다는 것을 의미한다. 다른 변수들을 상수화한 상태에서 우리가 관심을 가지는 변수의 기울기(영향)를 확인하는 개념이다.

- 다중선형회귀는 데이터를 설명하기 위해 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 2개라면 평면을, 3개라면 공간을 회귀모형으로 가짐

2) 모수의 추정 - 최소제곱법(LSE)

단순선형회귀에서는 추정해야 할 모수가 2개(β_0, β_1) 였기에 각각에 대한 편미분을 진행해서 연립방정식을 풀어주었다. → 하지만 다중선형회귀에서는 추정해야 할 모수가 $p+1$ 개이기 때문에 편미분을 통해 구할 경우 계산식이 복잡해져서 행렬식을 이용한다.

단순선형회귀와 동일하게 **최소제곱법**을 이용한다.
오차의 특정 조건 아래에서
Best(분산이 제일 작은) Linear(선형) Unbiased Estimator(불편추정량) 성질이 성립하고, 오차의 정규분포 가정이 있다면 **최대가능도추정법(MLE)도 같은 결과**를 도출한다.

- 최소제곱법을 이용한 모수 추정

  $\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\epsilon}$ 꼴의 행렬을 통해 우리는 회귀식을 표현할 수 있다.

  - 회귀식

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\epsilon} \quad \Longleftrightarrow \quad \begin{pmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{pmatrix} = \begin{pmatrix} 1 & x_{11} & \cdots & x_{1p} \\ 1 & x_{21} & \cdots & x_{2p} \\ \vdots & \ddots & & \vdots \\ 1 & x_{n1} & \cdots & x_{np} \end{pmatrix} \begin{pmatrix} \beta_0 \\ \beta_1 \\ \vdots \\ \beta_p \end{pmatrix} + \begin{pmatrix} \epsilon_0 \\ \epsilon_1 \\ \vdots \\ \epsilon_n \end{pmatrix}$$

$$\begin{bmatrix} y_1 \\ y_2 \\ y_3 \\ . \\ . \\ . \\ y_n \end{bmatrix} = \begin{bmatrix} \beta_0 + \beta_1 X_{1,1} + \beta_2 X_{1,2} + \cdots + \beta_p X_{1,p} + \epsilon_1 \\ \beta_0 + \beta_1 X_{2,1} + \beta_2 X_{2,2} + \cdots + \beta_p X_{2,p} + \epsilon_2 \\ \beta_0 + \beta_1 X_{3,1} + \beta_2 X_{3,2} + \cdots + \beta_p X_{3,p} + \epsilon_3 \\ . \\ . \\ . \\ \beta_0 + \beta_1 X_{n,1} + \beta_2 X_{n,2} + \cdots + \beta_p X_{n,p} + \epsilon_n \end{bmatrix}$$

  - 목적함수

  $J(\boldsymbol{\beta}) = \sum_{i=1}^{n} e_i^2 = \boldsymbol{e^T e} = (\mathbf{y} - \mathbf{X}\boldsymbol{\beta})^T (\mathbf{y} - \mathbf{X}\boldsymbol{\beta})$

  목적함수를 β에 대해 미분하여 '**미분값=0**'으로 두고 β의 추정량 $\hat{\boldsymbol{\beta}}$을 구할 수 있다.

  - 추정량

  $\hat{\boldsymbol{\beta}} = (\mathbf{X^T X})^{-1} \mathbf{X^T y}$

$$RSS = y^T y - y^T X\hat{\beta} - \hat{\beta}X^T y + \hat{\beta}^T X^T X\hat{\beta}$$

$$\frac{\delta(RSS)}{\delta\hat{\beta}} = \frac{\delta\left(y^T y - y^T X\hat{\beta} - \hat{\beta}X^T y + \hat{\beta}^T X^T X\hat{\beta}\right)}{\delta\hat{\beta}} = 0$$

$$\frac{\delta(y^T y)}{\delta\hat{\beta}} - \frac{\delta\left(y^T X\hat{\beta}\right)}{\delta\hat{\beta}} - \frac{\delta\left(\hat{\beta}X^T y\right)}{\delta\hat{\beta}} + \frac{\delta\left(\hat{\beta}^T X^T X\hat{\beta}\right)}{\delta\hat{\beta}} = 0$$

$$0 - y^T X - (X^T y)^T + 2\hat{\beta}^T X^T X = 0$$

$$0 - y^T X - y^T X + 2\hat{\beta}^T X^T X = 0$$

$$2\hat{\beta}^T X^T X = 2y^T X \qquad \hat{\beta}^T X^T X = y^T X$$

$$\hat{\beta}^T = y^T X(X^T X)^{-1} \qquad \hat{\beta} = (X^T X)^{-1} X^T y$$

추정량 구하는 과정!

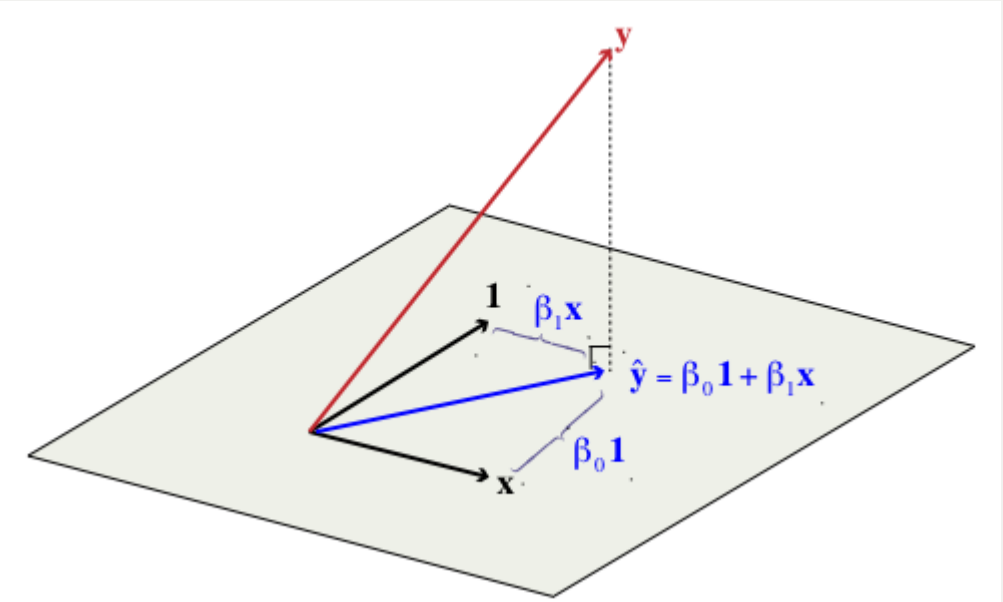
- 추정된 회귀식

$$\hat{\mathbf{Y}} = \mathbf{X}\hat{\boldsymbol{\beta}} = \mathbf{X}(\mathbf{X}^\mathbf{T}\mathbf{X})^{-1}\mathbf{X}^\mathbf{T}\mathbf{y} = \mathbf{H}\mathbf{y}$$

※ 여기서 $H = \mathbf{X}(\mathbf{X}^\mathbf{T}\mathbf{X})^{-1}\mathbf{X}^\mathbf{T}$ 는 투영 행렬(projection matrix)

✅ **투영 행렬(Projection matrix)** (선대팀 2주차 참고)

선형회귀분석에서의 목표는 $Y = X\boldsymbol{\beta}$를 만족하는 β를 찾는 것이다. 해가 존재하기 위해서는 y가 X의 열공간에 있어야 한다. 하지만 대부분의 경우 그렇지 않기 때문에, 우리는 y를 X의 열공간에 가장 가깝게 근사시켜야 하고 이때 투영 행렬 H가 사용된다. y를 X의 열공간에 투영시킴으로써 $\hat{\mathbf{Y}} = HY = \mathbf{X}\hat{\boldsymbol{\beta}}$ 를 만족시키는 근사해 $\hat{\boldsymbol{\beta}}$를 찾을 수 있게 된다.



y를 X가 존재하는 공간에 projection함으로써 해가 존재하는 선형방정식이 됨.

