Ch01 Regression

1. 회귀분석의 고전적 테크닉

(1) 선형회귀분석(Linear Regression)

선형 회귀분석의 기본 가정
- 두 변수 사이에 선형적 상관관계가 존재한다.

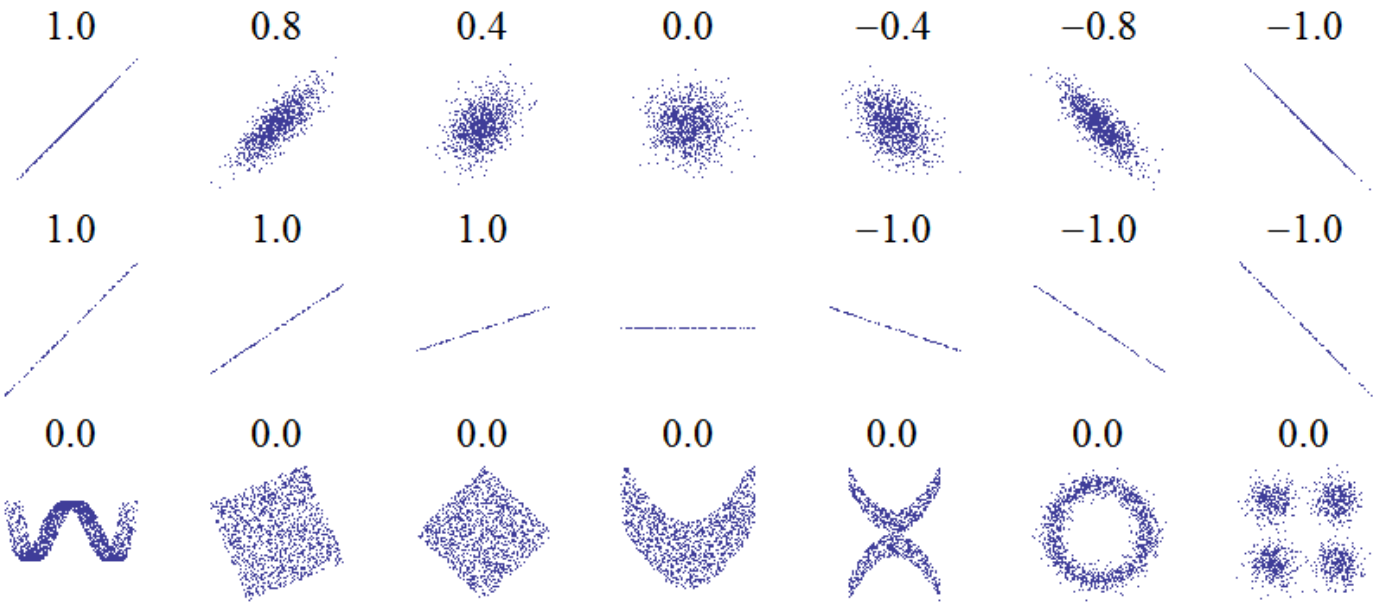
1) 공분산(Co-Variance)

$$Cov(X,Y) = \frac{\sum(x-\bar{X})(y-\bar{Y})}{n-1}$$

2) 피어슨상관계수

$$r(0 \le r \le 1) = \frac{Cov(X,Y)}{\sigma_X \sigma_Y}$$

r 값에 따른 두 변수의 분포



<이미지 출처 : http://cfile10.uf.tistory.com/image/20229C114BE007B6212B84>

피어슨 상관계수는 두 변수간 "선형적"상관정도를 수치화 한 것이다. r이 1에 가까울수록 강한 선형상관관계를, 0에 가까울수록 약한 선형상관관계를 뜻한다. 아무래도 회귀분석과 항상 같이 나오다보니 착각하는 부분이 존재하는데 바로 피어슨계수(r) 회귀 직선의 기울기와 같은 것으로 착각한다는 것이다.

