1주차 - 회귀분석의 기초

목차

Table of Contents

작성자 : 권남택 Phone : 010-7518-8810 github.com/Namtaek

0. 기본 수식

(1) 기초 수식

- 표본 평균 (Sample Mean)

$$\bar{x} = \frac{1}{n}\sum_i^n x_i\,, \qquad \bar{y} = \frac{1}{n}\sum_i^n y_i$$

- 표본 분산 (Sample Variance)

$$S^2 = \frac{1}{n-1}\sum(x_i - \bar{x})^2$$

- 표본 표준편차 (Sample Standard Deviation)

$$S = \sqrt{\frac{1}{n-1}\sum(x_i - \bar{x})^2}$$

- S_{xx}와 S_{xy} (변동)

$$S_{xx} = \sum(x_i - \bar{x})^2\,, \qquad S_{xy} = \sum(x_i - \bar{x})(y_i - \overline{y})$$

(2) 공분산 (Covariance)

공분산은 두 확률변수가 얼마나 상관성을 띄는지 나타내는 지표. 두 변수의 선형관계를 나타내지만, 얼마나 선형성을 갖는 지에 대한 '강도'를 표현하진 못한다. 값은 −∞에서 +∞까지 가질 수 있다.

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

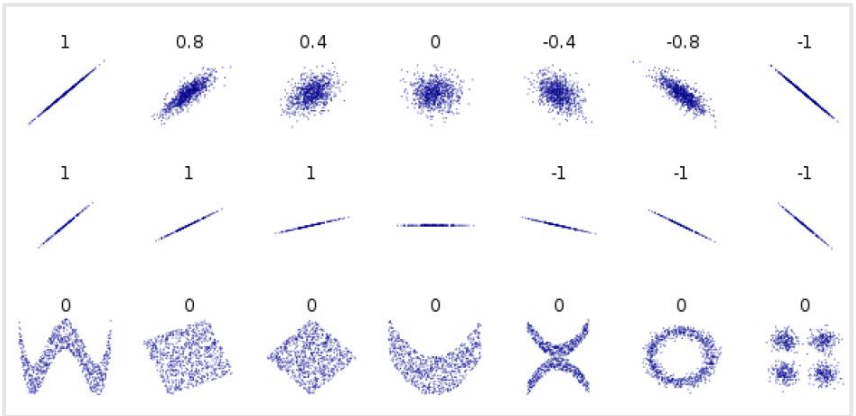
- 확률변수의 단위에 영향을 받는다. 공분산을 통해 상관성의 형태에 대한 짐작은 할 수 있지만, 정확한 정도를 표현하기에는 한계가 있다.

(3) 상관계수 (Correlation Coefficient)

공분산의 단점을 보완해줄 수 있는 '표준화된 공분산'

$$r_{xy} = \frac{Cov(X,Y)}{\sqrt{Var(X)}\sqrt{Var(Y)}} = \frac{\sum_i^n (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_i^n (x_i - \bar{x})^2}\sqrt{\sum_i^n (y_i - \bar{y})^2}}$$

- -1부터 1의 값을 갖고, 1이면 완전한 상향 직선, -1 이면 완전한 하향 직선의 형태다. 0이면 아무런 선형관계가 없다. 즉, Zero correlation이더라도 **선형관계 이외의 비선형 관계**가 있을 수 있다.



1. 회귀분석이란?

1) 회귀분석이란?

- 변수들 사이의 관계를 모델링하는 통계적 기법 by Duglas C. Montgomery
  - 예) 전통적인 주제 : 흡연과 폐암은 관련있을까?
  - 예) 주희의 음주량과 학점은 관련이 있을까?
- 회귀분석은 Francis Galton으로부터 시작되었다. 아버지의 신장(X)와 자녀의 신장(Y) 간의 관계식을 연구했는데, 산점도를 그려본 결과 완만한 직선을 중심으로 점들이 분포해 있었다. 그리고 그 추세선은 언제나 X와 Y의 평균을 지나고 있었다. 예를 들어 자녀 세대의 평균 신장이 172cm라고 하면, 아버지의 신장이 160cm일 때 자녀들의 평균 신장은 168cm 정도로 아버지보다 컸다. 또한 아버지의 신장이 180인 경우 자녀들의 평균 신장은 175cm로 아버지보다 작았다. 이렇듯 키가 큰 부모에게서 키가 큰 자녀가 태어나고 키가 작은 부모에게서 키가 작은 자녀가 태어나지만, 태어난 자녀들의 평균 키는 전체 평균 수준으로 회귀(돌아가다)하는 현상을 보인다.
- 우리가 관심있는 반응변수(Response) Y와 설명변수(Predictor, Feature) X의 방정식 형태
  - $Y = f(X_1, X_2, \cdots, X_p) + \varepsilon$
  - ε은 회귀모델의 오차항. 무작위(설명되지 않는) 오차의 형태
- 데이터가 주어졌을 때, 우리가 관심있는 것을 예측하고 해석하기 위한 기본적인 모델

2) 상관분석과의 차이

- 회귀분석은 기본적으로 변수들 사이의 상관관계에 기반한 모델이다. 그리고 회귀분석 이외의 머신러닝, 딥러닝 모델들도 다 상관관계에만 기반한 모델이다. 이렇게 상관관계에만 기반할거면 '상관분석'만 하면 되지 않을까?
- 상관분석의 한계
  - 상관관계는 두 변수간의 관계밖에 표현할 수 없다.
  - 두 변수의 선형적 상관성의 정도만 표현할 수 있고, 구체적 예측과 설명이 불가능
  - '주희의 음주량은 학점과 상관성이 높다' 만으로는 의미를 갖기 어려움.
    주희가 한 주에 술을 두 번 마신다면, 학점이 어떻게 변화할지를 알 수 있어야 유의미한 정보를 갖춘 것.
- 회귀 모델링의 과정

(1) 문제 정의
- 주희의 학점을 가장 잘 표현할 수 있는 변수들은 무엇이 있을까?

(2) 예상되는 적절한 변수들 선택
- X1~Xp : 통학 거리, 주당 술자리 횟수, 피셋 여부

(3) 데이터 수집 및 전처리
- 주희의 학점, 집주소와 학교 사이의 거리 계산, 술자리 사진으로 횟수 추정

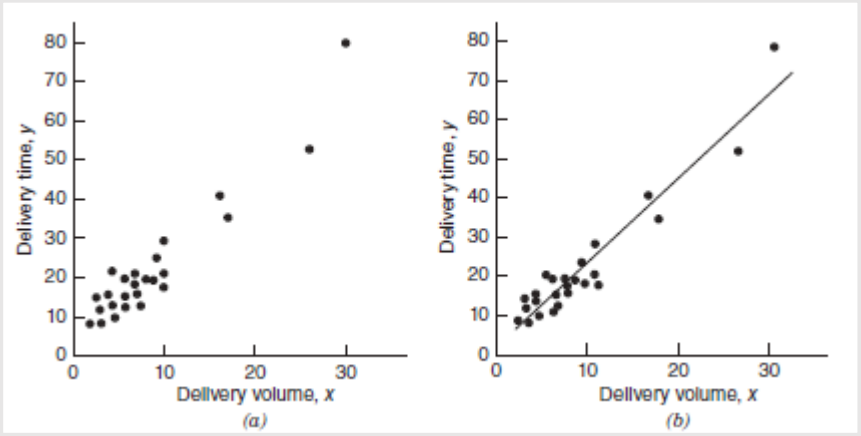
(4) 모형 설정과 적합
- 적절한 회귀분석 모델 선택.
- 선형 vs 비선형, 단순회귀 vs 다중회귀 등 고려
- $\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 X_1 + \cdots + \hat{\beta}_p X_p$

(5) 모형 평가와 해석
- 모형이 회귀 가정을 만족하는가? 만족하지 않으면 수정 (2주차 참고!)
- 매주 2회 술 먹고, 현재 거주지에서 피셋할 때 학점은 **평균적**으로 4.2 정도 나올 것이다

2. 단순선형회귀분석

1) 단순선형회귀식

- 단순 선형 회귀(Simple linear reg)에 대해 알아보자. X와 Y의 관계를 가장 잘 표현할 수 있는 직선을 찾는다. 주어진 데이터를 가장 잘 설명할 수 있는 직선을 찾아 수식화.



