1

산점도

1. 정의

2개의 측정형 변수 데이터를 2차원 공간에 표현하여 두 변수의 함수 관계를 예상함

- X- 축 : 결정의 요인, 설명변수, 독립변수, 예측변수
- Y-축 : Output, 종속변수, 목표변수

2. 활용

1) 함수관계

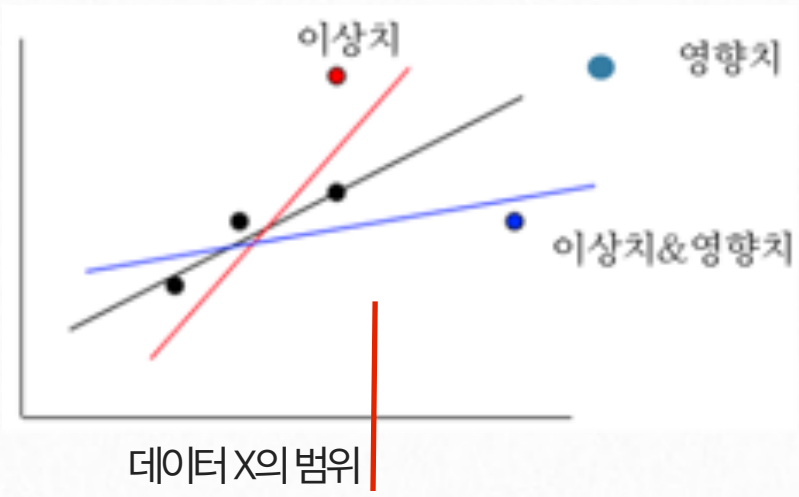
두 변수의 함수 관계를 판단 : 선형 모형에서는 직선의 관계를 판단한다.

예제 데이터

SAS Baseball 데이터 : [내용] [다운받기]

```
library(sas7bdat)
ds<-read.sas7bdat('baseball.sas7bdat')
names(ds) # List of Variables in DS
ds$Bat_yr<-ds[,10]/ds[,9] #Rescale per Yr
ds$Hit_yr<-ds[,11]/ds[,9]
ds$HR_yr<-ds[,12]/ds[,9]
ds$Run_yr<-ds[,13]/ds[,9]
ds$RBI_yr<-ds[,14]/ds[,9]
ds$BB_yr<-ds[,15]/ds[,9]
ds0<-ds[,c(9,22,24:30)] #Anaysis ready
sum(as.numeric(is.na(ds0))) #check missing
ds00<-na.omit(ds0) #data cleansing
```

2) 이상치 영향치 진단



(1) 이상치 outlier

선형 함수 관계에서 적합 직선을 많이 벗어난 관측값 - 실제 오차의 분산 기준 2* σ를 벗어남

설명변수 값은 관측 값의 범위 내에 있음

(진단) 오차의 추정치인 Studentized 잔차가 ± 2 벗어남 - 상세한 내용은 잔차 진단 참고

(해결) 삭제 - 물론 잔차분석 후에 실기

(2) 영향치 influential

설명변수 값이 극단 값(다른 관측치와 떨어져 있고 두 변수의 함수 관계에 영향을 주는 관측값

- 순수 영향치 : 함수 회귀 추정 식 상에 있어 함수 관계 (기울기 변동)에는 영향을 주지 않으나 결정계수 높여 설명변수의 설명 능력을 과다하게 높은 것으로 판단하게 하는 결과 왜곡
- 이상치&영향치 : 결정계수 왜곡, 함수관계 왜곡

(진단) 잔차분석 - Hat 통계량 활용

(해결) 영향치 주변의 관측값을 추가 수집 후 분석, 영향치 값이 실제 발생 가능하지 않은 경우 제외

3) 변수 분포

선형모형의 변수들은 좌우 대칭(정규분포, 특히 종속변수의 경우 오차의 정규성 가정으로 정규분포를 가져야 함)의 분포를 갖는 경우 선형 모형의 적합성과 결과 활용도가 높아짐