해는 $\hat{\beta}$

(열공간에 직교하도록 투영되기 위해서는 내적값이 0이어야 함)

3) 적합성(Goodness of fit) 검정

선택한 변수들로 적합한 회귀모델이 데이터에 얼마나 잘 들어맞는지 확인하기 위해 적합성 검정이 필요하다. 단순선형회귀처럼 SST, SSR, SSE를 이용하여 적합성을 판단하지만 다른 부분이 있다.

- **결정계수 (R square)** R^2

$$R^2 = \frac{SSR}{SST} = 1 - \frac{SSE}{SST}$$

단순선형회귀에서 R^2를 이용하여 설명력을 측정해 볼 수 있었다.

그러나 R^2에는 한계점이 있다. **바로 설명변수가 추가되면 반드시 R^2 값이 증가한다는 것**이다.

(총 변동(SST)은 고정되어 있는데, 독립변수가 추가되면 회귀식으로 설명되는 변동이 조금이라도 증가한다) 따라서 R^2을 다중선형회귀에 그대로 활용한다면 설명변수를 가장 많이 포함한 모형을 최적의 모형으로 판단할 것 이다.

하지만 무의미한 변수의 추가는 모델에 대한 해석도 어렵게 하고 예측에도 좋지 않은 영향을 끼칠 수 있다. 이 문제를 보완하기 위해 등장한 것이 **수정결정계수**이다.

- **수정결정계수 (Adjusted R square)** R^2_{adj}

$$R^2_{adj} = \frac{SSR/p}{SST/(n-1)} = 1 - \frac{SSE/(n-p-1)}{SST/(n-1)}$$

- 변수가 추가됨에 따라 증가하는 결정계수에 **변수 개수에 대한 페널티**를 부과한 형태
- R^2_{adj}가 높은 회귀식이 더 좋은 회귀식

- 수정결정계수(R^2_{adj})의 활용
  - AIC(Akaike Information Criterion), BIC(Bayesian Information Criterion)처럼 **변수의 개수가 다른 두 회귀식을 비교**할 때 유용하게 사용할 수 있다. (회귀팀 2주차 예정)
  - 그러나 결정계수처럼 '전체 변동 중에 회귀식이 설명하는 변동'으로 해석할 수는 없다.

4) 유의성 검정

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

- **F-test (전체 회귀계수에 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

$H_0 : \beta_0 = \beta_1 = \cdots = \beta_p = 0$ (모든 회귀계수는 0이다)

$H_1 : not\ H_0$ (적어도 한 개의 회귀 계수는 0이 아니다) → 귀무가설이 기각되어야 의미있는 모형!

2. 검정통계량

$$F_0 = \frac{(SST-SSE)/p}{SSE/(n-p-1)} = \frac{SSR/p}{SSE/(n-p-1)} = \frac{MSR}{MSE}$$

MSR : 평균회귀제곱, MSE : 평균오차제곱

- 검정통계량의 의미

단순선형회귀의 적합성 검정에서 살펴본 SST, SSR, SSE를 유의성 검정에도 적용해보자. 귀무가설을 기각시키기 위해서는 검정통계량 F_0가 충분히 커야한다. F_0이 크다는 것은 분자(SSR, 추정된 회귀식이 설명하는 부분)가 분모(SSE, 추정된 회귀식이 설명하지 못하는 부분)보다 꽤 크다는 것이다. (자유도로 나눠주는 것은 일단 배제하고 이해하자)

→ 검정통계량 F_0는 회귀식의 전반적인 계수가 얼마나 설명력을 갖는지를 의미한다.

3. 임계값

$F \geq \ F_{(1-a/2,\ p,\ n-p-1)}$: 검정통계량이 임계값보다 크다면 귀무가설을 기각할 수 있다.

다음은 R에서 summary 함수를 통해 확인할 수 있는 회귀분석 결과이다.

```
> lm.fit <- lm(Sales ~ TV + Radio + Newspaper, data=advertising)
> summary(lm.fit)

Call:
lm(formula = Sales ~ TV + Radio + Newspaper, data = advertising)

Residuals:
    Min      1Q  Median      3Q     Max
-8.8277 -0.8908  0.2418  1.1893  2.8292

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)  2.938889   0.311908   9.422   <2e-16 ***
TV           0.045765   0.001395  32.809   <2e-16 ***
Radio        0.188530   0.008611  21.893   <2e-16 ***
Newspaper   -0.001037   0.005871  -0.177     0.86
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 1.686 on 196 degrees of freedom
Multiple R-squared:  0.8972,    Adjusted R-squared:  0.8956
F-statistic: 570.3 on 3 and 196 DF,  p-value: < 2.2e-16
```

R-squared결과를 통해 모델이 주어진 데이터에 대해 얼마만큼의 설명력을 지니는지(독립변수가 종속변수를 얼마나 설명해주는지), P-value를 통해 전체 데이터를 예측하는데 얼마만큼의 성능이 있는지 파악할 수 있다.

이 예시의 경우 p-value가 작아 귀무가설을 기각할 수 있고, 적어도 하나의 회귀계수는 의미가 있어 모델이 틀리지 않았음을 알 수 있다 (*의 개수로 유의미함의 정도를 표현)

- **Partial F-test (일반화된 F-test, 일부 회귀계수에 대한 검정)**

F-test는 다중회귀모델 내의 모든 회귀계수에 대해 한번에 유의성을 검정하지만, Partial F-test는 전체 계수 중 일부에 대해서만 유의성을 검정한다.
(특정 변수의 조합에 대해서도 유의성 검정을 할 수 있어 일반화된 F-test)

모든 변수를 사용한 다중 회귀모형을 완전모형(FM: Full Model), 일부 회귀계수를 특정한 값(보통 0)으로 두는 축소모델(RM: Reduced Model)로 설정하고, 이때 FM이 좋은지 RM이 좋은지 검정하는 방식이다.