공식과 함께 생각해보자. 피어슨 상관계수 공식에서 X와 Y의 표준편차의 곱이 분모에 위치해있는 것을 발견할 수 있다. r 값과 $\sigma_X \sigma_Y$ 반비례관계로 $\sigma_X \sigma_Y$ 이 작을수록 r이 증가할 것이다. 이는 두 변수간 분산이 감소할 수록 큰 r의 값을 갖는다는 의미이다. 즉 관계의 유무는 공분산이 결정하게 되는 것이지 피어슨 상관계수가 r 관계의 여부를 결정하는 것은 아니며 피어슨 상관계수는 두 변수의 관계에서 얼마나 몰려있는지를 강도로 측정할 뿐인것이다. 따라서 기울기가 변하더라도 몰려있다면 피어슨상관계수(PCC)가 1이며, 직선에 몰려있지만 공분산이 0일경우 계수는 0의 값을 갖는 것이다.

기타 상관관계 추정법
* 스피어만 : 순서를 통한 상관관계 측정법

3) 선형 회귀모델

$$\boldsymbol{y} = w_1 \boldsymbol{x} + w_0$$

상관성 유무보다 사실 중요한것은 두 변수간의 구체적 관계이다. 즉 X가 변할때 얼마만큼 Y가 변하지 설명할 수 있어야한다. 이것이 모델이며 선형회귀모델이란 두 변수간의 구체적 관계를 수학적으로 설명하고 있는것이다.

3-1) 선형 회귀모델의 추청: 최소자승법(OLS)

실제 값을 담고 있는 벡터를 $\boldsymbol{y}$ 모델을 통해 추정한 값을 $\hat{\boldsymbol{y}}$ 라 할 때 , 이 둘의 차이를 잔차(residual,e)라고 부르며 다음과 같이 표현 할 수 있다.

$$e = \boldsymbol{y} - \hat{\boldsymbol{y}}$$

이 때 모델의 성능은 잔차제곱합(RSS)으로 표현할 수 있는데,

$$RSS = e_1^2 + e_2^2 + \ldots e_n^2$$

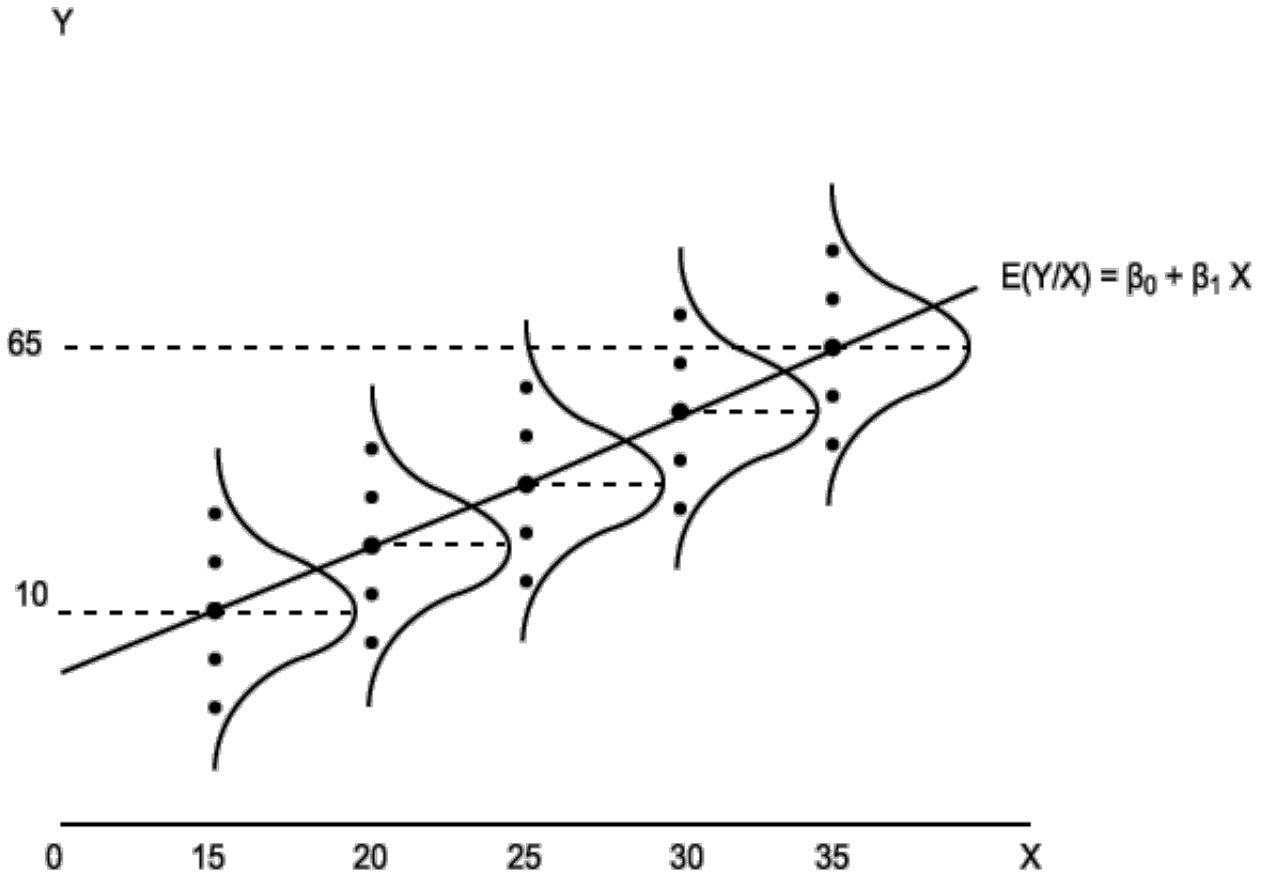
최소자승법(OLS)이란 이 RSS를 최소화하는 계수를 추정하는 알고리즘이다.OLS는 이상치에 매우 민감한 특징을 갖기 때문에 이상치가 1개라도 존재할 시 모델의 성능이 저하된다. 따라서 이 한계를 극복하기 위한 방법으로 L2노름 방식의 거리 측정방법이 아닌 L1노름인 LAD방식이 존재한다. 이 경우