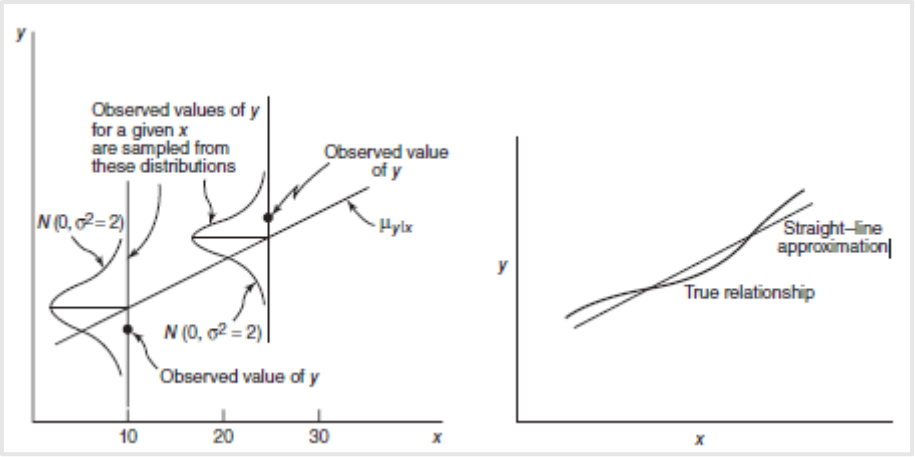
Population regression model : $y = \beta_0 + \beta_1 x + \varepsilon$, 모집단(population)의 관점
Sample regression model : $y_i = \beta_0 + \beta_1 x_i + \varepsilon_i, \quad i = 1, 2, \ldots, n$, 관측치(sample)의 관점

Mean of model : $E(y|x) = \beta_0 + \beta_1 x$
- > 평균적으로 저 직선 주위에서 데이터들이 있다

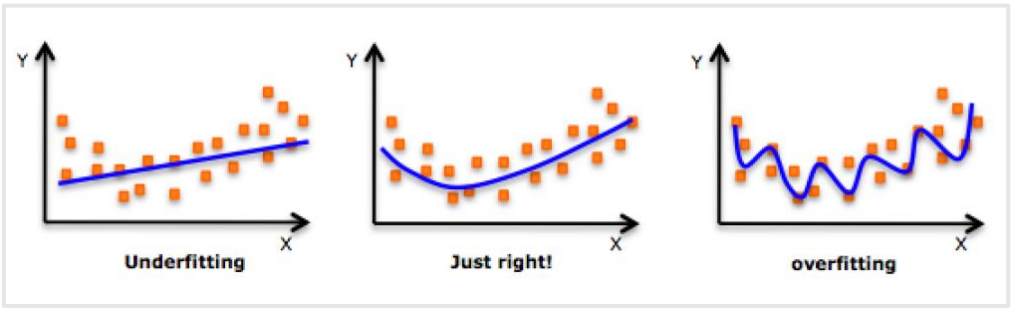
Variance of model : $Var(y|x) = Var(\beta_0 + \beta_1 x + \varepsilon) = \sigma^2$
- > 데이터들이 떨어진 정도는 일정하다.

- y_i : 반응변수 y의 i번째 관측값
- x_i : 예측변수 x의 i번째 관측값
- ε_i : i번째 관측값에 의한 랜덤 오차는 정규분포를 따르는데, 평균은 0이고 분산은 σ^2
  $\varepsilon_i \sim NID(0, \sigma^2)$, NID = Normally and Independently Distributed
- β_0, β_1 : 회귀계수라고 하고, 추정해야 할 모수.
  회귀계수를 잘 추정하는 것은 더 좋은 모델을 위해 필수적! 추정방법은 뒤에서 다뤄보자.

- 단순선형회귀 모델의 해석
  - x가 한 단위 증가할 때, y는 β_1 만큼 증가한다. 매우 간단함!! 마치 중학교 때 배운 일차함수나 다를 바 없다. 물론 딱 β_1만큼 증가하는 것이 아니라, **평균적으로 β_1만큼 증가**한다는 것.
- 왜 직선인가?
  - 변수의 영향력을 간단하게 모형화 할 수 있다. 아까 해석했죠?



- 선형근사를 넘어서 고차근사를 할 경우 모델의 복잡도가 높아진다.
  당장 주어진 데이터에는 잘 설명해도, 나중에는 설명 못함.

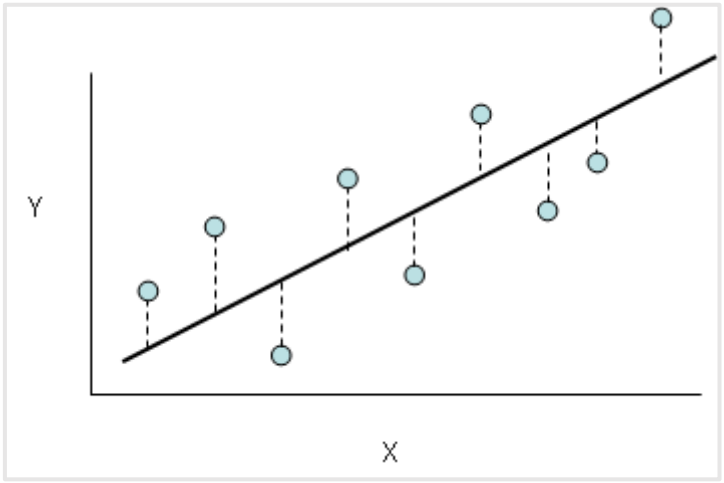


- 물론 현실(real world) 데이터는 선형적으로 생성되지 않는 경우가 많다. 따라서 단순한 선형 회귀식은 예측 성능이 떨어지는 경우도 많다. 하지만 기본적으로 선형회귀식의 원리와 가정, 변형을 통해 우리가 궁극적으로 하고자 하는 예측 모델링에 대한 전체적인 흐름을 이해할 수 있고, 머신러닝 방법들에 대한 총체적인 이해에 도움을 준다.

2) 모수의 추정 – 최소제곱법(Least Square Method)

- 우리가 알고 싶은, 가정하는 회귀식의 형태는 $y_i = \beta_0 + \beta_1 x_i + \varepsilon_i$이다. 하지만 이는 참회귀식(True Model)로 알 수 없다. 다만 실제 데이터가 생성되는 형태를 선형으로 가정했기 때문에, 주어진 데이터를 통해 데이터가 생성되는 형태를 '추정(estimation)'할 수 있을 뿐이다.

  - 추정해야 할 모수는 현재 $\beta_0, \beta_1, \sigma^2$, 하지만 β 추정에만 집중하자! 회귀직선만 잘 만들어내면 σ^2의 추정은 어렵지 않다. 원래 우리 목표는 데이터를 가장 잘 표현하는 직선을 찾는거니까!! 그렇다면 어떤 추정이 좋은 추정일까?



- 직관적으로 우리가 만들어낼 회귀 직선과 관측치 사이의 오차가 작으면 작을수록 좋다. 그렇다면 이 오차를 최소화하면 어떨까?

- 딱 보니까 오차의 합은 0이겠지? 절대적인 떨어짐(deviation)을 최소화하는 방법이 필요하다. 우리는 보통 이럴 때 절대값을 씌우기보다는 제곱한다. 따라서 '오차제곱합'을 최소화하고, 이를 '**최소제곱법(LSE)**'라고 한다.

Least-Squares Estimation of β_0 and β_1

- Use least squares (LS) method: estimate β_0 and β_1 to minimize

$$S(\beta_0, \beta_1) = \sum_{i=1}^{n}(y_i - \beta_0 - \beta_1 x_i)^2 = \sum_{i=1}^{n}\varepsilon_i^2$$

- 이런 제곱꼴의 함수는 볼록Convex(아래로 볼록) 함수다. 고등학교때 어떤 '미분가능한 함수'의 최솟값, 최댓값을 찾을 때, 미분값=0을 통해 찾았다. 현재 형태가 이차함수 $y = x^2$과 같은 형태이므로, 다른 고려 없이 미분값=0을 취하면 된다.

