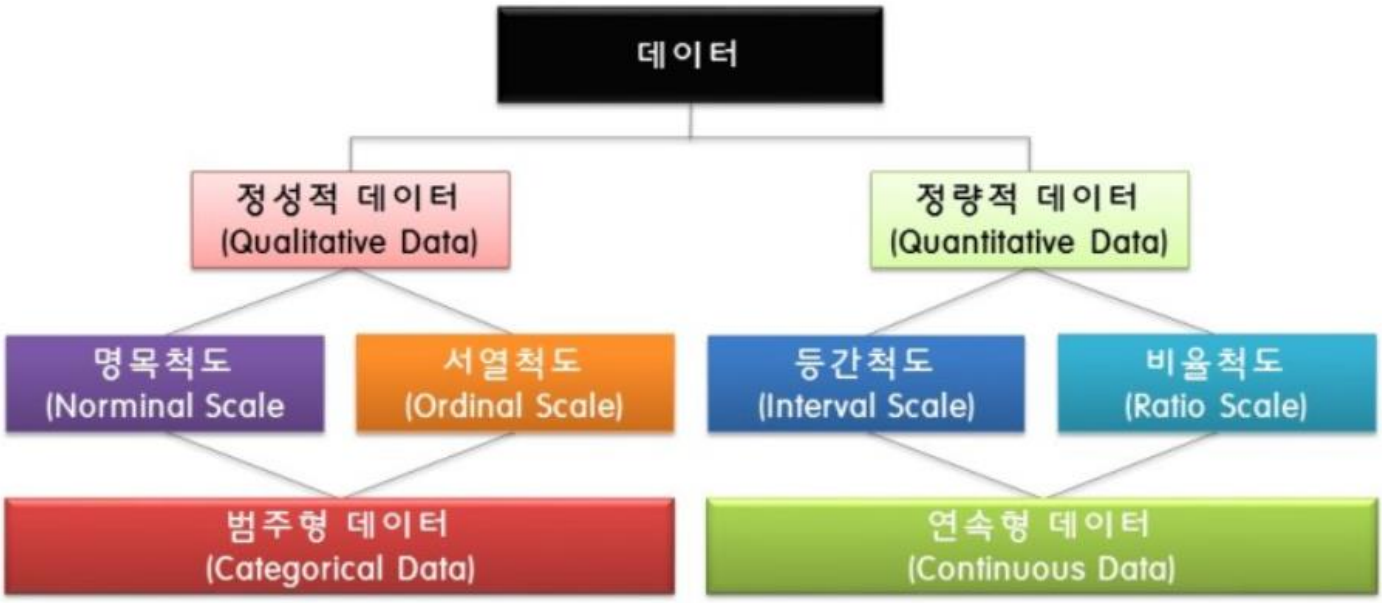


전통적인 통계분석

모수를 추정하기 위해 샘플링하고 추정하는 작업

주로 추론 분석을 진행함

빅데이터 통계분석

이미 모수를 가지고 있다. 데이터 안에 숨겨진 패턴, 규칙을 찾아서 원하는 분석 결과를 도출

데이터를 파악하려는 단계(EDA)에서 기술통계 분석을 진행한다.

[기술통계(descriptive statistics)]

데이터의 특성을 이해하기 쉽게 기술하는 통계라고 한다.

수집한 데이터 정리, 그래프나 숫자 등으로 요약, 표현하는 등 데이터의 특성을 규명하는 통계적 방법이다.

평균, 중앙값, 최빈값, 범위, 분산, 표준편차, 사분위수, 도수분포표, 왜도 첨도 등을 활용한다.

[상관분석]