예전에는 각 설명변수의 분포를 산점도와 개별적으로 분석하였지만 소프트웨어의 발달로 산점도에 분포를 함께 나타낼 수 있음

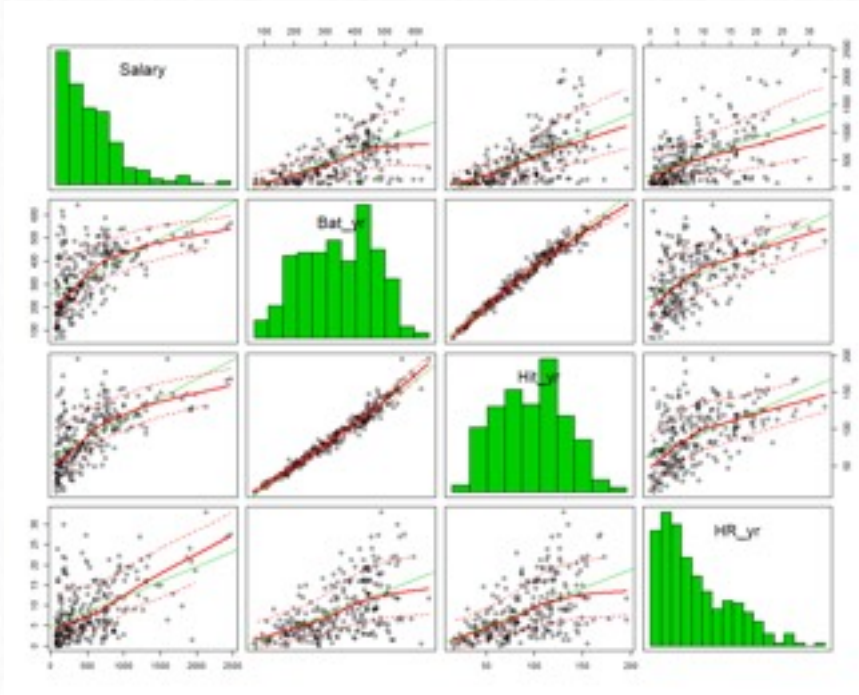
정규성 검정

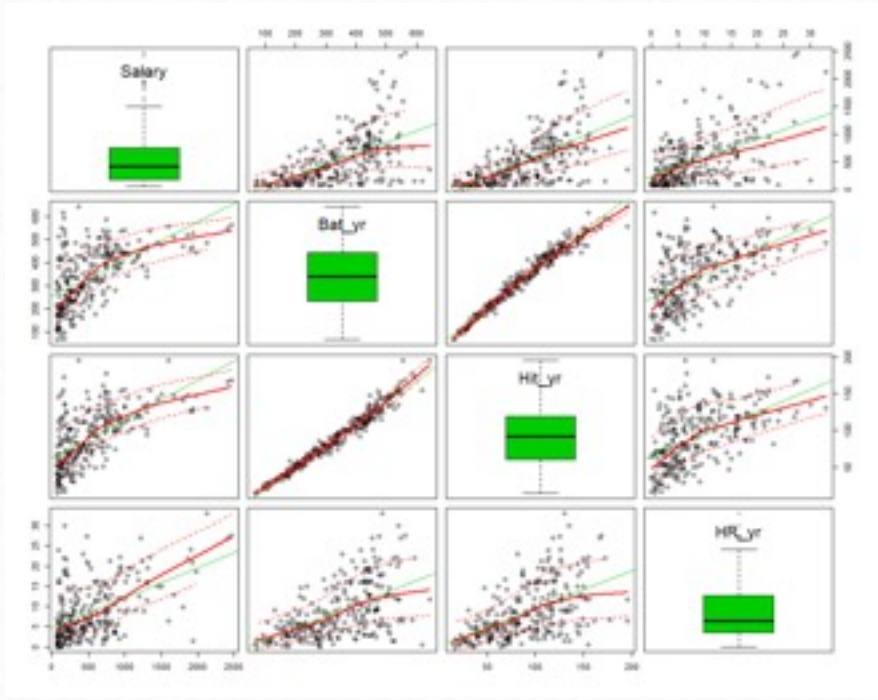
(진단) Shapiro Wilks 검정, Anderson Darling 검정

(해결) Tukey Power 변환, 정규변환

3. 산점도 행렬 그리고 해석

```
library(car)
scatterplotMatrix(data=ds00,~Salary+Bat_yr+Hit_yr+HR_yr, diag='hist')
scatterplotMatrix(data=ds00,~Salary+Bat_yr+Hit_yr+HR_yr, diag='boxplot')
```