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

$$\begin{aligned}\frac{\partial S}{\partial \beta_0}\Big|_{\hat{\beta}_0,\hat{\beta}_1} &= -2\sum_{i=1}^{n}(y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i) = 0\\ \frac{\partial S}{\partial \beta_1}\Big|_{\hat{\beta}_0,\hat{\beta}_1} &= -2\sum_{i=1}^{n}(y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i)x_i = 0\end{aligned}$$

- 그래서 이 방법으로 β_0, β_1의 추정치를 얻을 수 있고,
이를 '**최소제곱추정치(Least Square Estimator)**'라고 한다.

- The solution to the normal equation is

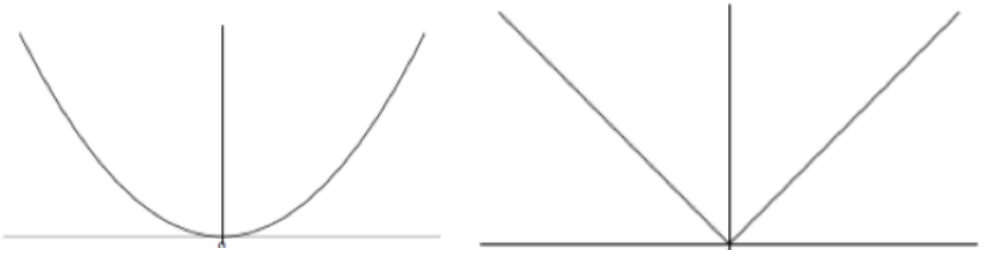
$$\begin{aligned}\hat{\beta}_0 &= \bar{y} - \hat{\beta}_1 \bar{x}\\ \hat{\beta}_1 &= \frac{S_{xy}}{S_{xx}},\end{aligned}$$

where $\bar{y} = \sum_{i=1}^{n} y_i/n$, $\bar{x} = \sum_{i=1}^{n} x_i/n$,

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

- 왜 오차제곱합을 최소화할까?

  - 미분이 편리하다. 미분값=0을 통해 바로 추정할 수 있기 때문에 매우 간단.
  물론 절대값의 합을 최소화하는 방법도 있지만 그것은 뒤에서!



  - 오차가 클수록 더 큰 페널티를 부여할 수 있다. 중심에서 멀어질수록 기울기가 상승하고 있죠?
  인정할 수 있는 정도의 오차의 영향력은 줄이고, 인정할 수 없는 오차의 영향력은 높이는 방법.

- LSE의 가정과 특징

  - LSE 방법은 그냥 사용이 가능하다. 하지만 다음의 조건이 갖춰질 경우 특별하고도 매우 유용한 성질을 지닌다.
    1) 오차들의 평균은 0,
    2) 오차들의 분산은 σ^2으로 동일 (등분산)
    3) 오차간에는 자기상관이 없다(uncorrelated)

  - 이 조건들을 만족하면 최소제곱추정량은 **BLUE**가 된다. Best(분산이 제일 작은) Linear(선형) Unbiased Estimator(불편추정량)으로, 다른 선형불편추정량보다 분산이 늘 작다!
    BLUE만을 생각했을 때는 정규성은 필요하지 않다.

  - 심화적으로 들어가면 통계적 추정이론에서는 가능하면 불편추정량을 구하고, 이때 추정량의 분산을 최소화하려 한다. 이를 '최소제곱불편추정량(MVUE)'라고 하는데, 그것과 살짝은 다르지만 거의 같다.
    어떻게 같은지는 잠시 뒤에 설명.

  - 아무튼 BLUE의 장점이 무엇인지는 알겠죠?
    저 3가지 조건만 만족하면 우리 추정량은 정말 좋은 추정량이라는 것!

- 최대가능도(ML) 추정과 최소제곱법의 차이

  - 최대가능도 추정은 확률적인 방법에 근거해서, 어떤 모수가 주어졌을 때, 우리 데이터가 나올 '**가능도**'를 최대로 하는 모수를 선택하는 방법이다.

  - 그런데 이런 ML방법은 언제나 분포 가정이 들어간다.
    회귀분석에서는 오차 $\varepsilon_i \sim N(0, \sigma^2)$ 이라는 가정이 들어가면, ML방법 적용 가능!

  - 그래서 MLE와 LSE가 다르냐? 정규분포 가정이 있다면 **완전히 동일한 추정량**을 산출한다.
    그냥 구하려는 아이디어만 다를 뿐, 수식마저 같은 형태다.

  - 심화적으로 접근하면, 통계적 추정이론에서는 MLE를 먼저 정의하고 MVUE를 정의한다. 우리가 알고 있는 간단한 분포에서는 MLE의 불편성만 만족시킬 경우 MVUE와 동일해지는 경우가 많다. 이런 점을 고려했을 때, LSE는 MVUE와 비슷한 방식으로 이해할 수 있지 않을까?

  - 실제로 최소제곱추정량이 MVUE가 되는 경우가 있다. 원래 3가지 가정만 있다면 LSE는 BLUE였다. 추가적으로 4) 오차항이 정규분포를 따르고, 5) X가 full rank라면, $X^t y$가 function of CSS(Complete Sufficient Statistics)이기 때문에, beta가 불편추정량이므로 Lemahn-Scheffe 정리에 의해 MVUE까지 성립하게 된다. 별로 중요하지 않지만 궁금해할거 같아서…

3) 적합성(Goodness of fit) 검정

- 잔차(Residual)

  우리는 지금 회귀계수 beta를 추정했다. 그렇다면 이 추정된 회귀계수를 바탕으로 회귀식을 만들 수 있고, 이 회귀식이 얼마나 우리 데이터를 잘 설명하는지가 궁금하다. 지금까지 오차에 대해 이야기했지만, 오차 ε_i도 β처럼 참회귀식에서 만들어지는 오차다. 우리 데이터들의 공간에서는 이 오차가 실현되는게 아니기 때문에, '오차의 추정량'으로서의 잔차를 정의한다. 결과적으로 오차와 잔차는 다를 바 없지만, 모집단과 표본의 차이라고 생각하면 된다.

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

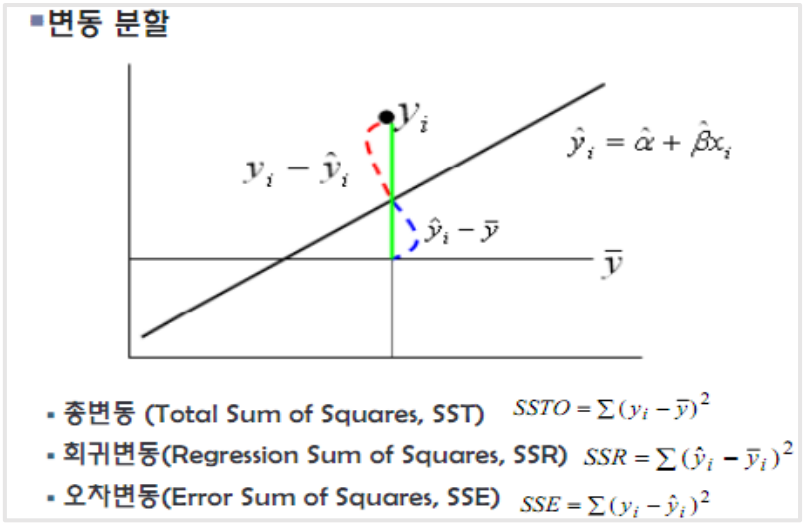
- 그렇다면 이 잔차를 통해 우리 회귀식이 얼마나 적절한지 알아보자!

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

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

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

- SST = SSR + SSE, 증명은 생략! 대신 그림으로 이해!



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

- 총 변동에서 회귀식이 설명하는 부분, 1에 가까울수록 좋다!
  즉, Y가 X에 의해 설명되는 비율이라고도 이해할 수 있다.