1. 가설설정

$$FM : y = \beta_0 + \beta_1 x_1 + \ldots \qquad \ldots + \beta_p x_p$$

$$RM : y = \beta_0 + \beta_1 x_1 + \ldots + \beta_{j-1} x_{j-1} + \beta_{j+q} x_{j+q} + \ldots + \beta_p x_p$$

$H_0 : \beta_j = \beta_{j+1} = \cdots = \beta_{j+q-1} = 0$ (RM이 맞다)

$H_1 : not\ H_0$ (FM이 맞다. 적어도 한 개의 회귀계수는 0이 아니다)

2. 검정통계량

$$F_0 = \frac{(SSE(RM)-SSE(FM))/(q)}{SSE(FM)/(n-p-1)} = \frac{(SSR(FM)-SSR(RM))/(q)}{SSE(FM)/(n-p-1)} \sim F_{q,n-p-1}$$

- 검정통계량의 의미

Partial F-test에서 귀무가설을 기각시키기 위해서는 검정통계량 F_0이 충분히 커야한다. F_0이 충분히 커지기 위해서는 $SSE(RM)$이 $SSE(FM)$보다 충분히 커야한다.

변수가 제거된다면 당연히 SSE는 커진다. 그러나 제거된 변수가 의미있는 변수라면, SSE는 매우 커질 것이며 이는 F_0가 귀무가설을 기각시킬 만큼 **충분히** 커질 것이다.

즉, 귀무가설 기각을 위해서는 기본적으로 $SSE(RM) >>>> SSE(FM)$이 되어야 한다.

3. 임계값

$F \geq F_{(1-a/2,\ q,\ n-p-1)}$

F-test와 마찬가지로 검정통계량이 임계값보다 크면 귀무가설을 기각할 수 있다.
Partial F-test가 더 일반화된 검정이기는 하지만, 보편적으로는 회귀식 전체에 대한 검정을 사용함

- **t-test : 개별 회귀계수의 유의성 검정**

1. 가설설정

$H_0 : \beta_j = 0$ (다른 변수들이 다 적합된 상태에서 x_j는 통계적으로 유의하지 않다.)

$H_1 : \beta_j \neq 0$ (다른 변수들이 다 적합된 상태에서 x_j는 통계적으로 유의하다.)

2. 검정통계량

$$t_j = \frac{\hat{\beta}_j}{s.e.(\hat{\beta}_j)}$$

3. 임계값

$$|t_j| \geq t_{(\alpha/2,\, n-p-1)} \quad , \quad reject \quad H_o$$

- 해석

t-test는 해당 변수 자체가 유의미한지를 확인하는 것이 아니라, 다른 변수들이 다 적합된 상태에서 x_j**를 추가적으로 적합하는 것이 유의미한 회귀식의 설명력 증가를 가져오는지**를 확인하는 것이다

→ 회귀분석에서 개별 변수에 대한 검정은 한 변수에 대한 Partial F-test와 완전히 동일하다.
(t분포를 제곱하면 F분포와 같은 형태임!)

t-test로 변수를 선택하는 것은 매우 위험하다. 다른 변수들이 이미 고정되어 있는 상황에서 변수의 유의성을 판단하는 것이기때문에, 다른 회귀식을 가정했을 때는 해당 변수가 유의할 수도 있기 때문이다. 그러므로 변수 선택법을 이용하는 것이 더 좋은 방법이다. (3주차 예정)

```
> lm.fit <- lm(Sales ~ TV + Radio + Newspaper, data=advertising)
> summary(lm.fit)

Call:
lm(formula = Sales ~ TV + Radio + Newspaper, data = advertising)

Residuals:
    Min      1Q  Median      3Q     Max
-8.8277 -0.8908  0.2418  1.1893  2.8292

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)  2.938889   0.311908   9.422   <2e-16 ***
TV           0.045765   0.001395  32.809   <2e-16 ***
Radio        0.188530   0.008611  21.893   <2e-16 ***
Newspaper   -0.001037   0.005871  -0.177     0.86
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 1.686 on 196 degrees of freedom
Multiple R-squared:  0.8972,    Adjusted R-squared:  0.8956
F-statistic: 570.3 on 3 and 196 DF,  p-value: < 2.2e-16
```

아까의 예시를 한번 더 살펴보자. F-test 값을 먼저 확인했을 때, 회귀식에 대한 귀무가설을 기각할 수 있었다. 이후 개별 변수에 대한 t-test 결과를 보면, TV와 Radio 변수는 ‘$\beta_j = 0$’이라는 귀무가설을 기각하고, Newspaper는 기각하지 못한다는 사실을 알 수 있다. 이는

‘다른 변수들(Radio, Newspaper)이 적합된 상황에서, TV를 추가적으로 적합하는 것은 회귀식 설명력을 통계적으로 유의미하게 증가시킨다.’
‘다른 변수들(TV, Radio)이 적합된 상황에서, Newspaper를 추가적으로 적합하는 것은 회귀식 설명력을 통계적으로 유의미하게 증가시키지 않는다.’ 로 해석할 수 있을 것이다.

- **F-test vs. t-test**

  F-test와 t-test 중에서는 회귀 모델 전체에 대한 F값을 먼저 확인해야 한다. 전체 모델에 대한 F값을 확인함으로써 모델 전체가 통계적으로 유의한지 확인할 수 있으며, 더욱 일반적인 검정이기 때문. 추가로 아래의 이유 때문에 F-test를 먼저 수행해야 한다.

  1) 전체 회귀식에 대한 검정이 더 엄격(rigorous)하기 때문에 F를 기각하지 못한 상태에서 t-test를 보는 것은 아무 의미가 없다.

  2) 또, F를 기각하지 못해도 t는 기각하는 경우가 있을 수 있기 때문이다

---

4. 데이터 진단

설정된 회귀모형을 적합시키는 것으로 회귀분석이 끝나는 것은 아니다. 설정된 회귀모형이 여러 측면에서 타당한 것인지를 검토해야 하는데, 데이터 중에 일반적인 경향에서 벗어나는 점들이 있기 때문이다. 이러한 점들은 최소제곱 회귀모형을 크게 바꾸거나, 그에 따라 성능을 저하시키기도 한다. 일부 점의 영향력이 큰 상황을 다루는 법을 알아보자.

설정된 회귀모형을 적합시키는 것으로 회귀분석이 끝나는 것은 아니다. 설정된 회귀모형이 여러 측면에서 타당한 것인지를 검토해야 하는데, 데이터 중에 일반적인 경향에서 벗어나는 점들이 있기 때문이다. 이러한 점들은 최소제곱 회귀모형을 크게 바꾸거나, 그에 따라 성능을 저하시키기도 한다. 일부 점의 영향력이 큰 상황을 다루는 법을 알아보자.