$$|e_1| + |e_2| + \ldots |e_n|$$

를 최소화하는 계수를 추정하게 된다.
그 밖에도 L1과 L2방법을 혼합한 "hurb", quantile의 개념을 도입한 "quantile regression" 등이 존재한다.

3-2) 회귀직선의 계수, 기울기와 절편

기울기



<이미지 출처:http://www.aistudy.com/math/regression_lee.htm>

"기울기란 X의 한 유닛 증가에 연관된 Y의 평균증가"** 로 정의 할 수 있다. 이러한 특성 때문에 (평균으로)회귀 한다하여 회귀모델이라 부르는 것이다. 이 때의 평균이란 X의 값에 따른 Y의 평균이기 때문에 조건부평균이라 부른다. 통계적 회귀에서 그렇게 정규분포에 목을 메는 이유는 사실 여기에 있다. 회귀모델 이라는 것이 X에 따른 조건부 평균들을 연결한 것이기 때문인데, 문제는 평균이라는 개념 자체가 정규분포를 대표하는 값이기 때문이다.
사실 실제 모델을 개발할 때 모든 변수에 대하여 정규분포를 가정할 필요는 없다. 중요한 것은 오차항이 0을 평균으로 하는 정규분포인지 아닌지이다. 그럼에도 불구하고 정규성의 의미에 대해서는 충분히 고려해볼만하다. 이러한 평균과 정규분포 모델의 강건성에 대하여 잘 설명한 글이 있기 때문에 아래 보충을 통해 소개한다.

3-3) 회귀계수의 추정과 해석

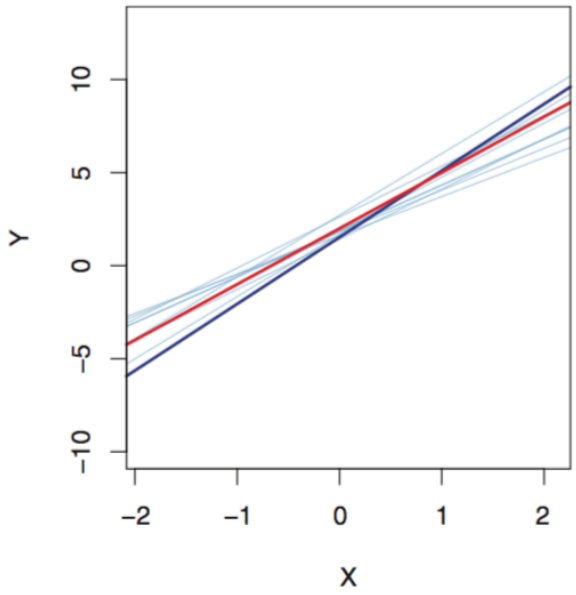
	Coefficient	Std. error	t-statistic	p-value
Intercept	7.0325	0.4578	15.36	< 0.0001
TV	0.0475	0.0027	17.67	< 0.0001