4) 유의성 검정

- 우리는 오차항이 평균은 0, 분산은 σ^2인 정규분포를 따른다고 가정했다. $\varepsilon_i \sim N(0, \sigma^2)$ 라는 **정규분포 가정**하에서 개별 베타 계수에 대한 통계적 검정을 할 수 있다.

  - 귀무가설은 $\beta = 0$이다. 밑에 사진의 $b = \hat{\beta}$ 으로 생각하고 보면 된다.

$$b_1 \sim N\left(\beta_1, \frac{\sigma^2}{\sum_{i=1}^{n}(X_i - \bar{X})^2}\right) \qquad b_0 \sim N(\beta_0, var(b_0))$$

$$\frac{b_1 - \beta_1}{s(b_1)} \sim t_{(n-2)}. \qquad \frac{b_0 - \beta_0}{s(b_0)} \sim t_{(n-2)}.$$

Hypothesis Tests
Two sided: $H_0 : \beta_1 = \beta_{tst}$ vs. $H_0 : \beta_1 \neq \beta_{tst}$.
Let

$$T = \frac{b_1 - \beta_{tst}}{s(b_1)}.$$

If $|T| > t_{(1-\alpha/2;n-2)}$, then reject H_0 at α level.
One sided: $H_0 : \beta_1 = \beta_{tst}$ vs. $H_a : \beta_1 > \beta_{tst}$.
If $T > t_{(1-\alpha;n-2)}$, then reject H_0 at α level.

  - β_0 에 대한 검정도 동일한 방법으로 진행하면 된다.

  - 귀무가설을 기각하지 못하면, 개별 회귀계수는 0이라는 것이고, X와 Y 사이에 아무 의미가 없다는게 아니다!
    단지 선형적 관계가 없다고 할 수 있을 뿐이다.

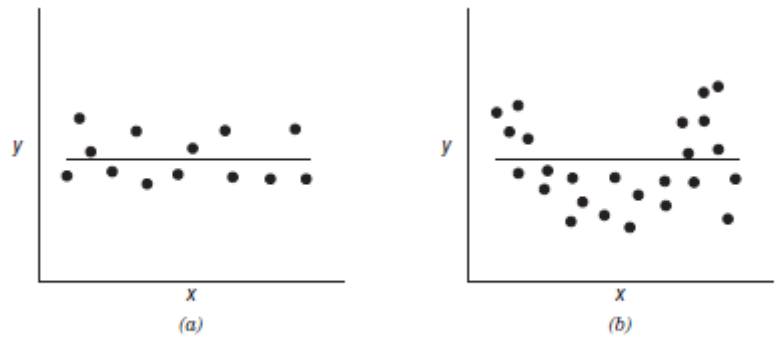


Figure 2.2 Situations where the hypothesis H_0: $\beta_1 = 0$ is not rejected.

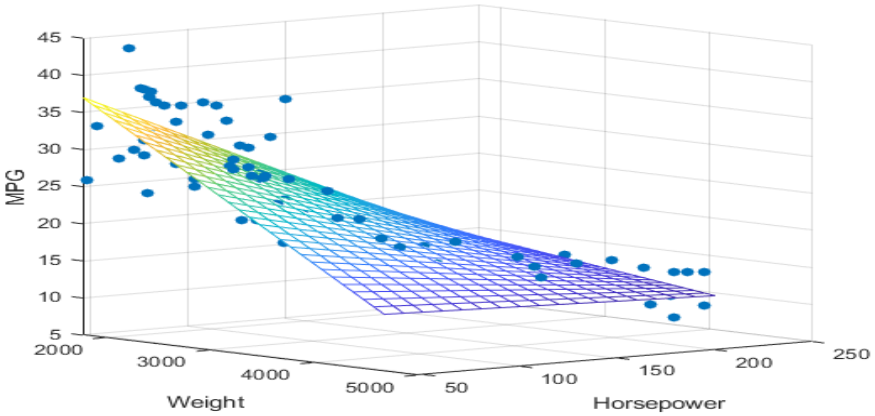
3. 다중회귀분석

1) 다중회귀분석이란?

- 단순화귀분석은 하나의 설명변수 X와 반응변수 Y 사이의 회귀선을 찾는 것이었다면, 다중회귀분석은 설명변수 X개 한 개가 아닌 여러 개인 상황이다.

$$y = \beta_0 + \beta_1 x_1 + \cdots + \beta_p x_p + \varepsilon$$

  - 변수의 개수는 p개고, 회귀계수는 p+1개
  - 설명변수 X가 두 개라면 평면을 찾고, 세 개라면 공간을 찾을 것이다.



- 회귀계수 추정
  - 단순선형회귀에서는 각각 beta0와 beta1에 대해 편미분을 통해 회귀계수를 추정했다. 하지만 현재 추정해야 할 beta의 개수가 p+1개이기 때문에, 일일이 편미분하는 것은 매우 불편한다. 따라서 우리는 행렬을 도입하기로 한다.
  - $y = X\beta + \varepsilon$ 로 행렬을 통해 우리의 회귀식을 표현할 수 있다. 이때도 동일하게 최소제곱법, 최대가능도추정법을 사용할 수 있고, 단순회귀의 경우와 동일하게 정규분포 가정하에서 둘은 **동일한 결과**를 만들어낸다. 최소제곱법의 목적함수를 보고 직접 유도해보자.

$$\begin{aligned} S(\boldsymbol{\beta}) &= \sum_{i=1}^{n} \varepsilon_i^2 \\ &= \boldsymbol{\varepsilon}'\boldsymbol{\varepsilon} \\ &= (\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta})'(\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta}) \end{aligned}$$

- The least-squares estimators of $\boldsymbol{\beta}$ is

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

provided that $(\boldsymbol{X}'\boldsymbol{X})^{-1}$ exists.

  - 별로 어렵지 않으므로 해당 내용은 간단히 증명해볼게요!

- 해석 (매우 중요!)

  - 단순회귀와는 다르다. 단순회귀에서 회귀계수에 대한 해석은 매우 단순했다. 하지만 다중회귀에서 β_j에 대한 해석은 x_j를 제외한 **나머지 X 변수들을 고정시킨 상태에서 x_j가 한 단위 증가할 때 y가 증가하는 양**을 말한다. 즉, 이미 다른 변수들이 설명하는 부분 이외에 나머지를 x_j가 설명하는 것이다.

  - 다른 변수들이 고정되어 있다면 이는 곧 다른 변수들이 상수처럼 취급된다는 것! 다른 변수들을 상수화 하고 이 상태에서 우리가 관심있는 변수의 기울기를 확인하는 형태이다.

  - 이렇게 다른 변수들을 고정시킨 상태에서 x_j의 증분을 정의하는 것이 가설검정의 설계를 보게 되면 매우 타당해진다. 조금 어려울수는 있다…

2) 유의성 검정 – F-test와 t-test

- F-test (수식 좀 많고 어렵지만 가보자!)

  - 보통 회귀분석 문제에 있어서 F-test는 '전체 베타 계수는 0이다'를 귀무가설로 하는 모델 전체에 대한 검정을 실행한다. 만약 $\beta_1 = \beta_2 = \cdots = \beta_p = 0$ 이라는 귀무가설을 기각하지 못하면, 우리 모델링이 무의미함을 나타낸다. R에서 Summary를 통해 확인하는 F-statistic에 대한 출력이 이 내용이다.

$$H_0 : \beta_1 = \beta_2 = \cdots = \beta_p = 0\,, \qquad H_1 : at\ least\ one\ \beta_j \neq 0$$

```
> 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
```

  - 하지만 좀 더 일반화된 검정을 다루려 한다. 이를 'Partial F-test'라고 한다. 원래의 모든 변수를 사용한 다중회귀모형을 완전모형(Full Model : FM), 일부 회귀계수를 특정한 값(보통 0)으로 두는 축소모형(Reduced Model : RM)이라고 하자. 이때 FM이 좋은지, RM이 좋은지 검정하고 싶다.

$$RM: y = \beta_0 + \beta_1 + \cdots + \beta_{j-1} + \beta_{p-j} + \beta_{p-j+1} + \cdots + \beta_p$$
$$FM: y = \beta_0 + \beta_1 + \cdots \quad \ldots + \beta_p$$

$$H_0: RM이\ 맞다.\ \beta_j = \beta_{j+1} = \cdots = \beta_{p-j} = 0, \qquad H_1: FM이\ 맞다.\ \ at\ least\ one\ \beta \neq 0$$

$F = \frac{SSR(FM)-SSR(RM)/(p-q)}{SSE(FM)/(n-p-1)} \sim F_{p-q,n-p-1}$ 혹은 $F = \frac{SSE(RM)-SSE(FM)/(p-q)}{SSE(FM)/(n-p-1)} \sim F_{p-q,n-p-1}$의 F-test를 진행하면 된다.

