자료 재표현 목적

개별 변수

o '평균'에 관련된 추론 통계 : 치우침이나 이상치는 표본평균을 왜곡하고 표본분산을 크게하여 신뢰구간이나 가설검정은 적절하지 않음
o 이에 치우침이 있는 표본데이터는 좌우 대칭 형태로 변환하여 추론을 해야 함.

함수 관계

o 두 변수 간의 선형 함수관계 : $Y = a + bX$ => easy to use and interprete
o 선형관계가 무너진 경우 선형화

분포 대칭화 (2 장 정리)

Power 변환방법 $X^* = X^p$

차수 p	예제	해결내용
∞	$X^* = \exp(x)$	Severe left skewed
2, 3, 4	$X^* = X^2$	Mild left skewed
1/2, 1/3, 1/4	$X^* = \sqrt{X}$	Mild right skewed
0	$X^* = Ln(x)$	Right skewed
-1/2, -1, -2	$X^* = -\frac{1}{X}$	Severe Right skewed

변환 실습

(1) 우로 치우친 데이터를 50 개 생성하고 평균에 대한 95% 신뢰구간을 구하고 치우침을 판단하시오. 그리고 치우침을 해결하는 적절한 방법을 사용한 후 95% 신뢰구간을 구하고 해석하시오.
(2) 좌로 치우친 데이터를 50 개 생성하고 평균에 대한 95% 신뢰구간을 구하고 치우침을 판단하시오. 그리고 치우침을 해결하는 적절한 방법을 사용한 후 95% 신뢰구간을 구하고 해석하시오.

치우침 진단은 (1) Rule of Thumbs 시각적 : 히스토그램 => (값에 의한 판단) 평균과 중앙값 크기 비교, 수학 왜도, EDA 왜도 (2) 검정통계량 : AD 방법, SW 방법 등

Professor Kwon, Sehyug | Dept. of Statistics, HANNAM Univ. 010.6365.7622 http://wolfpack.hnu.ac.kr

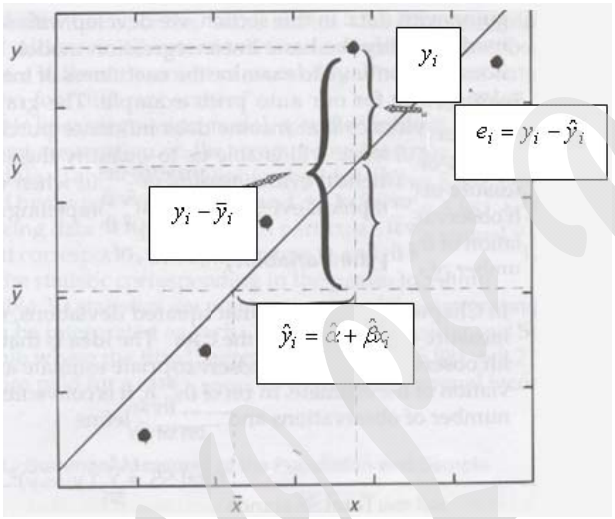
함수관계 선형화

산점도

(정의)두 측정형 변수의 함수관계를 알아보기 위하여 2 차원 공간에 데이터를 표현한 그래프
(활용) 두 변수의 함수관계 시각적 판단, 관계에서의 이상치 판단

모수적 분석방법

회귀분석 실시하여 (1)회귀계수의 유의성 및 (2)결정계수 값을 비교
(회귀분석) $Y = a + bX + e$ 회귀모형에서 선형 관계 성립여부 판단
(회귀계수 a, b 추정) OLS 추정치 +=> 최소자승추정치



$$\min_{a,b} \sum e^2 = \min_{a,b} \sum (y - \boldsymbol{a} - bx)^2$$

추정 회귀계수 $\hat{b}$ 를 이용하여 ($H_0 : b_0 = 0$ ⇔ 설명변수 X 는 유의하지 않음, 종속변수 Y 와 설명변수 X 는 선형 관계 없음) 가설을 검정한다.

검정통계량 $TS = \dfrac{\hat{b} - b_0}{s(\hat{b})} \sim t(n-2)$

결정계수 Determinant Coefficient $0 \le \boldsymbol{R}^2 = \dfrac{SSR}{SST} \le 1$

- 설명변수가 하나인 경우 상관계수의 제곱이 결정계수
- 결정계수는 종속변수의 총변동(SST) 중 설정된 모형 $Y = a + bX$ 이 설명하는 변동이므로 모형의 설명력 척도

선형화 방법 $Y = a + bX^{\lambda}$ (Tukey 변환 사다리)

Table 1. Tukey's Ladder of Transformation

λ	-2	-1	-1/2	0	1/2	1	2
y	$\frac{1}{x^2}$	$\frac{1}{x}$	$\frac{1}{\sqrt{x}}$	$\log x$	$\sqrt{x}$	x	x^2

종속변수 Y가 음의 값을 갖는 경우

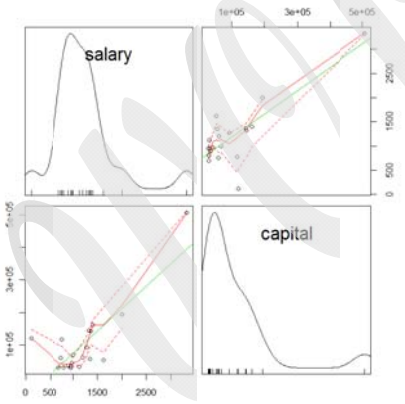
Table 2. Modified Tukey's Ladder of Transformation