두 변수간에 어떤 선형저ㄱ 또는 비선형적 관계를 갖고 있는지를 분석하는 방법

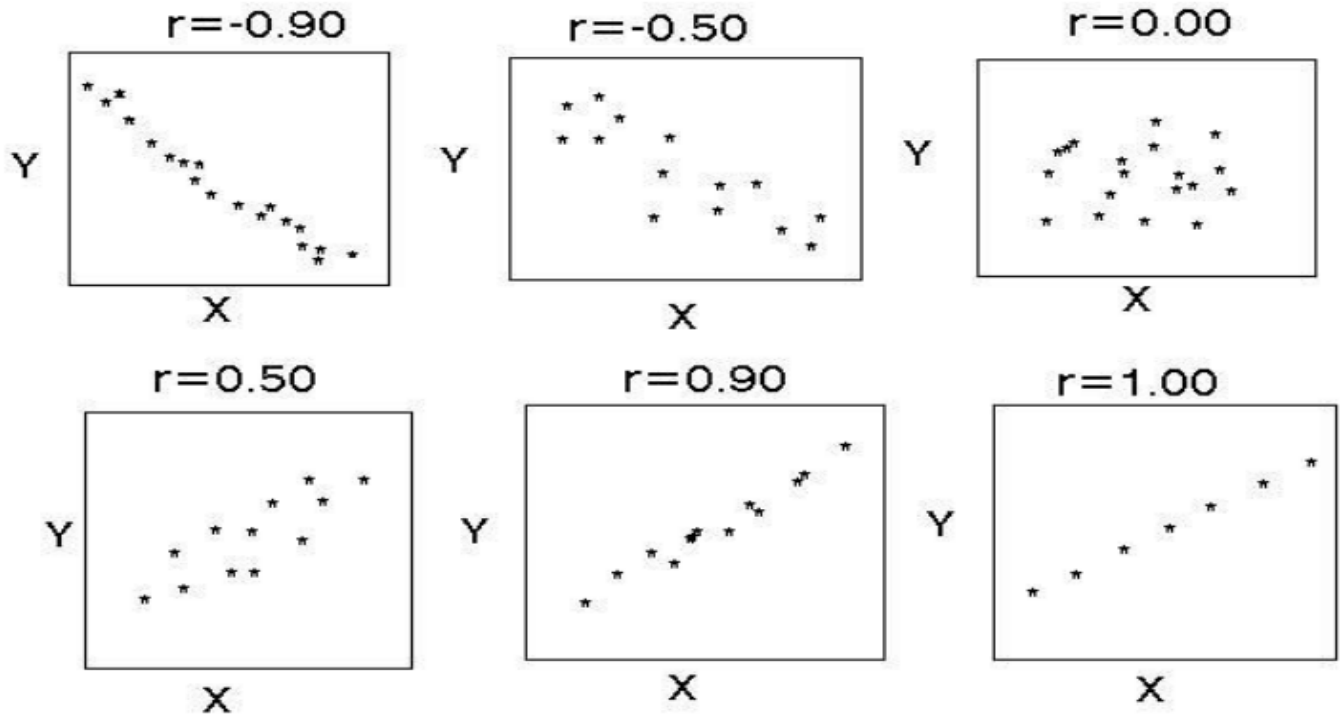
두 변수는 서로 독립적인 관계이거나 상관된 관계일 수 있으며 이때 두 변수 간의 관계의 강도를 상관계수라고 한다.

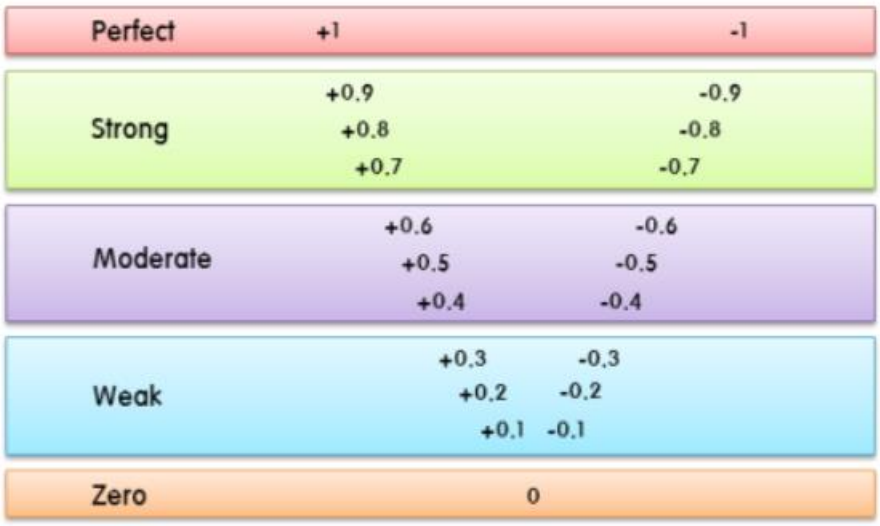
상관관계의 정도를 파악하는 상관계수는 두 변수의 연관 정도만을 나타낸다.

피어슨 상관계수 : 두 연속성 변수간의 선형적인 관계를 측정하여 -1 과 1 사이의 값을 갖는다. 1 또는 -1에 가까울수록 뚜렷한 선형적인 상관관계를 갖는다. 산점도를 그려서 두 변수간의 상관관계를 시각적으로 파악할 수 있다.

X변수와 Y 변수의 상관계수 = X와 Y의 공분산/X의 표준편차 * Y의 표준편차

(X와 Y가 함께 변하는 정도를 X와 Y가 각각 변화는 정도로 나눈다.)