$F \geq F_{p-q,n-p-1;\alpha}$ 이면 귀무가설을 기각한다. -> FM이 더 적절하다.

- 이런 정의가 갖는 의미는 무엇일까? 분자의 값이 분모에 비해 많이 크다면 귀무가설을 기각할 수 있고, FM이 더 적절해진다. 분자의 값을 키우려면 SSR(FM)의 값이 SSR(RM)보다 많이 커야한다. 즉 더 많은 변수를 넣는만큼, 회귀식의 설명력이 충분히 커져야 의미가 있다는 것! 원래 SSR(FM) > SSR(RM) 이지만, SSR(FM) >>>> SSR(RM) 이어야 분자가 충분히 커진다. 변수를 넣었는데 설명력이 거의 상승하지 않는다는 것은 중요도가 떨어지는 변수라는 것이겠지?

  또한 다시 확인하자면 이런 설계는 RM에서 살아있는 베타계수를 기반으로 하는 검정이다. 즉 다른 변수들이 이미 존재하는데, FM에 들어가는 변수들을 추가할 때 회귀식의 유의미한 설명력 증가가 있는지를 확인하는 것이다.

- 이를 기본으로 F-test를 정의하게 될 경우, 조금 더 다양한 방식의 회귀계수에 대한 검정이 가능해진다. 몇 개의 베타계수를 0이라고 가정한 상태에서 검정하는 것도 가능하고, 혹은 몇 개의 베타계수를 1, 2와 같은 상수로 고정하는 검정도 가능해진다. 이에 대한 검정은 행렬을 통해 정의하는 것이 매우 편리한데, 어려워지니 생략. 관심있으면 몽고메리 책 참고!

- Partial F-test는 좀더 일반화된 검정이나, 사실 우리들이 사용할 일은 거의 없다.
  일반적으로 회귀식 전체에 대한 검정을 중심으로 이해하자.

$$H_0: RM이\ 맞다, \quad y = \beta_0 + \varepsilon, \quad VS \quad H_1: FM이\ 맞다, \quad y = \beta_0 + \beta_1 x_1 + \cdots + \beta_p x_p + \varepsilon$$

$F = \frac{SSE(RM)-SSE(FM)/(p-q)}{SSE(FM)/(n-p-1)}$ 이지만 RM의 SSE가 회귀식 SST고 q=0이기 때문에,

$F = \frac{SST-SSE/p}{SSE/(n-p-1)} = \frac{SSR/p}{SSE/(n-p-1)} = \frac{MSR}{MSE}$ 라는 간단한 공식이 만들어진다. 이거만 기억하자!!!!!

- t-test : 개별 회귀계수의 유의성 검정

  - F-test의 가설을 다시 보자.

$$H_0: RM이\ 맞다, \quad y = \beta_0 + \varepsilon, \quad VS \quad H_1: FM이\ 맞다, \quad y = \beta_0 + \beta_1 x_1 + \cdots + \beta_p x_p + \varepsilon$$

  - 이런 상황에서 귀무가설을 기각한다고 해도, 개별 회귀계수 중에 유의하지 않은 회귀계수가 존재할 수 있다. 따라서 어떤 회귀계수가 유의한지를 알고 싶다.

  - $H_0: \beta_j = 0, \quad vs \quad H_1: \beta_j \neq 0$
    H0 : 다른 변수들이 다 적합된 상태에서 x_j는 통계적으로 유의하지 않다.
    H1 : 다른 변수들이 다 적합된 상태에서 x_j는 통계적으로 유의하다.

  - $t_j = \frac{\hat{\beta}_j}{s.e.(\hat{\beta}_j)}$이고, $|t_j| \geq t_{n-p-1;\,\alpha/2}$ 이면 H0을 기각한다. 즉, x_j는 다른 변수들이 고정된 상태에서, 통계적으로 유의하다(0이 아니다).

  - 계속 '다른 변수들을 고정한 상태에서'를 강조하는 이유가 뭘까? t-test는 해당 변수 자체가 유의미한 지를 확인하는 것이 아니다. 다른 변수들이 다 적합된 상태에서, x_j를 추가적으로 적합했을 때 이게 유의미한 회귀식의 설명력 증가를 가져오는지를 확인하는 것이다. 표현이 어디서 들어본 것 같은데?

  - 회귀분석에서 개별 변수에 대한 검정은 한 변수에 대한 Partial F-test와 완전히 동일하다.

들어본 사람은 알텐데, t분포를 제곱하면 F분포와 같은 형태다. 아무튼 같음!!!

- 그래서 하고 싶은 말을 말하자면, **t-test로 변수 선택하는 것은 매우 위험**하다. 다른 변수들이 이미 고정되어 있는 상황에서 변수의 유의성을 판단하는 것이기 때문에, 다른 회귀식을 가정했을 때는 해당 변수가 유의할 수도 있다. 어떤 결과가 나올지 모르므로, t-test로 변수를 선택해선 안되고, 3주차에 다룰 변수선택법으로 선택하는 것이 더 나은 방법이다.

```
> 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
```

- 아까 봤던 예시를 다시 보자. 여기서 F-test 값을 먼저 확인하면, 회귀식에 대한 귀무가설을 기각할 수 있다. 이후 t-test를 확인한다. TV와 Radio는 해당 회귀계수가 0이라는 귀무가설을 기각하고, Newspaper는 기각하지 못한다. 이를 적절하게 해석해보면,

  '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 vs t

  - 한 변수에 대한 Partial F-test와 해당 변수에 대한 t-test의 결과는 동일하다. 그렇다면 회귀식 전체에 대한 F값과 개별 변수에 대한 t값 중에 무엇을 먼저 보아야 할까?

  - F를 먼저 보는 것이 적절하다. 전체 회귀식에 대한 검정이 더 rigorous한 검정이기 때문에, F를 기각하지 못하면 t-test를 보는 것은 의미가 없다. 하지만 대부분의 경우 F를 기각하지 못하는 경우는 없다.

  - 또한 F를 먼저 보라고 하는 이유는, F를 기각하지 못해도 t는 기각하는 경우가 있을 수 있기 때문이다. 개별 변수들에 대한 동시검정을 진행하다 보면 유의수준이 α로 고정되지 않고 넘어선다. 그에 따라 가설을 기각하기 더 쉬워진다.

3) 적합성(Goodness of fit) 검정

- $R_a{}^2$: Adjusted R square(수정결정계수)

  - $R^2 = \frac{SSR}{SST} = 1 - \frac{SSE}{SST}$ 을 정의했었다. 총 변동에서 회귀식이 설명하는 변동의 비율이었다.

- 근데 이 지표의 단점은, **변수가 늘면 자연적으로 증가**한다. 진짜 쓸데없는 변수가 들어가도 의미없이 증가해버린다. 왜냐하면 총 변동은 고정되어 있는데, x 변수가 추가되면 회귀식으로 설명되는 변동이 아주 조금이라도 증가하겠지? 그러면 R^2 값은 증가해버린다. 그런데 이런 무의미한 변수추가는 해석도 어렵게 하고, 예측에도 안좋은 영향을 끼칠 가능성이 높다. 따라서 이런 무의미한 변수추가로 인한 R^2의 상승을 방지하고자 새로운 지표를 만든다.