1) 표준화 잔차

- **잔차**

  잔차(e)는 **설명할 수 없는 오차(ϵ)의 추정치**로, 관측된 종속변수(y)와 추정된 종속변수($\hat{y}$)의 차($y - \hat{y}$)으로 구해진다. 설명할 수 없는 오차 ϵ을 실제로 관측할 수는 없으므로, 그에 대한 추정치를 통해 데이터를 진단하거나 오차에 대한 검정을 진행한다.

$$\mathbf{e} = \mathbf{y} - \hat{\mathbf{y}} = \mathbf{y} - \mathbf{X}\hat{\boldsymbol{\beta}} = \mathbf{y} - \mathbf{X}(\mathbf{X'X})^{-1}\mathbf{X'y} = \mathbf{y} - \mathbf{Hy} = (\mathbf{I} - \mathbf{H})\mathbf{y}$$

$$Var(\mathbf{e}) = Var((\mathbf{I}-\mathbf{H})\mathbf{y}) = (\mathbf{I}-\mathbf{H})\sigma^2(\mathbf{I}-\mathbf{H})' =$$
$$\sigma^2(\mathbf{I}-\mathbf{H})(\mathbf{I}-\mathbf{H})' = \sigma^2(\mathbf{I}-\mathbf{H})$$
$$(\because \mathbf{I}-\mathbf{H} \text{ is symmetric and idempotent})$$

→ idempotent : 멱등성 (연산을 여러 번 반복해도 결과값이 달라지지 않는 성질)

$$Var(e_i) = (I - h_{ii})\sigma^2$$
$$\sigma^2(\mathbf{I}-\mathbf{H}) = \begin{pmatrix} 1-h_{11} & \cdots & -h_{1n} \\ \vdots & \ddots & \vdots \\ -h_{n1} & \cdots & 1-h_{nn} \end{pmatrix} \cdot \sigma^2$$

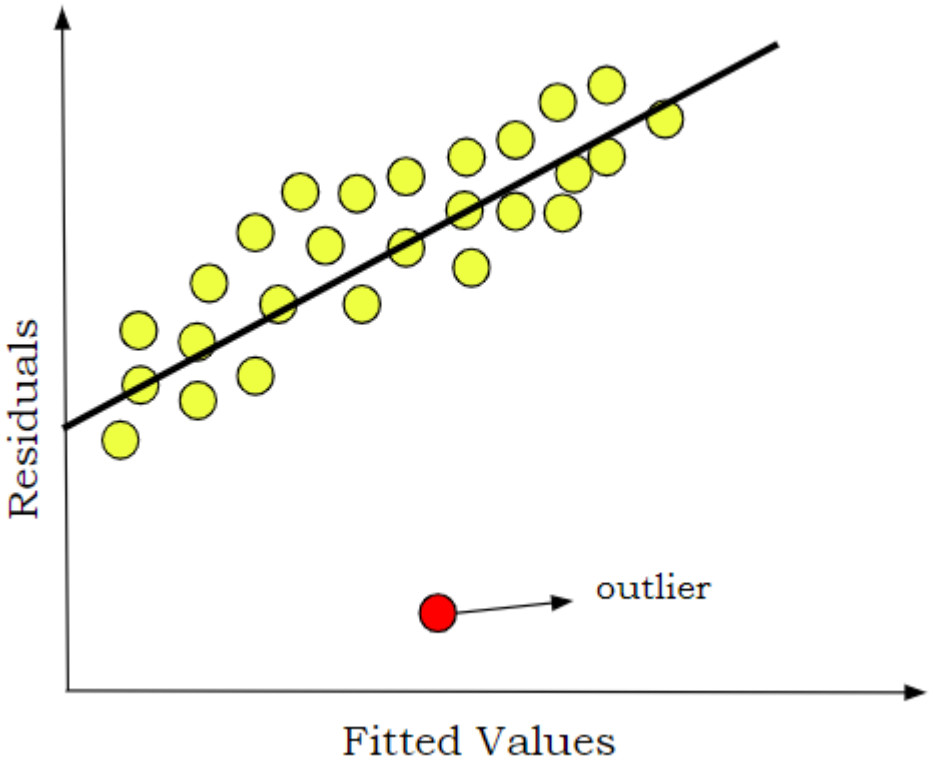
- **스튜던트화 잔차(Studentized residual)**

  잔차는 y값의 단위에 영향을 많이 받기 때문에, 일반화해서 적용할 수 있도록 표준화가 필요하다. σ 는 모수이므로 알 수 없어 이에 대한 **추정량인 $\hat{\sigma}$ (s)을 넣어주게 된다.**

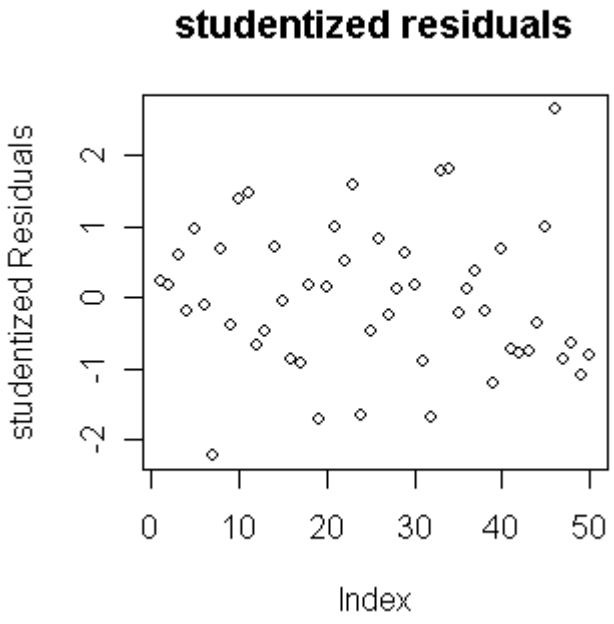
$$r_i = \frac{e_i}{\hat{\sigma}\sqrt{1-h_{ii}}}\ , \quad \hat{\sigma} = \sqrt{\frac{SSE}{n-p-1}}$$

2) 이상치 (Outlier)

표준화 잔차가 매우 큰 값을 의미한다. (y를 기준으로 절댓값이 큰 값)



- 보통 $|r_i| > 3$ 이면 이상치라고 판단한다. 편의상 r_i가 정규분포를 따른다고 가정하여 이 값을 설정하는데, 사실 정규분포를 따르지는 않는다. 자세한 내용은 부록 확인!!



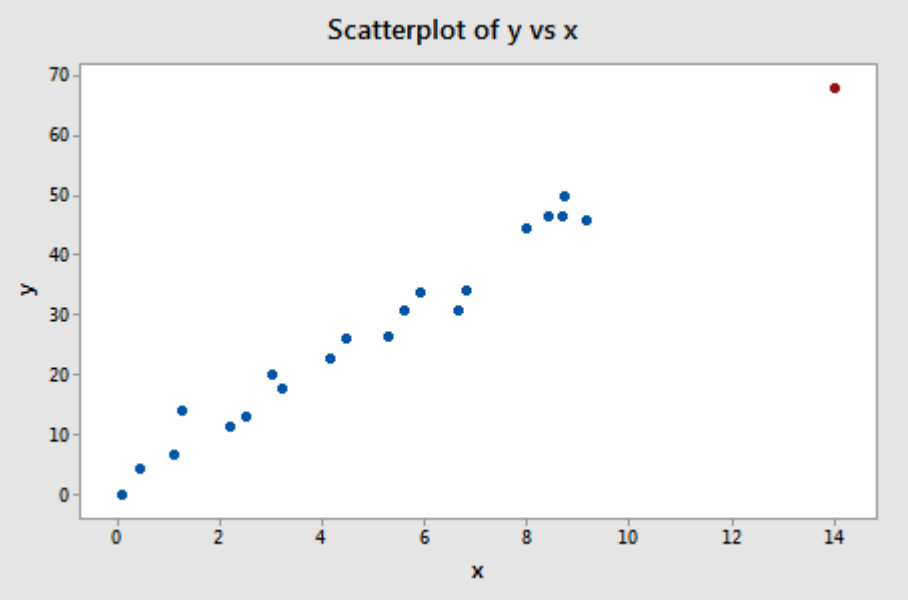
3) 지렛값 (Leverage point)

x의 평균 $\bar{x}$에서 멀리 떨어져 있어 기울기에 큰 영향을 주는 값

Outlier가 y의 관점이었다면, Leverage point는 (표준화했을 때) x의 기준에서 절댓값이 큰 값!