R에서 지원하는 피어슨 상관계수를 계산하는 함수 : cor(X, Y)

cor(백터 또는 매트릭스 또는 데이터 프레임, 벡터 또는 매트릭스 또는 데이터프레임), method = c("pearson", "kendall", "spearman"))

--- 여러 변수에 대해서 상관계수를 구하려는 경우에는 corrplot() 이나 corrgram() 등의 함수를 이용해서 시각화는 방법이 더욱 효과적이다.

스피어만 상관계수 : 두 변수에 대한 비선형 관계의 연관성을 파악할 수 있다. 데이터가 서열 척도인 경우 즉, 값 대신 순위를 이용하는 경우에 사용된다. 데이터의 갯수가 적을 때 그리고 데이터의 동률이 많을 때 유용하다.

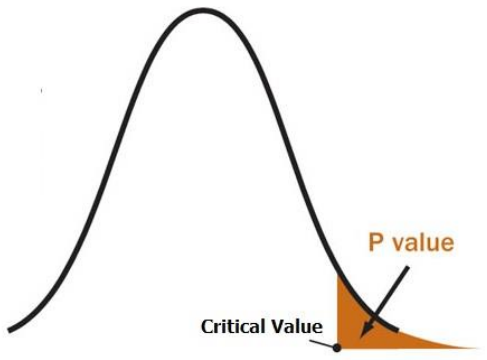
켄달 상관계수 : 두 변수에 대한 비선형 관계의 연관성을 파악할 수 있다. 데이터가 서열 척도인 경우 즉, 값 대신 순위를 이용하는 경우에 사용된다. 일반적으로 켄달 상관계수보다 높은 값을 갖는다.

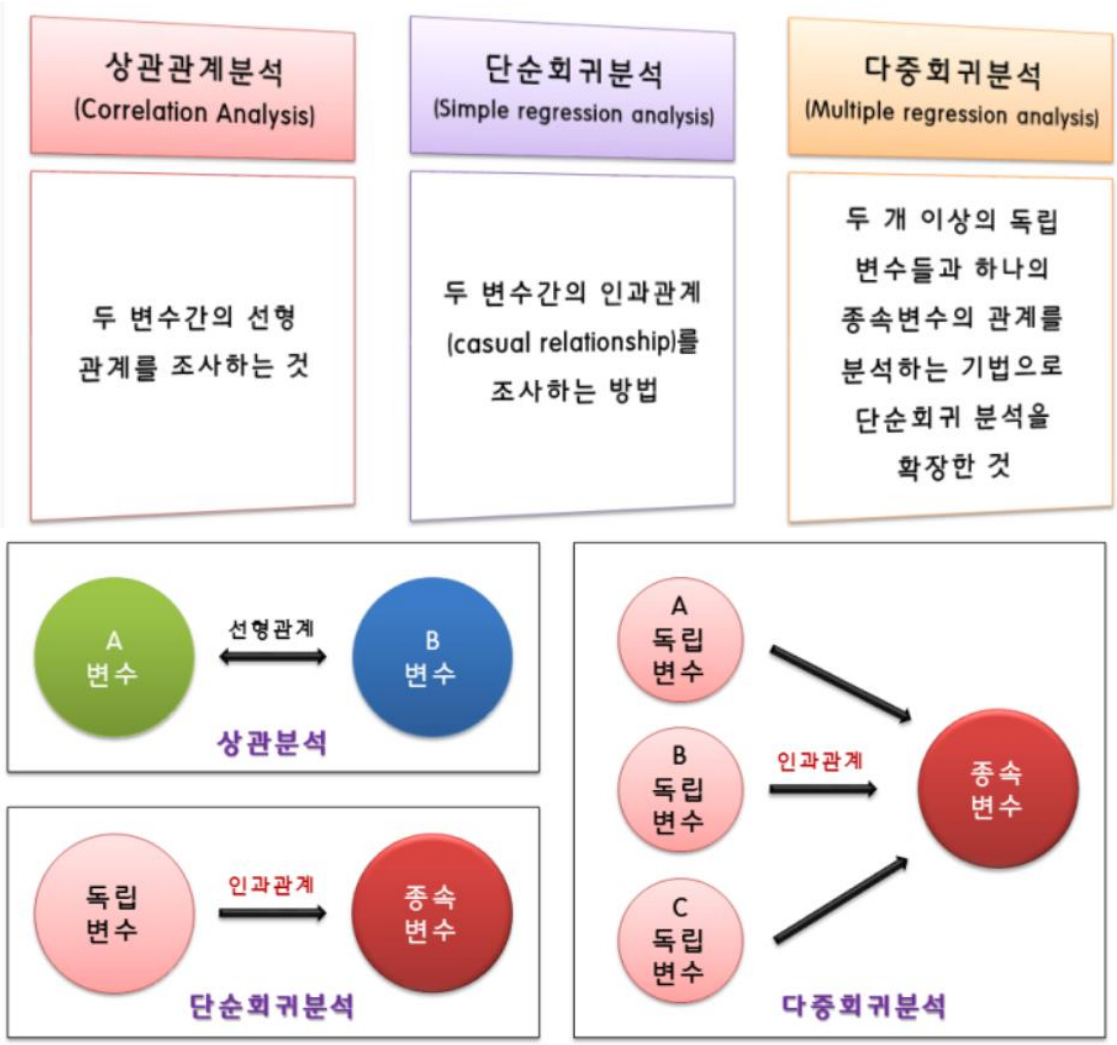
[상관계수 검정]