산점도 행렬이 그려지고 대각선에 각 변수의 분포를 알 수 있는 "히스토그램" 혹은 "상자-수염 그림"이 그려짐

어느 것이 좋나? 각 변수의 이상치 진단이 가능한 상자 수염 그림이 적절함

연봉, 홈런은 우로 치우친 경향을 보임

(1) 정규성 검정

```
library(nortest)
ad.test(ds00$Salary)
shapiro.test(ds00$Salary)
```

```
> ad.test(ds00$Salary)

        Anderson-Darling normality test

data:  ds00$Salary
A = 9.3077, p-value < 2.2e-16

> shapiro.test(ds00$Salary)

        Shapiro-Wilk normality test

data:  ds00$Salary
W = 0.85111, p-value = 3.307e-15
```

(2) 정규변환

1) 일반적 접근

- 우로 치우침 : $\sqrt{x} \rightarrow x^{1/3} \rightarrow ln(x)$
- 좌로 치우침 : $\sqrt{k-x} \rightarrow (k-x)^{1/3} \rightarrow ln(k-x)$, 상수 k, 혹은 $x^2 \rightarrow x^{1/3}$

2) Tukey Ladder of Power $x' = x^{\lambda}$

- 좌로 치우침 : $\lambda = 2(x^2) < 3(x^3)$

우로 치우침 : $\lambda = 1/2(\sqrt{n}), 1/3(x^{1/3}), 0(ln(x))$

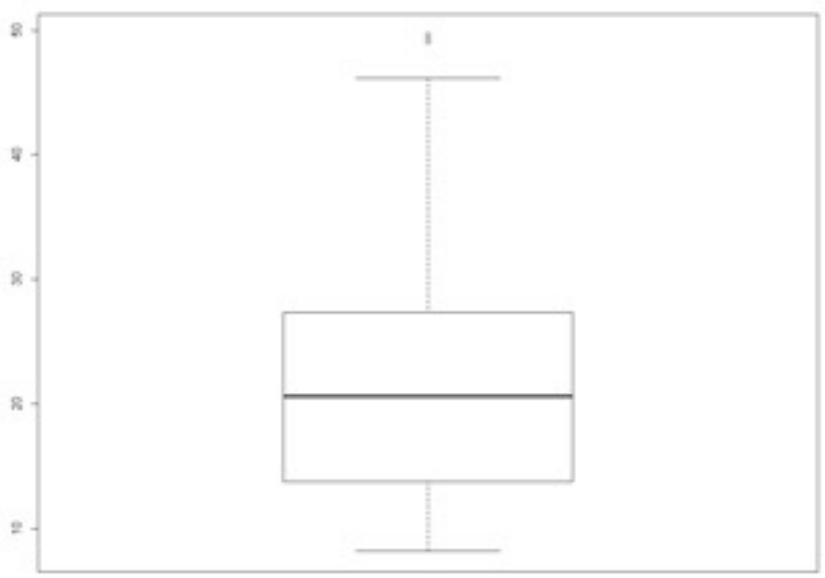
```
shapiro.test(sqrt(ds00$Salary))
boxplot(sqrt(ds00$Salary))
shapiro.test(log(ds00$Salary))
boxplot(log(ds00$Salary))
library(rcompanion)
transformTukey(ds00$Salary)
```

(제곱근 변환) - 여전히 우로 치우침, 이상치 진단

```
> shapiro.test(sqrt(ds00$Salary))

        Shapiro-Wilk normality test

data:  sqrt(ds00$Salary)
W = 0.95398, p-value = 2.174e-07
```