$H = X(X^TX)^{-1}X^T, \qquad h_{ii} = {x_i}^t(X^tX)^{-1}x_i$ 일때,

- 투영행렬 H의 대각요소를 $h_{ii} = \frac{1}{n} + \frac{(x_i-\bar{x})^2}{\sum(x_i-\bar{x})^2}$ 라고 표현할 수 있다.
- x_i 값과 $\bar{x}$의 차이가 클수록 h_{ii} 가 커진다. → x 평균에서 멀수록 레버리지 값이 상승한다.
- $h_{ii} > \frac{2(p+1)}{n}$ 이면 지렛값으로 판단한다.



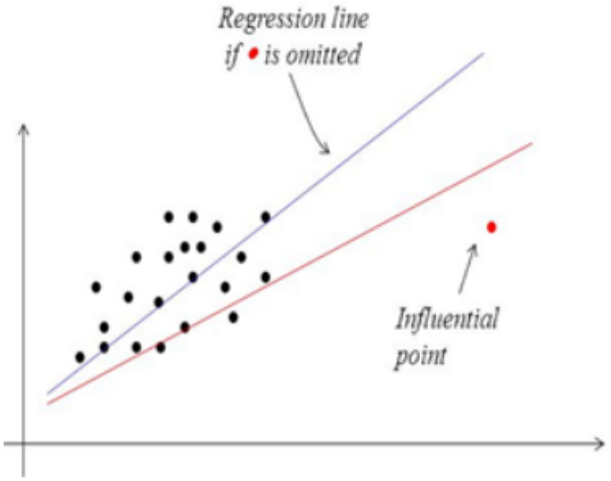
4) 영향점 (Influential point)

한 관측치가 회귀직선의 기울기에 상당한 영향을 줄 때, 이 점을 영향점이라고 한다.

사실 Outlier와 Leverage point의 진단만으로 회귀직선이 변한다고 말하기가 어렵다.

- Outlier일지라도 x평균 주위에 위치할 경우 기울기를 변화시키지 못하고,
- Leverage point라고 하더라도 회귀직선의 연장선에 있을 수 있기 때문이다.

→ Outlier와 Leverage point를 동시에 고려하는 지표가 필요하고, 이를 **Cook’s distance** 라고 한다.



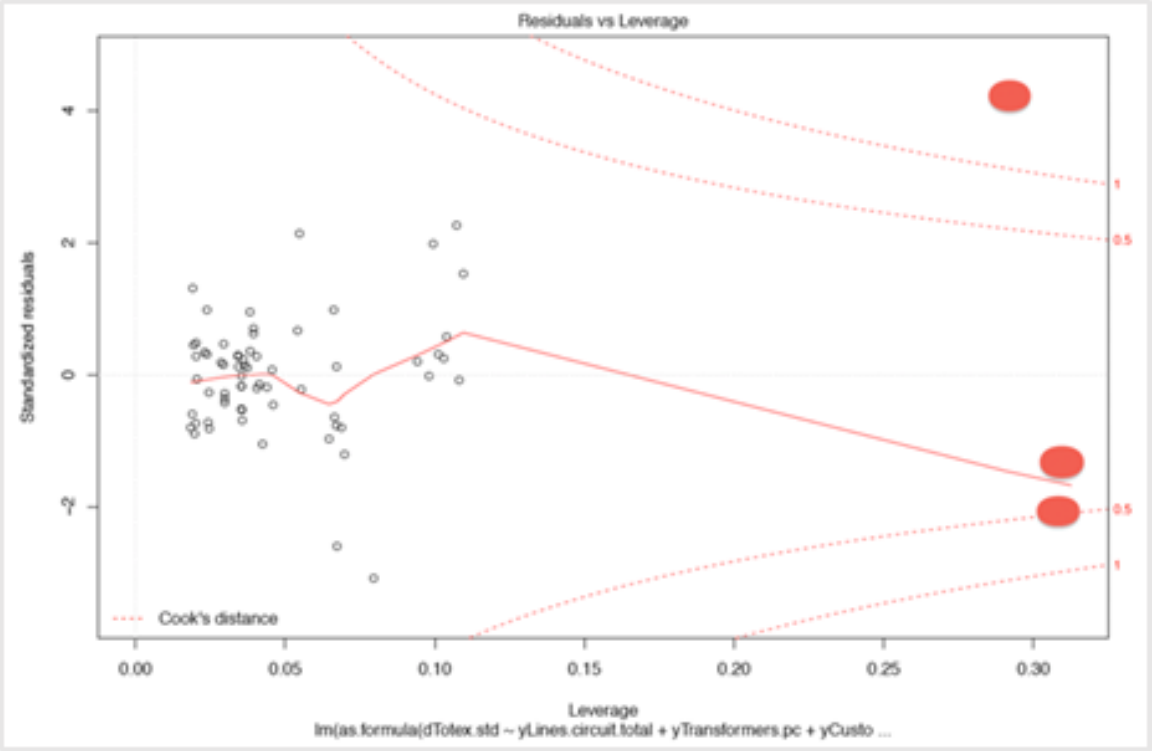
- Cook’s distance

: 영향점을 확인하는 표준적인 지표로, 특정 데이터를 지웠을 때 회귀선이 변하는 정도를 가리킴

$$C_i = \frac{r_i^2}{p+1} \times \frac{h_{ii}}{1-h_{ii}}$$

- Outlier와 Leverage 각각이 커지면 커질수록 C_i가 커진다.

- 보통 $C_i > 1$ 이면 영향점으로 판단한다. ($\frac{4}{(n-p-1)}$를 사용하기도 한다.)



- R에서 회귀식을 적합하고, 적합식에 대해 plot을 그리면 쉽게 Cook's distance를 확인할 수 있다.

 - 그림에서 볼 수 있듯이 Leverage point라고 하더라도 모두 Influential point인 것은 아니다.

(그러나 Leverage point중에 Cook's distance가 1보다 커서 Influential point인 점이 하나 존재하는 것을 plot에서 확인할 수 있다.)

- 영향점의 처리

영향점은 추정량을 불안정(분산을 크게)하게 만들고, 이는 잘못된 모델의 해석과 예측 성능 저하를 일으킬 수 있기 때문에 적절한 처리가 필요하다.