cor.test() 함수를 사용하여 상관계수의 통계적 유의성 즉, 통계적으로 의미가 있는지 검증하게 된다. 상관분석과 관련해서 확인하고자 하는 대립 가설과 귀무가설은 다음과 같다.
귀무가설 : 상관계수가 0이다.
대립가설 : 상관계수는 0이 아니다.

대립가설이란 주장하려는 가설, 새로운 가설을 의미하며 귀무가설은 대립가설의 반대의 가설로서 통상적으로 인정되는 일반적인 가설을 의미한다. 두 가설을 모두 검증하는 것이 아니라 기존가설(귀무가설)이 잘못됐다는 것을 증명함으로써 새로운 가설(대립가설)을 채택하는 방식을 사용한다. 이 때 사용되는 지표 중 하나가 바로 p-value(유의확률)이다.

p-value(유의확률)은 귀무가설이 참이라는 가정하에 얻은 통계량이 귀무가설을 얼마나 지지하는지를 나타낸 확률로서 일반적으로 0.05 이하이면 귀무가설을 기각하고 이상이면 귀무가설을 채택한다. p-value 값이 적을수록 대립가설이 통계적의미를 갖는다고 할 수 있다. p-value는 단지 주어진 분석에 대해서 귀무가설이 통계적으로 의미가 있는지 여부를 판단하는 기준일 뿐이다.





[선형회귀분석]

선형회귀는 종속변수 Y와 한개 이상의 독립변수(설명변수) X와의 관계를 모델링하는 회귀분석 기법이다.

한 개의 설명변수에 기반한 경우에는 단순선형회귀, 두 이상의 설명변수에 기반한 경우에는 다중선형회귀라고 한다.
추세선을 이용하여 종속변수의 값을 예측하는 모델을 선형회귀 모델이라고 한다.
lm() 함수를 사용하여 선형회귀 모델을 생성할 수 있다.
residuals() 함수로 잔차를 확인할 수 있다.
predict.lm() 함수로 종속변수 데이터를 예측할 수 있다.
결정계수는 추정한 선형 모델이 주어진 데이터에 적합한 정도를 재는 척도다.
summary() 함수로 결정계수, 수정된 결정계수 및 F 통계량, 잔차, 사분위수, 회귀계수를 확인할 수 있다.
coef() 함수를 이용하여 회귀계수만 출력하여 볼 수 있다.

회귀분석은 지금처럼 빅데이터가 등장하기 전에 비교적 적은 데이터로 독립변수나 종속변수의 관계를 수식으로 표현할 수 있어서 논문과 실무에서 가장 많이 사용하는 예측기법이다.

종속변수(반응변수,관심변수) : 영향을 받는 변수

독립변수(설명변수) : 영향을 주는 변수

- 단순선형회귀

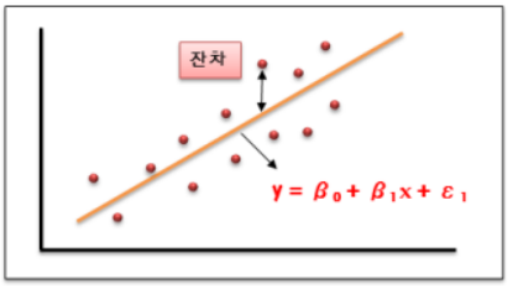
하나의 독립변수와 하나의 종속변수 간의 회귀 분석을 단순 회귀분석이라고 하며 독립변수가 여러 개인 경우를 다중 회귀분석이라고 한다.