λ	-2	-1	-1/2	0	1/2	1	2
y	$\frac{-1}{x^2}$	$\frac{-1}{x}$	$\frac{-1}{\sqrt{x}}$	$\log x$	$\sqrt{x}$	x	x^2

CEO.xls

```
ds=read.csv("CEO.csv")

library(car) #산점도 행렬 함수 라이브러리
scatterplot.matrix(~salary+capital,data=ds) #산점도 행렬
summary(lm(salary~capital,data=ds)) #회귀분석결과 출력
```



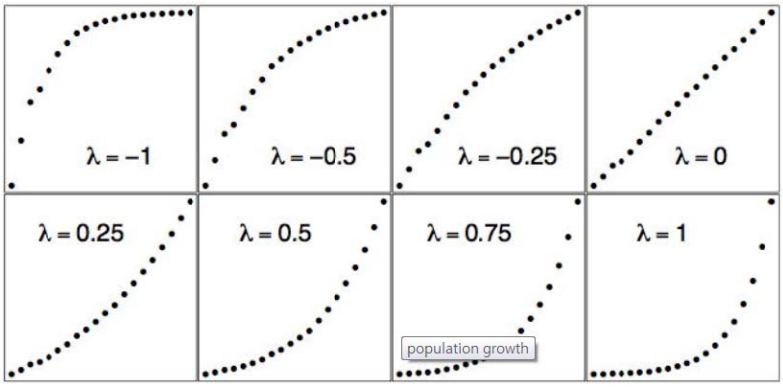

```
Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) 7.070e+02  1.180e+02   5.994 1.14e-05 ***
capital     4.815e-03  8.076e-04   5.963 1.22e-05 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.

Residual standard error: 379.8 on 18 degrees of freedom
Multiple R-squared: 0.6639,     Adjusted R-squared: 0.6
F-statistic: 35.55 on 1 and 18 DF,  p-value: 1.216e-05
```

(해결방안)

```
scatterplot.matrix(~log(salary)+log(capital),data=ds) #산점도 행렬
summary(lm(log(salary)~log(capital),data=ds)) #회귀분석결과 출력
```

회귀계수의 유의확률 판단, 결정계수 값 비교

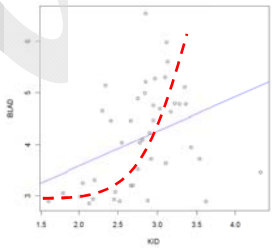


CANCER.CSV

1. CIG = Number of cigarettes smoked (hds per capita)
2. BLAD = Deaths per 100K population from bladder cancer
3. LUNG = Deathes per 100K population from lung cancer
4. KID = Deaths per 100K population from bladder cancer
5. LEUK = Deaths per 100 K population from leukemia

```
ds=read.csv("cancer.csv")
attach(ds)
plot(BLAD~KID)
abline(lm(BLAD~KID),col="blue")
summary(lm(BLAD~KID))
```

Sqrt 변환하기 => sqrt(KID), log(KID)



```
Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)   2.2572     0.7607   2.967  0.00494 **
KID           0.6670     0.2677   2.491  0.01676 *
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1

Residual standard error: 0.9113 on 42 degrees of freedom
Multiple R-squared: 0.1287,     Adjusted R-squared: 0.108


Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)   2.0727     0.7228   2.868  0.00644 **
log(KID)      2.0281     0.7031   2.884  0.00616 **
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 '

Residual standard error: 0.892 on 42 degrees of freedom
Multiple R-squared: 0.1653,     Adjusted R-squared: 0.1455
```

표준화 변환

방법 $z(x) = \dfrac{x - \mu(x)}{\sigma(x)} \sim (0,1)$

활용

o 표준화는 평균과 표준편차가 다른 집단들의 상호 비교가 필요할 때 (예) 수능 표준점수
o 서로 다른 업무 (시험)를 담당하는 부서원(학생)들을 평가하고자 할 때 (예) 경제학과와 통계학과 학생들에게 '통계학개론'을 강의하는 교수가 중간고사를 보려고 한다. 같은 날 시험을 볼 수 없어 시험 문제를 달리하여 시험을 보았다. 두 학과 학생을 합하여 성적을 부여하고자 할 때 => 두 학과 서로 표준화 (가정) 두 학과 학생의 능력은 동일하며, 시험점수 평균의 차이와 표준편차 차이는 시험 문제의 상이하기에 발생하였다.

문제점

두 학과(집단) 시험 점수의 평균과 표준편차의 차이가 학생들의 능력에서 온 것이라면?

Robust 표준화 $\tilde{z}(x) = \dfrac{x - \tilde{x}(\text{중위수})}{\tilde{\sigma}(x)}$, $\tilde{\sigma}(x) = \dfrac{IQR}{1.35}$

변환 실습

경제학과 학생 50 명의 시험 점수는 $N(50,10)$ 에서 생성하고 통계학과 40 명 학생 중 30 명은 $N(60,10)$ 그리고 10 명은 $N(80,5)$ 에서 생성하여 두 학과 학생들의 표준화 점수 분포를 히스토그램으로 나타내시오.

Professor Kwon, Sehyug | Dept. of Statistics, HANNAM Univ. 010.6365.7622 http://wolfpack.hnu.ac.kr