<표 출처:R로 실습하는 : 가볍게시작하는 통계학습(2016),마이클 옮김,루비페이퍼>

TV와 타겟변수간의 선형관계에 대해 회귀모델의 기울기는 0.0475며 p값은 0.0001로 유의하다는 것이다. 이 때 $H_0 : \beta_1 = 0$ 이기 때문에 p값에 의해 H_0를 기각하고 $H_0 : \beta_1 \neq 0$을 채택하게 된다.

말은 어렵지만 그냥 쉽게 해석해서 타겟변수와 TV변수의 관계는 우연에 의한것이 아니라고 보는게 맞으며 TV가 1단위 증가할 때 안할때 보다 0.0475 증가한다는 소리이다.

3-4) 회귀 계수의 추정과 신뢰구간



<이미지출처:R로 실습하는 : 가볍게시작하는 통계학습(2016),마이클 옮김,루비페이퍼>

통계학이 항상 그러하듯 내가 갖고 있는 데이터는 전수데이터가 아니다. 그렇기 때문에 표본추출을 통한 대표성을 확보하곤 한다. N이 충분히 클경우(보통 30이상) 중심극한 정리에 의해 표본 평균이 정규분포로 수렴하고 이 때 표본평균의 평균은 모평균에 수렴하게 되는데 이는 우리가 만든 회귀직선과 모 회귀직선간의 관계로 확장할 수 있다.

회귀 직선 또한 중심극한 정리에 의해 여러 표본에서 추정한 표본 회귀계수들을 평균할 경우 모 회귀직선에 수렴하는 것을 확인할 수 있는데 이를 통해 내가 추정한 회귀계수에 대한 신뢰구간과 p값을 통해 회귀직선에 대한 추가적 정보를 얻을 수 있다.

3-5) 모델의 적합성과 정확도 평가

추정된 계수에대한 검정외에도 모델자체의 정확도를 측정하는 방법이 여러개 있다. 그 중 한가지는 잔차표준오차 RSE다 RSE는

$$RSE = \sqrt{\frac{RSS}{n-2}}$$

로 예측값과 실제값이 가까울수록 이 값은 줄어들게 될 것이다.

또다른 모델 적합성 지표로는 R^2 통계량이 있다. R^2 은 X를 사용하여 Y를 설명할 수 있는 변동비율이며 이는 아래와 같이 표형 할 수 있다.

$$R^2 = 1 - \frac{RSS}{TSS}$$

R^2 이 높다는 소리는 독립변수가 반응변수의 많은 부분을 설명하고 있다는 것을 말한다. 따라서 모델이 잘 적합한다는 것이다. 하지만 이 R^2 은 비율은 분산을 설명하는 특성상 F검정을 따른다.따라서 F검정을 이용해서 R^2 통계량을 검정하게 된다.

TSS 와 RSS, ESS

TSS는 종속변수($\boldsymbol{y}$)의 분산을 의미한다. 이에 반해 ESS는 회귀분석 결과 예측된 $\hat{\boldsymbol{y}}$의 분산을 의미한다.