독립변수(X) ------→ 종속변수(Y)

독립변수(X1)+ 독립변수(X2)... ------→ 종속변수(Y)

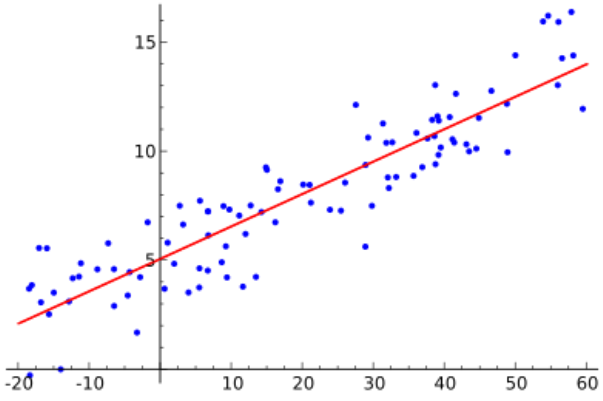
- 독립변수(예측변수) : 영향을 미칠 것으로 생각되는 변수
- 종속변수(기준변수) : 영향을 받을 것으로 생각되는 변수
- 전제조건
  → 독립변수와 종속변수의 설정은 **논리적 타당성을 토대로** 해야함
- 단순회귀분석의 가장 기본적인 과업
  → β_0과 β_1을 구하는 것
  → 기본식

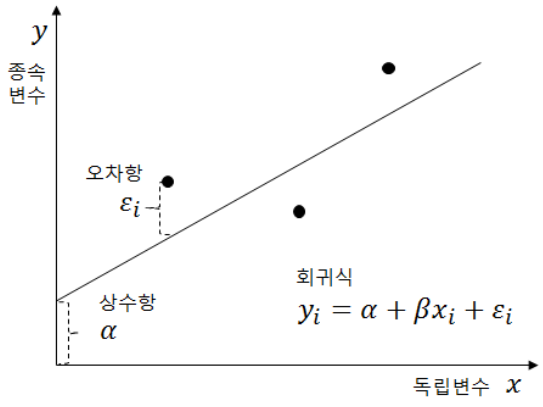
$$y = \beta_0 + \beta_1 x + \varepsilon_1$$



회귀분석은 변수간의 관계를 추정하는 통계방법으로서 선형 회귀분석은 독립변수에 대한 종속변수 값들을 이용해 두 변수간의 선형 관계를 설명하는 회귀선인 직선의 방정식(y = a + bx)를 만들고 임의의 x값에 대한 y값의 추정치를 추론하는 방법이다. 이 때 회귀선은 x에 대응하는 실제 y값과 추정된 y값 사이의 오차인 잔차를 최소화하는 직선을 잘 정해야 설명력이 좋은 회귀 방정식이 된다.

회귀선(추세선)을 찾는 방법은 범위탐색, 통계적 방법 그리고 머신러닝 방법이 사용되며 일반적으로 많이 사용되는 통계적 방법은 잔차들의 제곱합이 최소화 하는 최소 제곱법을 이용한다. 머신러닝 방법은 경사하강법이라는 방법을 이용한다.





lm(종속변수(결과) ~ 독립변수(원인),데이터)

회귀모델 만들기

model <- lm(circumference ~ age, Orange)

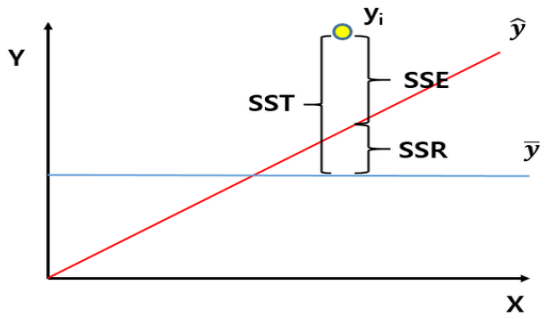
예측하기

predict.lm(model, newdata = data.frame(age = 100))

- 결정계수와 수정된 결정계수

결정계수(R-squared : R^2)는 추정한 선형모델이 주어진 데이터에 적합한 정도를 재는 척도이다. 전체 분산값(SST)과 설명되는 분산값(SSR)으로 계산할 수 있다. 0~1 값을 가지며 1에 가까울수록 설명력이 좋은 모델이다. summary() 함수로 확인할 수 있다.

$R^2 = SSR/SST$



- 단순회귀 모델의 시각화 :

plot(Orange$age, Orange$circumference)

abline(coef(model))

- 다중선형회귀

하나의 종속변수(Y)와 두 개 이상의 독립변수(X)가 있고 오차항()이 있는 선형 관계이다.