- $R_a{}^2 = \frac{SSR/p}{SST/(n-1)} = 1 - \frac{SSE/(n-p-1)}{SST/(n-1)}$ 이렇게 새로운 지표 수정결정계수를 만든다. 수정결정계수는 변수의 개수가 다른 두 회귀식을 비교할 때 드러난다. 변수의 개수가 다른 회귀식의 결정계수를 비교하면, 변수가 많은 쪽의 결정계수가 회귀식이 더 유의미한지 여부와 관련없이 더 높을 수 있다. 이를 고려해 수정결정계수를 사용하는 것이고, 변수 개수의 다름을 보정하는 지표다.

- 따라서 변수의 개수가 다른 경우 $R_a{}^2$를 사용하여 비교하고, $R_a{}^2$가 높은 회귀식이 더 좋은 회귀식이다. 다만, 이때 $R_a{}^2$는 R^2처럼 전체 변동중에 회귀식이 설명하는 변동으로는 해석할 수 없다.

4. 데이터 진단 - 이상치, 지렛값, 영향점

데이터 중에 일반적인 경향에서 벗어나는 점들이 있을 것이다. 이러한 점들은 최소제곱 회귀모형을 크게 바꾸거나, 그에 따라 성능을 저하시키기도 한다. 일부 점의 영향력이 너무 큰 상황을 어떻게 다루면 될까?

1) 잔차

- $e = y - \hat{y} = y - X\hat{\beta} = y - X(X^tX)^{-1}X^ty = y - Hy = (I-H)y$

  $Var(e) = Var\big((I-H)y\big) = (I-H)\sigma^2(I-H)^t = \sigma^2(I-H)(I-H)^t = \sigma^2(I-H)$

  $Var(e_i) = (I - h_{ii})\sigma^2$

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

  - 그냥 잔차는 y-yhat의 값. y값의 단위에 따라 잔차가 날뛸 수 있으니, 좀더 일반화된 상황에서 적용할 수 있도록 표준화해줘야 한다.

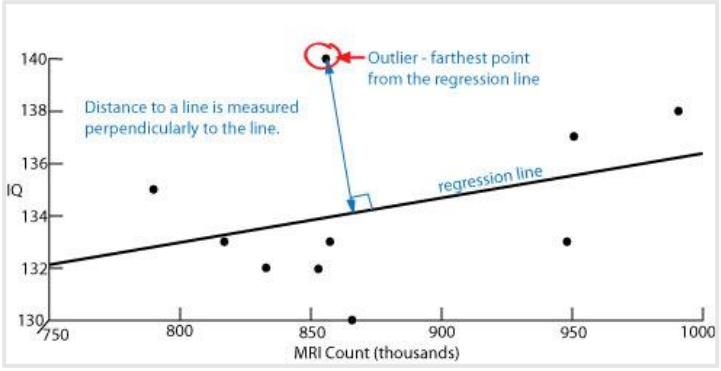
  $$z_i = \frac{e_i}{\sigma\sqrt{1-h_{ii}}}$$

  - 근데 이때 시그마는 모수이므로 알 수 없기 때문에, 시그마의 추정량을 넣어준다. 이를 표준화잔차(Studentized Residual)라고 한다.

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

2) 이상치(Outlier)

- 이상치는 표준화 잔차가 매우 큰 값을 의미한다. y의 기준에서 절대값이 큰 값!



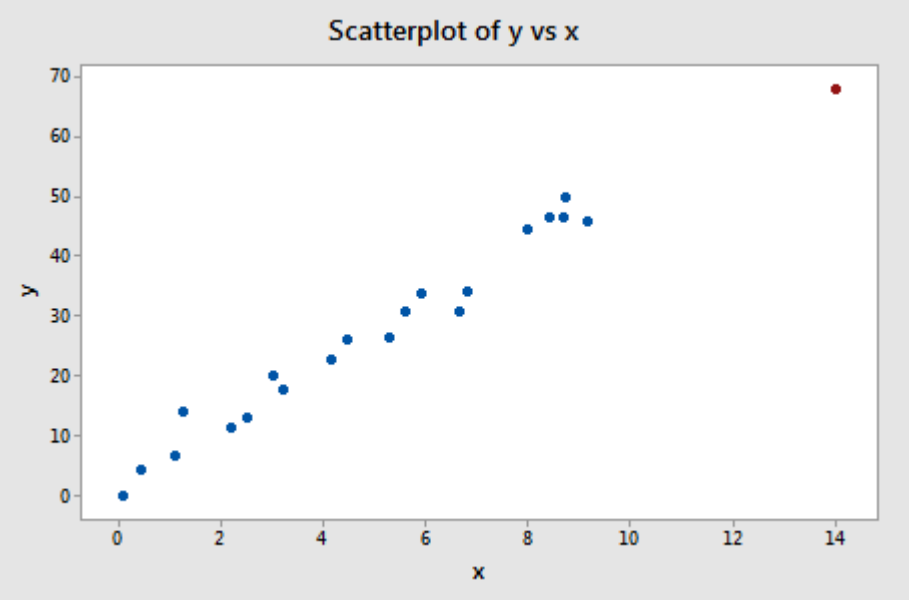
- 보통 $|r_i| > 3$ 이면 이상치라고 판단한다.

3) 지렛값(Leverage point)

- Outlier가 y의 관점이었다면, 지렛값(leverage point)는 x의 관점이다.
  - $H = X(X^tX)^{-1}X^t, \quad h_{ii} = {x_i}^t(X^tX)^{-1}x_i$ 라는 것을 아까 확인했다. h_{ii}를 이렇게 표현 가능하다.

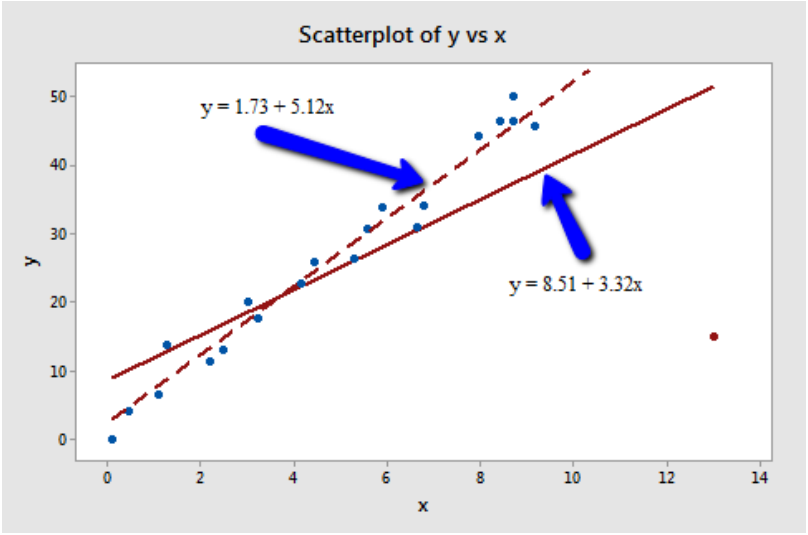
$$h_{ii} = \frac{1}{n} + \frac{(x_i - \bar{x})^2}{\sum(x_i - \bar{x})^2}$$

  - 저 식을 보자. x_i값이 $\bar{x}$에서 멀리 떨어져 있을수록 h_{ii}가 커진다고 볼 수 있다. 즉, x평균에서 멀수록 레버리지 값이 상승한다.
  - $h_{ii} > \frac{2(p+1)}{n}$ 이면 지렛값으로 판단한다.



4) 영향점 (Influential point)

- 영향점은 한 관측치가 회귀직선 기울기에 상당한 영향을 주는 점을 말한다. 기존의 Outlier와 Leverage Point의 진단만으로 회귀직선이 변한다고 말하기 어렵다. Outlier일지라도 x 평균 주위에 위치할 경우 기울기를 변화시키지 못하고, Leverage일지라도 회귀직선의 연장선에 있을 수 있다. 따라서 **Outlier와 Leverage를 동시에 고려**하는 지표가 필요하고, 이를 Cook's Distance라고 한다.