→ But! 의미가 있는 데이터일 수도 있으므로, 데이터를 삭제하는 것은 늘 조심해야 한다.

따라서 이를 고려한, **이상치에 강건한(robust) 모델링** 방법도 필요하다.

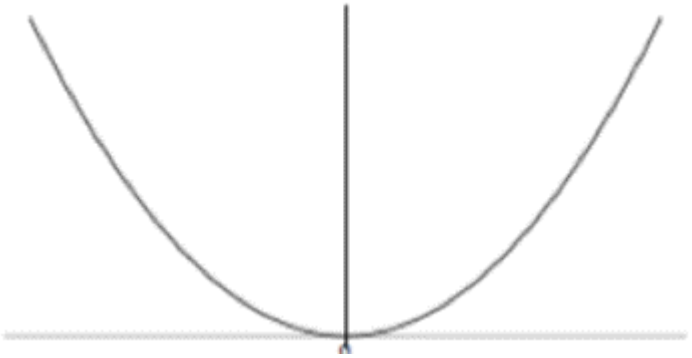
5. 로버스트 회귀

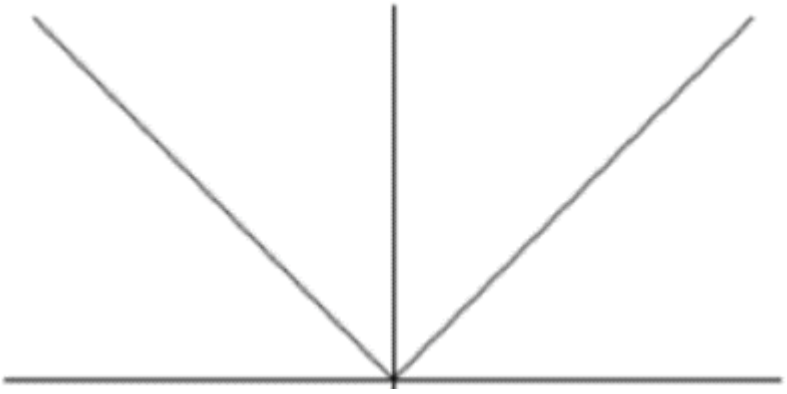
로버스트 회귀는 이상치의 영향을 줄이는 회귀분석 방법이다. 이상치의 영향을 줄이는 방법은 매우 다양하고, 이에 따라 다양한 모델들이 존재한다.

1) Median Regression

- 최소제곱회귀는 $\sum {\varepsilon_i}^2 = \quad (y - X\beta)^t \, (y - X\beta)$를 최소화하는 β를 찾는 방법이다.이는 제곱을 최소화하기 때문에 미분이 편리하지만, 이상치에 대해 너무 큰 가중치를 주는 경향이 있다.

  → 어떤 경우더라도 동일한 가중치를 주는 선택지도 고려해볼 수 있다.

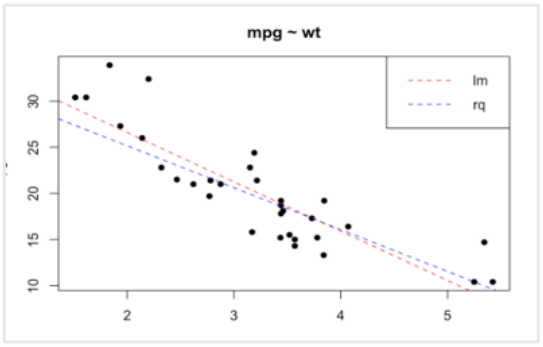


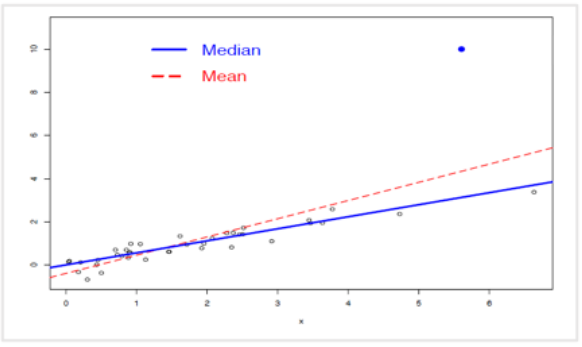


- 최소제곱법은 오차의 제곱합을 최소로 하는 추정량을 찾는다. 반면 Median regression은 오차의 절대값의 합을 최소화한다. 즉, $\sum |\epsilon_i|$을 최소화하는 것.
- 최소제곱회귀는 X에 따른 평균적인 Y를 반환한다면, median regression은 X에 따른 Y의 **중앙값**을 반환한다.

  → 다중선형회귀는 독립변수X의 변화에 따른 종속변수 Y의 조건부 평균($E(Y|X)$)을 추정하지만, Median regression은 **조건부 중앙값**을 추정한다. 이를 통해서 중심에서 멀리 떨어진 이상치에 덜 민감한 추정량을 가질 수 있다.

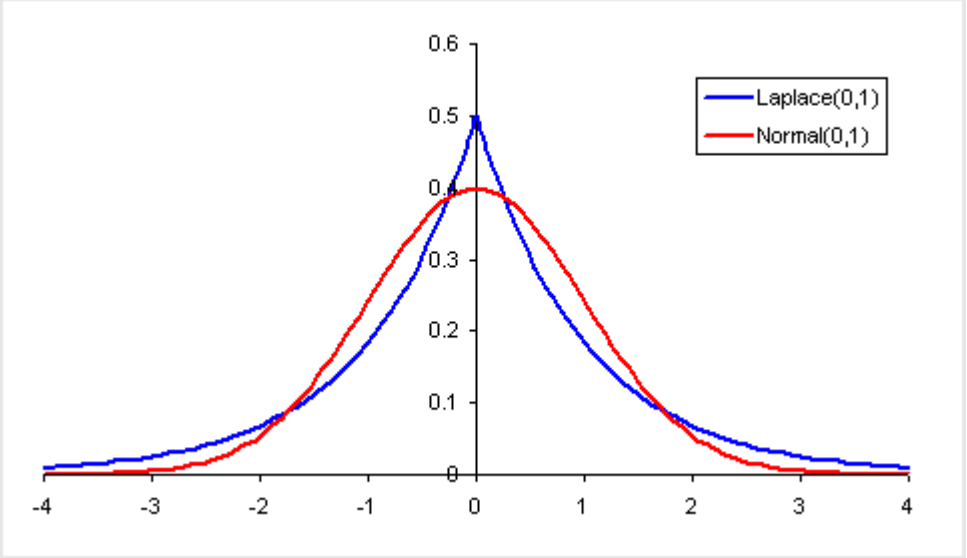
  (대표값인 평균, 중앙값, 최빈값에 대해 배울 때, 중앙값은 이상치의 영향을 덜 받는다고 배웠던 걸 상기해보면 좋다)





- 분포 가정과 등분산 가정이 없는 모델이다
- R에서는 'quantreg' 패키지의 `rq()` 함수 사용

전에 오차항이 정규분포를 따른다고 가정했을 때, LSE와 MLE가 같은 추정량을 산출한다고 했었다. 오차항이 정규분포가 아닌 Laplace(double exponential) 분포 $\left(f(x) = \frac{1}{2b}exp(-\frac{|x-u|}{b})\right)$ 를 따른다고 가정했을 때, MLE로 추정한 값과 Median Regression으로 추정한 값 또한 같다!