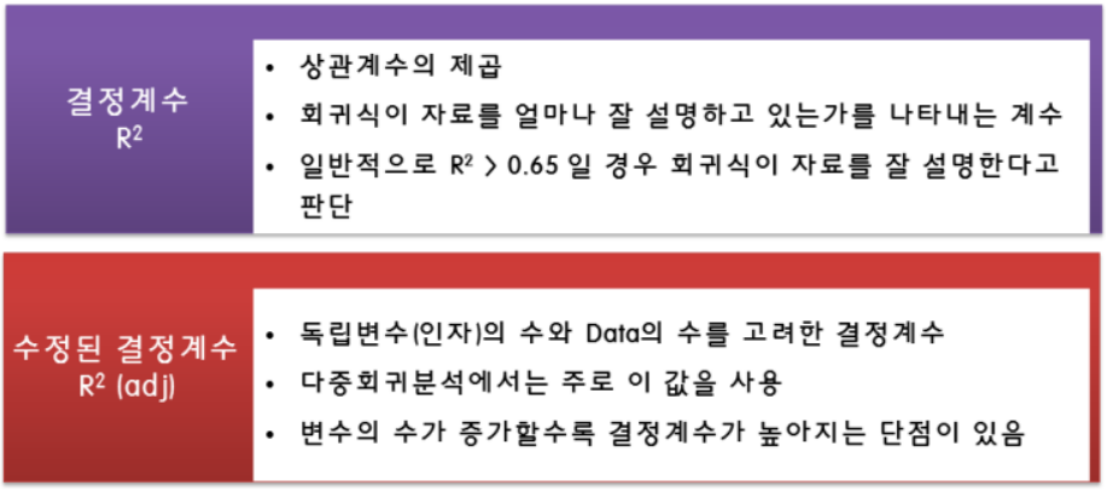
- 두 개 이상의 독립변수들과 하나의 종속변수의 관계를 분석하는 기법으로 단순회귀분석을 확장한 것
- 독립변수(예측변수) : 영향을 미칠 것으로 생각되는 변수
- 종속변수(기준변수) : 영향을 받을 것으로 생각되는 변수
- 다중회귀분석의 가장 기본적인 과업은 각 계수들을 구하는 것

단순회귀식	$\hat{Y} = \hat{\beta}_0 + \hat{\beta}_1 X$
다중회귀식	$\hat{Y} = \hat{\beta}_0 + \hat{\beta}_1 X_1 + \hat{\beta}_2 X_2 + \cdots\cdots + \hat{\beta}_k X_k$

height_father <- c(180, 172, 150, 180, 177, 160, 170, 165, 179, 159) # 아버지 키

```
height_mohter <- c(160, 164, 166, 188, 160, 160, 171, 158, 169, 159) # 어머니 키

height_son <- c(180, 173, 163, 184, 165, 165, 175, 168, 179, 160) # 아들 키

height <- data.frame(height_father, height_mohter, height_son)

head(height)

model <-lm (height_son ~ height_father + height_mohter, data = height)

predict.lm(model, newdata = data.frame(height_father = 170, height_mohter = 160))

predict.lm(model, newdata = data.frame(height_father = 170, height_mohter = 160)

          , interval = "confidence")
```

```
> summary(model)

Call:
lm(formula = height_son ~ height_father + height_mohter, data = height)

Residuals:
    Min      1Q  Median      3Q     Max
-7.9806 -0.8972  1.1166  1.4482  5.5113

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)    21.7437    28.7417   0.757  0.47402
height_father   0.5027     0.1420   3.540  0.00947 **
height_mohter   0.3891     0.1628   2.390  0.04815 *
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 4.116 on 7 degrees of freedom
Multiple R-squared:  0.8022,    Adjusted R-squared:  0.7457
F-statistic: 14.19 on 2 and 7 DF,  p-value: 0.003442
```



[설명 변수의 선택법]

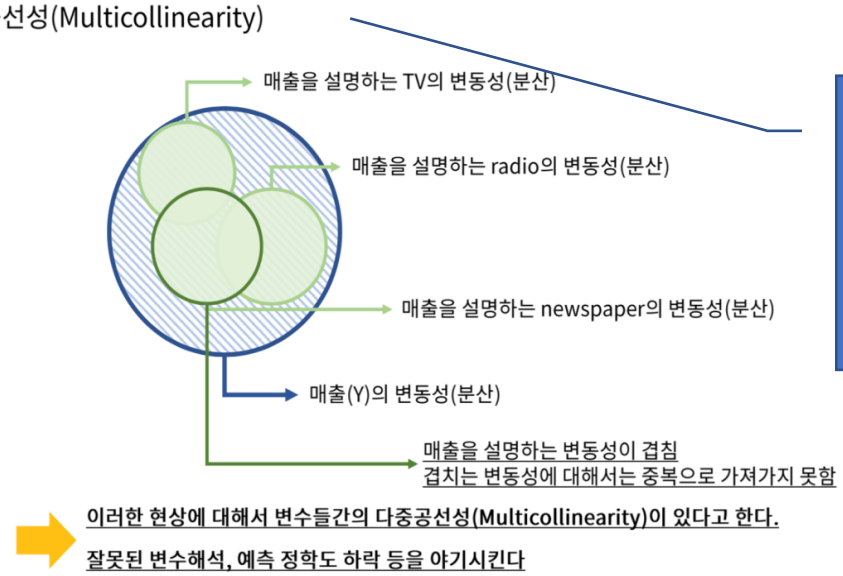
다중선형회귀 모델에서 종속변수에 영향을 주는 설명 변수를 선택하는 방법 - 전진선택법, 후진제거법, 단계적 방법

```
step(model, direction="forward|backward|both")
```