$$TSS = Var(\boldsymbol{y})$$

$$ESS = Var(\hat{\boldsymbol{y}})$$

$$RSS = Var(\boldsymbol{e})$$

따라서

$$TSS = ESS + RSS$$

가 되는것이 당연할 것이다. 즉 $ESS = TSS - RSS$ 이기 때문에

$$R^2 = \frac{TSS - RSS}{TSS} = \frac{ESS}{TSS}$$

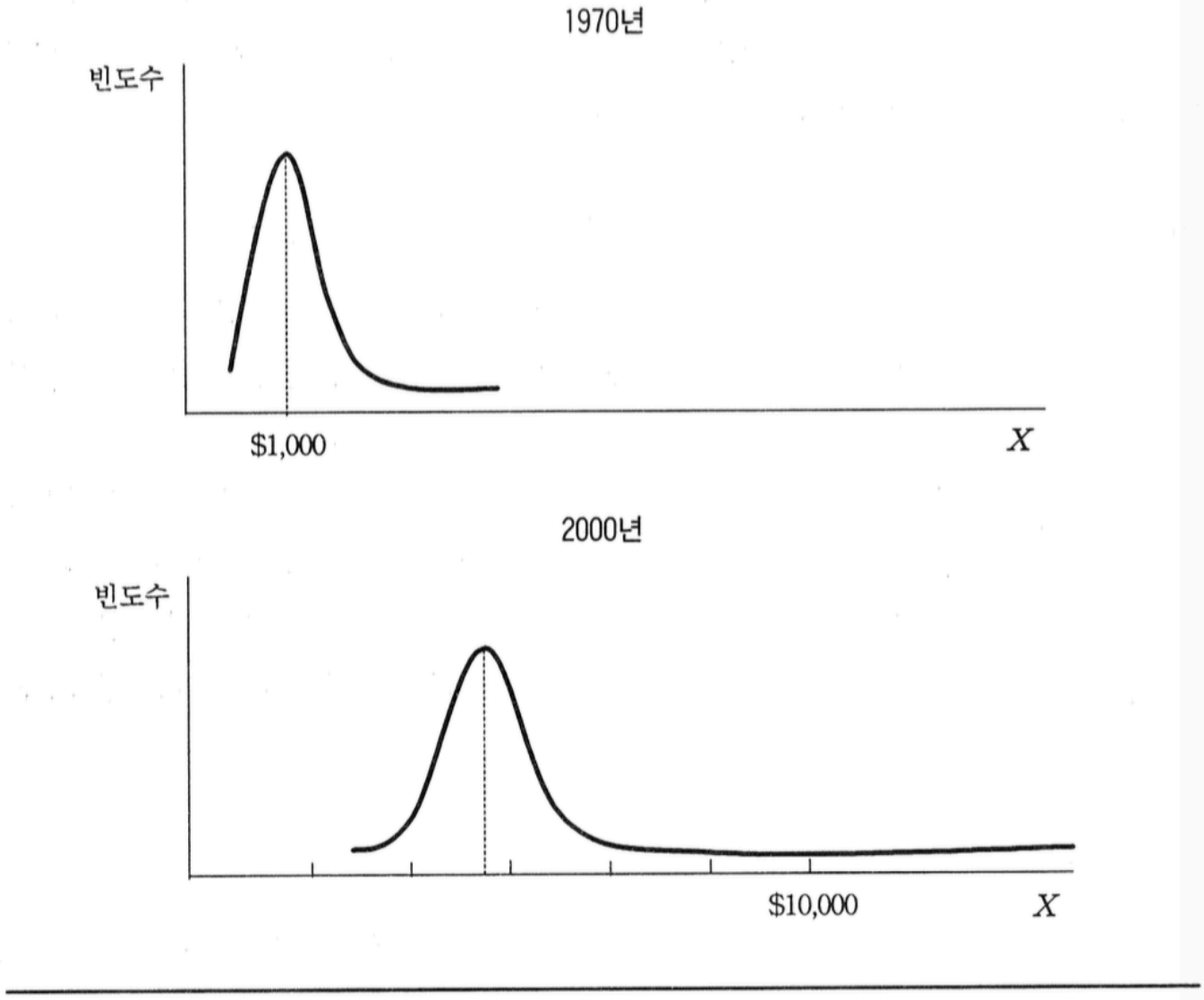
가 된다. 따라서 R^2 란 예측값이 실제값의 분산을 얼마만큼 설명하는지가 되는 것이다.