참고로 분포를 살펴보면 라플라스 분포가 정규분포보다 긴 꼬리분포를 지니는데, 정규분포의 관점에서 이상치가 많은 형태라고 볼 수 있기도 하다. 신기하죠? ㅎㅎ

2) M-estimation

최소제곱회귀를 위해 $\sum_{i=1}^{n}(y_i - \beta_0 - \beta_1 x_i)^2$ 를 최소화하며 계수를 추정했던 것을 기억할 것이다. 최소제곱회귀는 이상치에 지나치게 큰 패널티를 부여하지만, 동시에 적정 수준 안에서는 패널티를 완화시켜준다.

이때 최적화(최소화)하는 목적함수로 $\sum_{i=1}^{n}(y_i - \beta_0 - \beta_1 x_i)^2$ 대신 $\sum \rho$ 를 설정하게 된다.
(
ρ는 함수이고, 어떻게 설정하느냐에 따라 달라진다.) 이렇게 구한 추정량을 M-estimator라고 부른다.

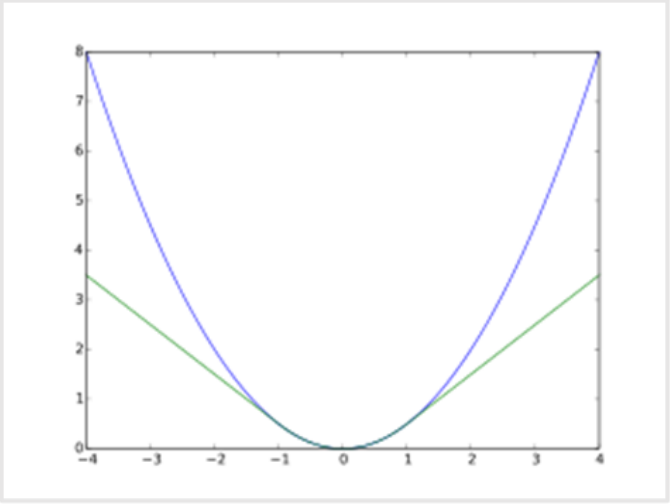
그중 대표적으로 회귀를 로버스트하게 모델링할 수 있는 **Huber's M-estimation**을 살펴보자.

ρ 함수를

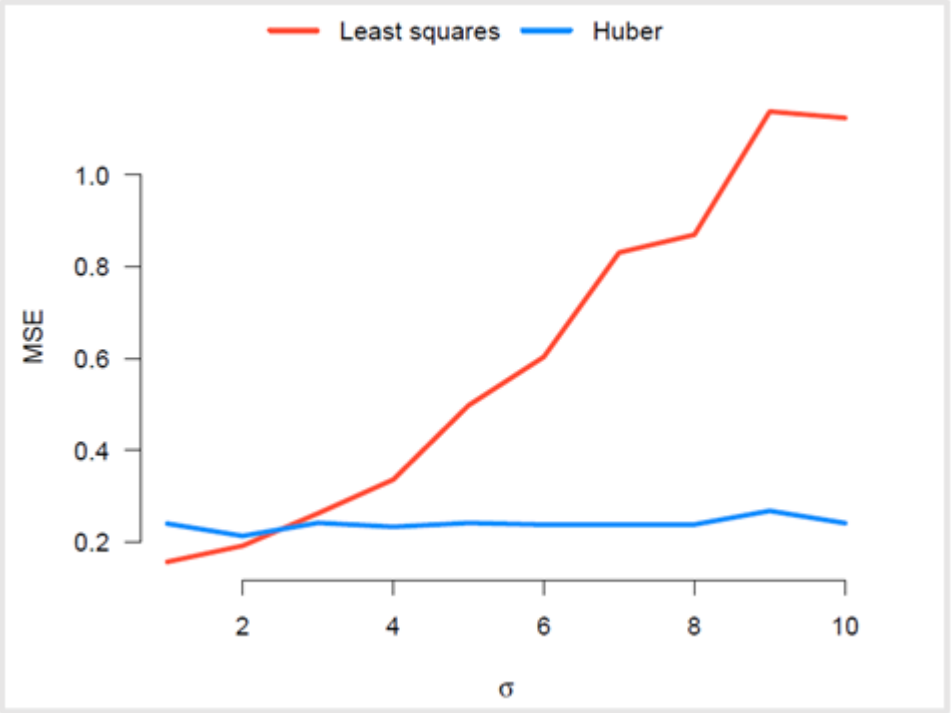
$$\rho(e) = \begin{cases} \frac{1}{2}e^2, & if\ |e| \leq\ c \\ c|e| - \frac{1}{2}c^2, & otherwise \end{cases}$$

다음과 같이 설정하게 되면, 잔차가 특정 상수값보다 크면 잔차의 '제곱'이 아닌 1차식으로 바꾸어서 이상치에 강건한 회귀계수를 추정할 수 있다. 잔차의 절대값이 c 이하면 기존 최소제곱추정법의 목적함수와 동일하지만, c 이상이면 이상치에 대한 큰 페널티를 부여하지 않도록 일차식의 형태를 적용하는 것.

이렇게 적정수준의 페널티를 완화시켜주는 형태는 유지하되, 이상치에 대해 지나친 페널티를 부여하는 것은 막을 수 있다.



파란색 : 최소제곱회귀 초록색 : Huber's M-estimation



이상치에 대한 패널티 완화로 MSE값이 작다는 것을 확인할 수 있음

- R에서는 'MASS' 패키지의 `rlm()` 함수 사용
- 이외에도 Ramsay, Tukey, Hampel 등 많은 ρ 함수를 적용할 수 있다. 그렇지만.. (부록에 계속)

3) Least Trimmed Square

통계적 기준에 따라 잔차가 너무 큰 관측치를 제거하고 회귀계수를 추정하는 방식

- 공식

  $$\hat{\beta} = \min \sum_{j=1}^{h} r_{(j)}^2 \quad \begin{cases} r_1 \leq r_2 \leq \cdots \leq r_h \\ \frac{n}{2} + 1 \leq h \end{cases}$$