고려하는 독립(설명)변수 모두를 회귀 모형에 포함하는 경우 독립변수들 중 일부만을 포함하는 회귀모형에 비해서 결정계수의 값이 항상 크므로 설명력을 최대화시킬 수 있는 반면에 독립변수들간의 상관관계가 커져서 생기는 다중공선성의 문제에 직면하는 경우가 많고, 따라서 모형의 안정성과 신뢰성에 의문의 생길 수 있다.

(다중공선성 : 다중 선형 회귀에서 독립변수들간에 강한 선형관계가 있을 때)

- 다중공선성(Multicollinearity)



이러한 현상에 대해서 변수들간의 다중공선성(Multicollinearity)이 있다고 한다.

잘못된 변수해석, 예측 정확도 하락 등을 야기시킨다

의심이 가는 독립변수들만을 가지고 회귀분석을 한 다음 vif() 함수로 분산팽창 계수를 확인한다. 10이 넘는 변수는 다중공선성이 존재한다고 간주한다.

```
tadata <- read.csv("data/TAccident.csv")
start.lm <- lm(Y~1, data=tadata)
full.lm <- lm(Y~., data=tadata)
```

(1) 전진선택법

```
Y ~ 1
Y ~ X9
Y ~ X9 + X1
Y ~ X9 + X1 + X4
Y ~ X9 + X1 + X4 + X8
Y ~ X9 + X1 + X4 + X8 + X12
```

(2) 후진 제거법

```
Y ~ X1 + X2 + X3 + X4 + X5 + X6 + X7 + X8 + X9 + X10 + X11 + X12 + X13
Y ~ X1 + X3 + X4 + X5 + X6 + X7 + X8 + X9 + X10 + X11 + X12 + X13
Y ~ X1 + X3 + X4 + X5 + X6 + X7 + X8 + X9 + X11 + X12 + X13
Y ~ X1 + X3 + X4 + X5 + X6 + X7 + X8 + X9 + X12 + X13
Y ~ X1 + X3 + X4 + X6 + X7 + X8 + X9 + X12 + X13
Y ~ X1 + X3 + X4 + X7 + X8 + X9 + X12 + X13
Y ~ X1 + X3 + X4 + X8 + X9 + X12 + X13
Y ~ X1 + X3 + X4 + X8 + X9 + X12
Y ~ X1 + X4 + X8 + X9 + X12
```

```
model <- lm(mpg ~ . , data = mtcars)
```

```
new_model <- step(model, direction = "both")
```

```
Start:  AIC=70.9
mpg ~ cyl + disp + hp + drat + wt + qsec + vs + am + gear + carb

       Df Sum of Sq     RSS     AIC
- cyl   1    0.0799 147.57 68.915
- vs    1    0.1601 147.66 68.932
- carb  1    0.4067 147.90 68.986
- gear  1    1.3531 148.85 69.190
- drat  1    1.6270 149.12 69.249
- disp  1    3.9167 151.41 69.736
- hp    1    6.8399 154.33 70.348
- qsec  1    8.8641 156.36 70.765
<none>                  147.49 70.898
- am    1   10.5467 158.04 71.108
- wt    1   27.0144 174.51 74.280

Step:  AIC=68.92
mpg ~ disp + hp + drat + wt + qsec + vs + am + gear + carb

       Df Sum of Sq     RSS     AIC
- vs    1    0.2685 147.84 66.973
- carb  1    0.5201 148.09 67.028
- gear  1    1.8211 149.40 67.308
- drat  1    1.9826 149.56 67.342
- disp  1    3.9009 151.47 67.750
- hp    1    7.3632 154.94 68.473
<none>                  147.57 68.915
- qsec  1   10.0933 157.67 69.032
- am    1   11.8359 159.41 69.384
+ cyl   1    0.0799 147.49 70.898
- wt    1   27.0280 174.60 72.297

Step:  AIC=66.97
mpg ~ disp + hp + drat + wt + qsec + am + gear + carb

       Df Sum of Sq     RSS     AIC
```

```
- carb   1      0.6855 148.53 65.121
- gear   1      2.1437 149.99 65.434
- drat   1      2.2139 150.06 65.449
- disp   1      3.6467 151.49 65.753
- hp     1      7.1060 154.95 66.475
<none>                    147.84 66.973
- am     1     11.5694 159.41 67.384
- qsec   1     15.6830 163.53 68.200
+ vs     1      0.2685 147.57 68.915
+ cyl    1      0.1883 147.66 68.932
- wt     1     27.3799 175.22 70.410

Step:  AIC=65.12
mpg ~ disp + hp + drat + wt + qsec + am + gear

       Df Sum of Sq      RSS      AIC
- gear   1       1.565 150.09 63.457
- drat   1       1.932 150.46 63.535
<none>                    148.53 65.121
- disp   1      10.110 158.64 65.229
- am     1      12.323 160.85 65.672
- hp     1      14.826 163.35 66.166
+ carb   1       0.685 147.84 66.973
+ vs     1       0.434 148.09 67.028
+ cyl    1       0.414 148.11 67.032
- qsec   1      26.408 174.94 68.358
- wt     1      69.127 217.66 75.350

Step:  AIC=63.46
mpg ~ disp + hp + drat + wt + qsec + am

       Df Sum of Sq      RSS      AIC
- drat   1       3.345 153.44 62.162
- disp   1       8.545 158.64 63.229
<none>                    150.09 63.457
- hp     1      13.285 163.38 64.171
+ gear   1       1.565 148.53 65.121
+ cyl    1       1.003 149.09 65.242
```