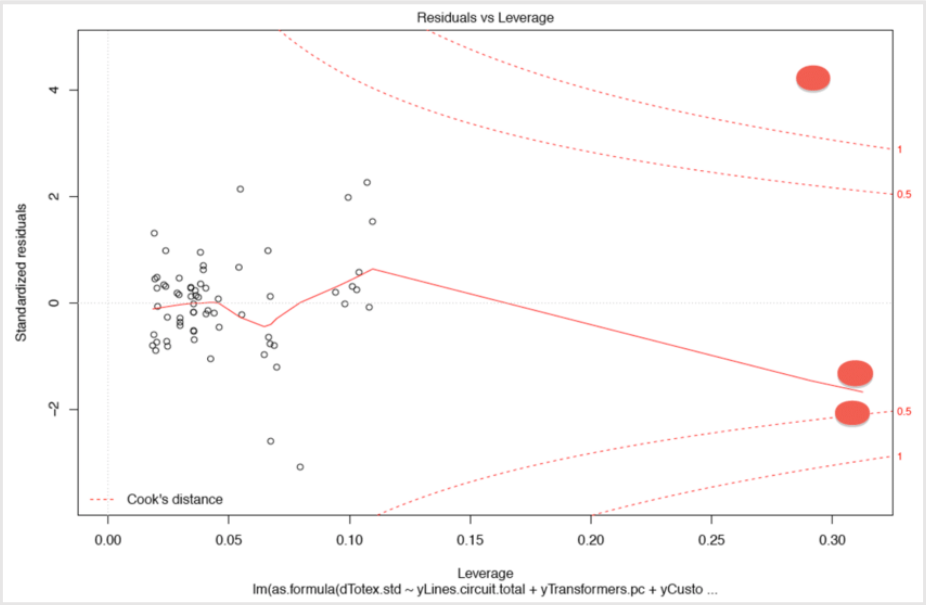


- Cook's Distance

- 영향점을 확인하는 표준적인 지표. 특정 데이터를 지웠을 때, 회귀선이 변하는 정도

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

- 식에서 확인할 수 있듯, outlier와 leverage를 동시에 고려하는 지표다. 각각이 커지면 커질수록 C_i가 커진다. 보통 C_i>1 이면 영향점으로 간주한다.



- R에서 회귀식을 적합하고, 적합식에 대해 plot을 그리면 쉽게 cook's distance를 확인할 수 있다. 그림에서 볼 수 있듯, leverage라고 언제나 영향점은 아니다. 다만 leverage 중에 cook's distance가 1보다 큰 점이 하나 존재한다.

- 영향점의 처리

  - 보통 이런 영향점은 제거할 수 있다. 추정을 불안하게 만들면 당연히 잘못된 해석이 나올 수 있고, 예측의 성능도 떨어질 수 있기 때문에 종종 제거하곤 한다.

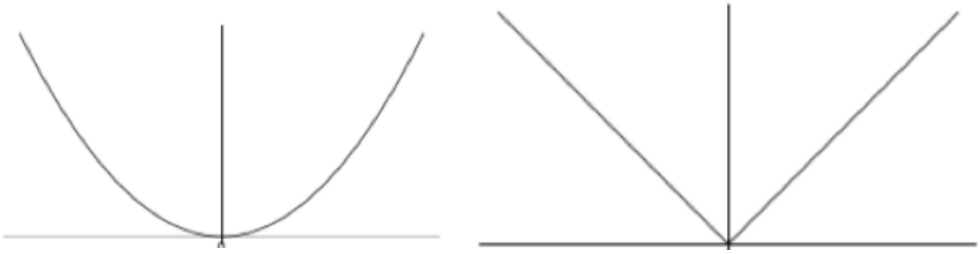
  - 하지만 **데이터를 삭제한다는 것은 늘 조심**해야하는 일이다. 그런 이상치에도 만약 이유가 있다면? 의미를 가지는 이상치라면? 이를 고려하는 **이상치에 강건한(robust)** 모델링이 필요하다.

5. 로버스트 회귀

로버스트 회귀는 이상치의 영향을 줄이는 회귀분석 방법이다. 이상치의 영향을 줄이는 방법은 매우 다양하고, 그에 따라 다양한 모델들이 존재한다. 이중에서 직관적인 이해가 가능한 두 가지 모델(Median Regression, Huber's M-estimation)을 확인하려 한다.

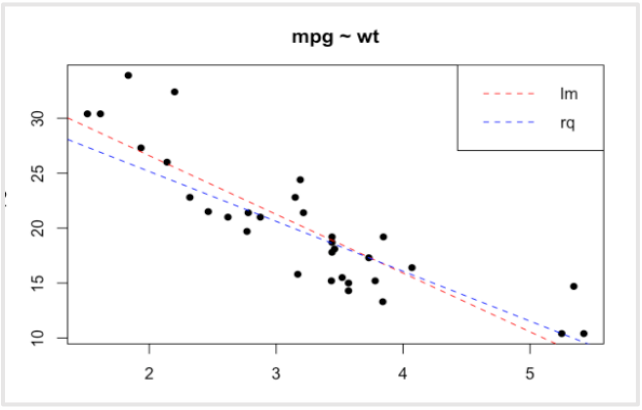
1) Median Regression

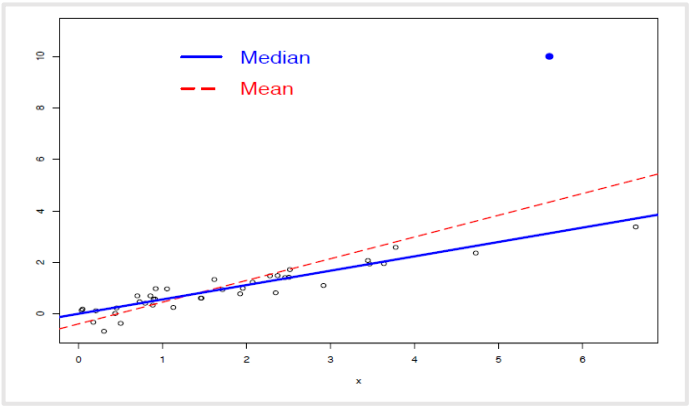
- 최소제곱회귀는 $\sum {\varepsilon_i}^2 = (y - X\beta)^t(y - X\beta)$ 를 최소화하는 beta를 찾는 방법이다. 아까 말했듯 제곱을 최소화하기 때문에 미분이 편리하지만, 이상치에 대해 너무 큰 가중치를 주는 경향이 있다. 그렇다면 어떤 경우에도 동일한 가중치를 주는 것은 어떨까?



- Median Regression은 다음과 같은 식을 최소화한다. $\sum|\varepsilon_i|$
  우리가 평균, 중앙값, 최빈값에 대해 배울 때, 중앙값은 이상치의 영향을 덜 받는다고 배웠죠?
  그런 아이디어를 가져왔다고 생각하면 됩니다.

  - 최소제곱회귀는 x에 따른 평균적인 y를 반환한다면, median regression은 **x에 따른 y의 중앙값**을 반환한다.

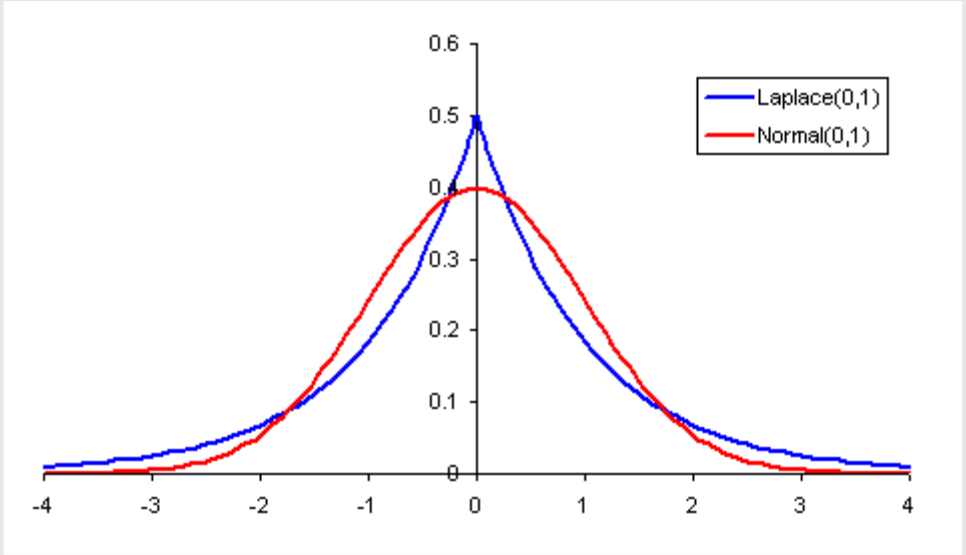




  - 분포가정, 등분산 가정이 없는 모델이다.