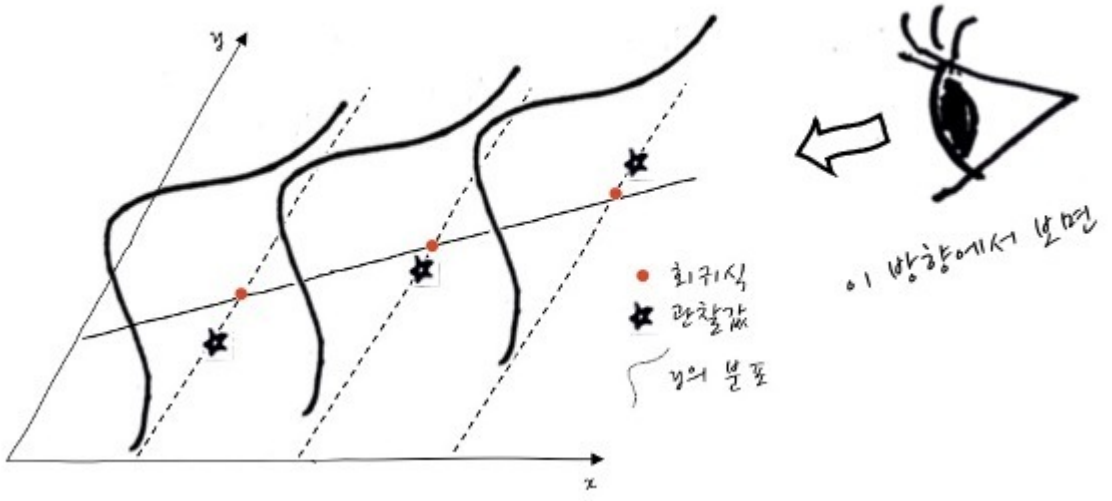
  이 때, $r_{(j)}$는 작은 순서부터 오름차순으로 나열한 잔차이다.
- n개의 obs 중 h개만 사용하여 회귀식을 만드는데, $\binom{n}{h}$개의 회귀식 중 가장 잔차제곱합이 작은 회귀식을 사용한다.
- obs가 별로 없는 경우, 혹은 영향점이 존재하지 않는 경우 주의해서 사용해야 한다.
- R에서는 'robustbase' 패키지의 `ltsReg()` 함수 사용

Appendix

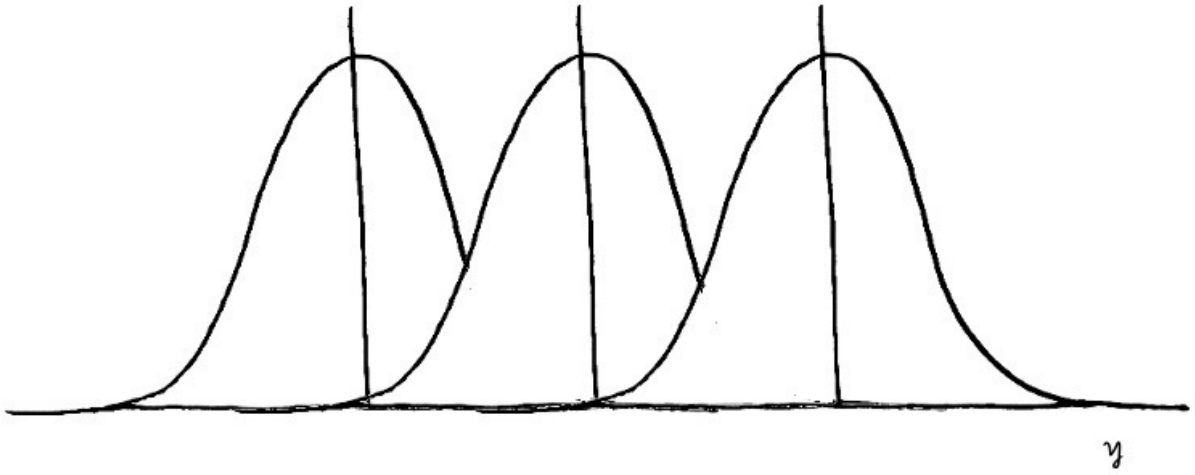
1. 유의성 검정과 ANOVA의 관계

$$F_0 = \frac{(SST - SSE)/p}{SSE/(n-p-1)} = \frac{SSR/p}{SSE/(n-p-1)} = \frac{MSR}{MSE}$$

실험계획법을 수강했거나 ANOVA에 대해 알고 있다면, 유의성 검정의 부분에서 다음과 같은 수식이 ANOVA에서의 F 통계량과 비슷하다는 것을 눈치챘을 것이다. 사실은 유사한 정도가 아니고 똑같은 수준인데, 이를 이해하기 위해 각 y의 분포를 3차원으로 표현해보자.



이를 오른쪽의 시선에서 바라보게 되면



이고, 이는 곧 집단 간 차이를 보게 되는 ANOVA와 같다는 것을 직관적으로 이해할 수 있다. 기울기가 존재한다면 저 연속형 집단들 간 차이가 분명 있을 것이다.

즉, Regression의 F검정은 연속형 집단에서 집단 간 최소한 1개의 변수에 의해 차이가 발생하는가? 와 같은 이야기를 다룬다는 것을 알 수 있다. ANOVA는 범주형 집단, Regression은 연속형 집단 차이를 다룬다고도 이해할 수 있는 것.

2. 내 표준화 잔차(Internally Studentized residual)와 외 표준화 잔차 (Externally Studentized residual)

우리가 4.-1)에서 배운 것은 내 표준화 잔차에 대한 것이었다.

그러나 Outlier가 존재하는 경우 그 이상치가 $\hat{\sigma}$에도 영향을 미쳐 그 값이 커지게 되고, 이상치 탐지가 제대로 되지 않을 수 있다. 이 경우 그 이상치 데이터를 제외한 후 n-1개의 관측치로 회귀모형을 적합시켰을 때 얻어지는 $\hat{\sigma}_{(i)}$를 표준화에 사용하는 아이디어를 생각할 수 있다. 즉

$$r_i^* = \frac{e_i}{\hat{\sigma}_{(i)}\sqrt{1-h_{ii}}}$$

를 정의할 수 있고, 이를 **외 표준화 잔차**(Externally Studentized residual)라고 한다.

내 표준화 잔차는 정규분포를 따르지 않는다. 편의상
$|r_i|$의 기준치를 정하기 위해 정규분포를 사용할 수도 있으나, 정확하게는 $r_i^2/(n-p)$가 Beta$(\frac{1}{2}, \frac{n-p-1}{2})$를 따른다고 알려져 있다. 반면 외 표준화 잔차는 우리가 잘 알고 있는 자유도 $(n-p-1)$의 t분포를 따른다는 장점이 있다. 또 관측치를 제거하여 다시 회귀모형을 적합시킬 필요 없이, 밑의 관계식을 통해 계산할 수 있다!

$$\hat{\sigma}_{(i)}^2 = \hat{\sigma}^2 \cdot \frac{n-p-r_i^2}{(n-p-1)}$$

3. 로버스트 회귀의 비용

Median Regression, M-estimation 모두 이상치에 강건하지만, 목적함수를 최적화하는 것이 어려워진다는 문제가 발생한다. 최솟값을 찾기 위해 수치적 방법을 사용해야 할 수도 있고, 반복 알고리즘을 사용해야 할 수도 있다. 또한 정규분포를 가정하지 않아 일반적인 해석이 어려워질 수도 있다.

분석 과정에서도 Trade-off 상황이 많이 발생한다. Robustness를 위해 계산비용을 포기할 수도 있고, 더 좋은 score를 위해 해석력을 포기하는 경우가 발생할 수도 있을 것이다. 따라

서 우리는 상황과 분석 목적을 항상 인지한 상태에서, 판단 및 근거의 설명이 가능해야 할 것이다.