보충 강건성(robustness)와 변수변환

학생들에게 통계학을 가르치면서 가장 신경이 쓰이는 것은 학생들이 데이터를 살펴보지도 않고 무조건 통계패키지를 사용해서 결과물을 얻고 이 결과물을 자세히 검토해 보지도 않고 문제를 해결했다고 생각하는 태도이다. 대부분의 통계분석기법들은 정규성(즉, 변수가 정규분포를 갖는다는 가정)을 바탕으로 하고 있다. 물론 많은 통계기법들의 이론이 정규성가정에서 이탈해도 성립된다는 강건성(robustness)를 가지고 있다고 해도 우리는 데이터를 받았을 때 우선 정규성 여부를 조사해보아야 한다. 만일 정규성이 없다고 생각되면, 변수들을 변환(transform)해서 정규성을 갖는 새로운 변수를 찾아 이를 분석하는 노력을 기울여야 할 것이다.

우선 왜 변환이 필요한가를 생각해 보자. 어느 나라의 일인당 국민소득이 1970년에는 1000달러이었다가 경제능력이 열 배가 되어 무척 잘 살게 되었다고 선전을 한다고 하자. 물론 전반적으로 잘 살게 된 것은 사실이지만 달러의 가치절하를 염두해 두더라도 어쩐지 열 배라는 말이 어색하다고 많은 국민들이 생각할 것이다. 통계적으로 보아 일인당 국민소득은 국민 개개인의 소득을 나타내는 타당한 측도인가? 우리는 달러의 가치변화가 없었다고 가정하고 이 문제를 생각해 보자.

【그림 6.1.1】 가상된 일인당 국민소득분포



위 그림은 1970년과 2000년의 국민소득 분포들이 그려져 있다.

일인단 국민소득은 이 두 분포들의 평균이다. 과연 이 두 분포에서 평균은 무엇을 의미하는가? 소득이 다른 세 사람이 있다고 하자. 1970년에는 각각 400달러, 600달러, 2000달러를 별어 평균이 1000달러가 되었다. 2000년에는 각각 1000달러 9000달러 20000달러를 벌어서 평균이 10000달러가 되었다. 이 세사람 모두가 그 동안 경제력이 여섯 배로 증가되었다는 데 동의하지는 않을 것이다. 평균값이란 정규분포를 바탕으로 했을 때의 대표값이다. 예를 들어 정규분포보다 분포의 꼬리 쪽이 무거운 이중지수분포(왜도가높은 분포를 말한다.)의 경우에는 메디안을 사용하는 것이 더 타당하다. 또한 베이즈 정리를 사용해서 정규분포가 평균값을 대표값으로 갖기 위한 확률분포임을 보일 수 있다. 또한 정규분포는 평균값이 주어졌을 때 최대엔트로피를 갖는 확률분포이다. 좀더 쉽게 설명하기 위해서 데이터값 $x_1, x_2, \ldots . x_N$을 생각해 보자. 평균m은 제곱합 $\sum_{i=1}^{N} (x_i - m)^2$을 최소화하는 값이다. 이 제곱합은 x_i값을 대표값 m으로 나타낼 때 오차들을 하나로 묶어놓는 여러 측도 중의 하나일 뿐이고, 또한 정규분포의 지수항에 해당한다. 따라서 오차들을 하나로 묶는데 다른 측도를 사용한다면, 당연히 다른 대표값과 다른 분포를 사용하는 것을 가정하여야 한다. 이러한 여러 가지 이유로, 평균값을 대표값으로 하기 위해서는 정규분포를 갖는다는 가정을 하는 것이 타당하다. 위 그림의 확률분포들은 정규성을 갖는다고 할 수 있을까?

어느 누구도 이에 동의하지 않을 것이다. 따라서 우리는 변수 X를 변환해서 정규분포에 가깝도록 만들어야 한다.

즉 정리하자면 대표값의 정의는 오차 제곱합을 최소화하는 지점을 말한다. 이 대표값으로 평균을 사용했다는 점 자체가 이미 데이터의 분포가 정규성을 띈다는 가정을 전제로 사용한다는 것이다. 즉 평균을 전제로 하는 모든 모형과 검정 추정 등등은 정규분포를 가정할 수 밖에 없는것이다.

인용

R로 실습하는 : 가볍게시작하는 통계학습(2016),마이클 옮김,루비페이퍼 67p
SAS를 이용한 현대통계학