  - 하지만 이를 분포로서 이해할 수도 있는데, error가 정규분포가 아닌 double expoenetial(laplace)분포를 따른다고 볼 수 있다. 라플라스 분포는 정규분포보다 긴 꼬리분포를 지니는데, 정규분포의 관점에서 이상치가 많은 형태이다. 따라서 라플라스 오차를 따른다고 보고 ML방법으로 접근할 수 있다.

$$f(x) = \frac{1}{2b}\exp\left(-\frac{|x-u|}{b}\right)$$



  - 다만 해를 찾는 과정이 매우 어렵다. 함수 자체는 convex 함수지만, 0에서 미분 불가능하니 보통의 방법으로는 찾을 수 없다. 비선형최적화 방법들이 필요한데 당연히 생략!!

  - R에서는 'quantreg' 패키지의 rq 함수를 사용한다.

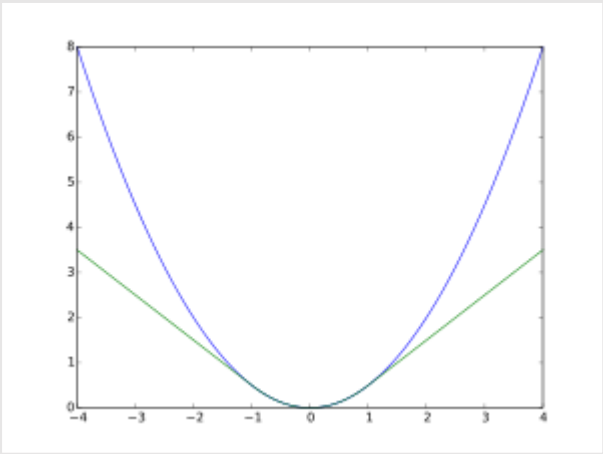
2) Huber's M-estimation

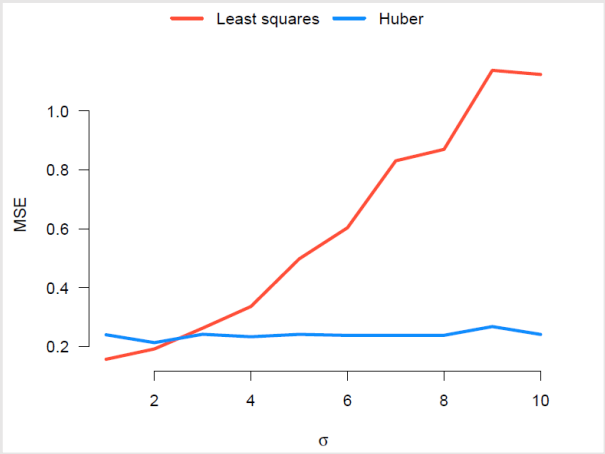
- 최소제곱회귀는 이상치에 지나치게 큰 페널티를 부여하지만, 동시에 적정수준 안에서는 페널티를 완화시켜준다. 그렇다면 적정수준의 페널티는 완화시켜주는 형태는 유지하되, **이상치에 대한 지나친 페널티 부여를 없애는 식**으로 가는 것은 어떨까?

- Huber's M-estimation의 아이디어는 잔차가 특정 상수값보다 크면 잔차의 '제곱'이 아닌 1차식으로 바꿔서 이상치에 강건한 회귀계수를 추정하는 것이다.

$$if\ |e| \le c, \rho(e) = \frac{1}{2}e^2, \quad otherwise\ \rho(e) = c|e| - \frac{1}{2}c^2$$

- 이때의 목적함수(최적화할 함수)는 $\sum \rho(e)$이고, 예를 들어 LSE의 경우는 $\rho(e) = \frac{1}{2}e^2$

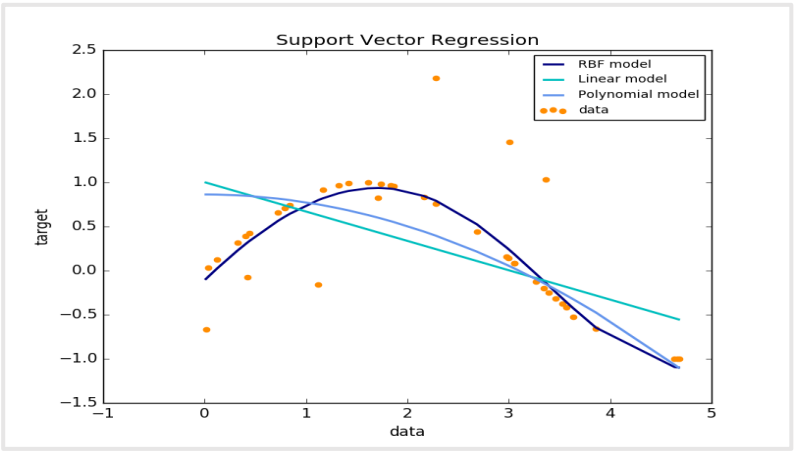




- R에서는 MASS 패키지의 rlm함수를 사용한다.

3) 다른 로버스트 방법

- Least Trimmed Square : 잔차가 큰 관측치를 제외하고 추정한다.
- **SVR(Support Vector Regression)** : 서포트벡터머신과 비슷한 알고리즘인데, SVM은 margin을 최대화한다면, SVR은 margin 안에 들어오는 점들의 개수를 최대화한다고 가슴으로 이해합니다! SVM을 가슴으로+수식적으로 이해한 상태에서 접근해야 그 차이를 알 수 있습니다… 한글자료도 많지 않고 코드만 있는 수준이니 이해하는 것이 좀 어려울 겁니다! 저도 잘 몰라요…ㅎㅎㅎ 다만 SVR의 장점은 robust하며 비선형적인 모델링이 가능하다는 점! 대신 추정하는 속도가 꽤나 느릴겁니다. 나중에 SVR이라는 fancy한 모델을 떠올려서 해볼 수도 있지 않을까요?



4) 더 좋은 추정은 많은 비용을 요구한다.

- 다 알려줘 놓고 왜 이렇게 말할까? 우리는 데이터에 맞게 다양한 모델들을 사용한다. Median regression이나 Huber's M 모두 이상치에 강건하지만 보면 알 수 있듯 목적함수를 최적화하는 것이 간단하지만은 않다. 미분이 불가능해지기도 하고, 아니더라도 느리다. 물론 더 좋은 성능을 발휘하는 경우도 있지만, 정규분포를 가정하지 않기 때문에 일반적인 해석이 매우 어렵고, 모수를 추정하는 데에도 추가적인 시간이 걸린다. 단순히 이상치에 강건해서 성능을 높이려는 이유 때문에 쓰기에는 포기해야 하는 것들이 많아진다.
- 상황에 따라서는 이상치에 영향을 받더라도 LSE를 사용하는 것이 나을 수도 있다. LSE가 제공해주는 BLUE 추정량, 정규분포 가정에 따른 베타 값의 Confidence Interval과 예측 값의 Prediction Interval의 제공 등등 되게 좋은 성질들이 많다. 더 좋은 추정(여기서는 robustness)를 위해서는 그만큼 포기해야 하는 것들이 생긴다. 그 중간에서 우리는 상황에 맞게 판단하고, 우리 분석의 분명한 리즈닝을 말할 수 있어야 한다.