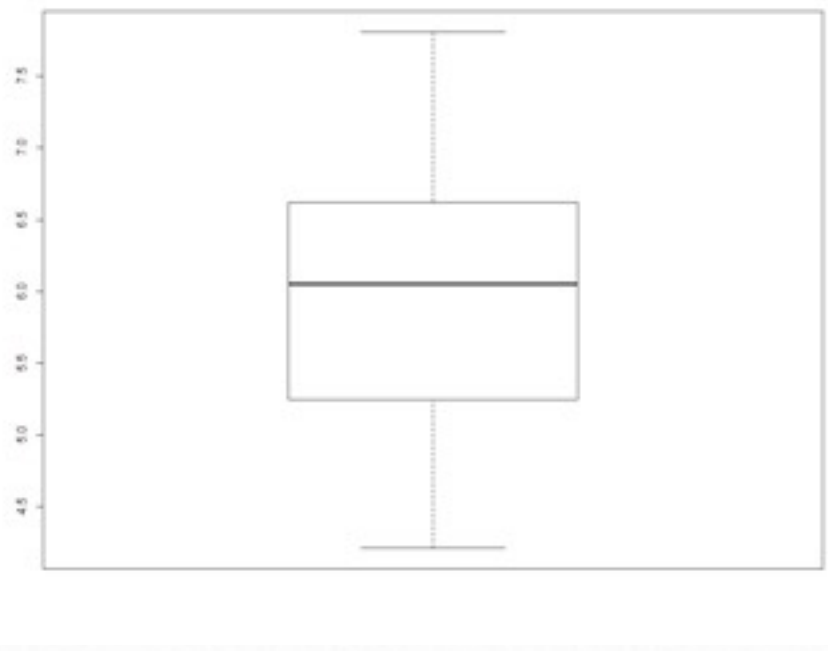


(로그 변환) 유의확률은 여전히 정규성 기각, 상자 그림은 좌우 대칭으로 보임 - 약간 좌로 치우침

```
> shapiro.test(log(ds00$Salary))

        Shapiro-Wilk normality test

data:  log(ds00$Salary)
W = 0.97101, p-value = 3.514e-05
```



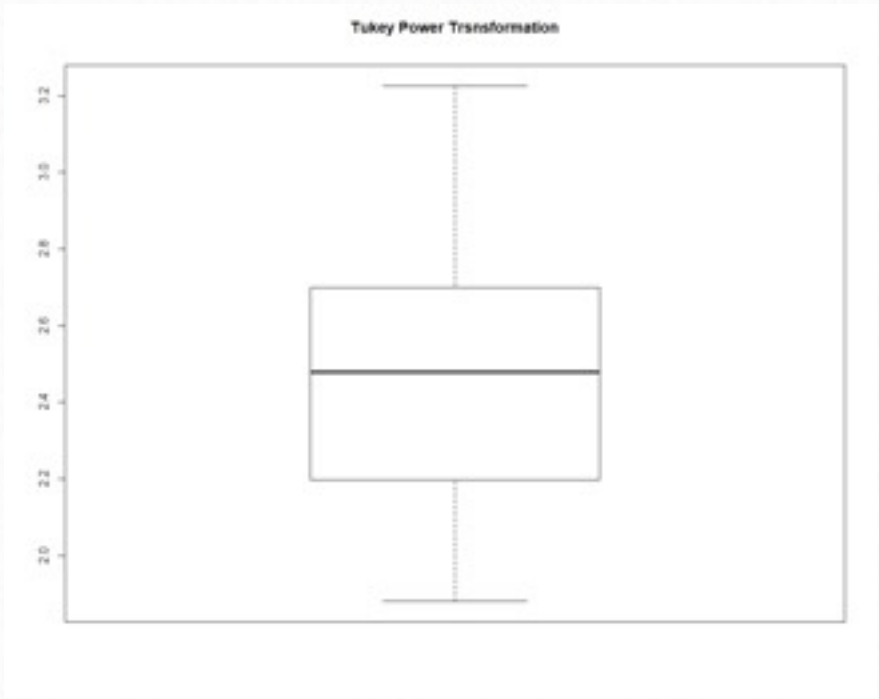
(튜키의 최적 파워 계수) $\lambda = 0.15$, 유의 확률은 0.00017로 여전히 정규성 파괴

```
> transformTukey(ds00$Salary)

    lambda      W Shapiro.p.value
407   0.15 0.9756        0.0001793

if (lambda >  0){TRANS = x ^ lambda}
if (lambda == 0){TRANS = log(x)}
if (lambda <  0){TRANS = -1 * x ^ lambda}
```

```
boxplot(ds00$Salary^0.15, main="Tukey Power Trsnsformation")
```



최종 변환 선택?

로그변환(LOG(x)) 혹은 $X^{0.15}$ 중 하나 선택하면 됨 -
SAS에서는 로그 변환을 택한 것임 -> 향후 회귀분석에서는 변환된 변수를 사용하여 분석함