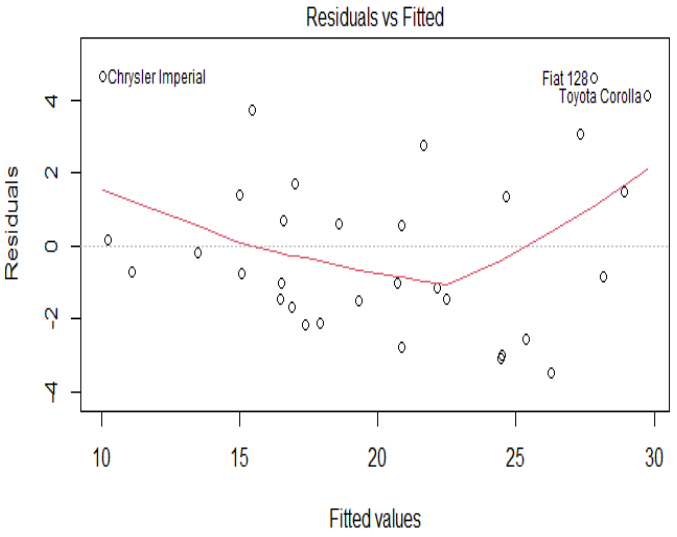
```
+ vs      1       0.645 149.45 65.319
+ carb    1       0.107 149.99 65.434
- am      1      20.036 170.13 65.466
- qsec    1      25.574 175.67 66.491
- wt      1      67.572 217.66 73.351

Step:  AIC=62.16
mpg ~ disp + hp + wt + qsec + am

          Df Sum of Sq      RSS      AIC
- disp    1       6.629 160.07 61.515
<none>                   153.44 62.162
- hp      1      12.572 166.01 62.682
+ drat    1       3.345 150.09 63.457
+ gear    1       2.977 150.46 63.535
+ cyl     1       2.447 150.99 63.648
+ vs      1       1.121 152.32 63.927
+ carb    1       0.011 153.43 64.160
- qsec    1      26.470 179.91 65.255
- am      1      32.198 185.63 66.258
- wt      1      69.043 222.48 72.051

Step:  AIC=61.52
mpg ~ hp + wt + qsec + am

          Df Sum of Sq      RSS      AIC
- hp      1       9.219 169.29 61.307
<none>                   160.07 61.515
+ disp    1       6.629 153.44 62.162
+ carb    1       3.227 156.84 62.864
+ drat    1       1.428 158.64 63.229
- qsec    1      20.225 180.29 63.323
+ cyl     1       0.249 159.82 63.465
+ vs      1       0.249 159.82 63.466
+ gear    1       0.171 159.90 63.481
- am      1      25.993 186.06 64.331
- wt      1      78.494 238.56 72.284
Step:  AIC=61.31
```

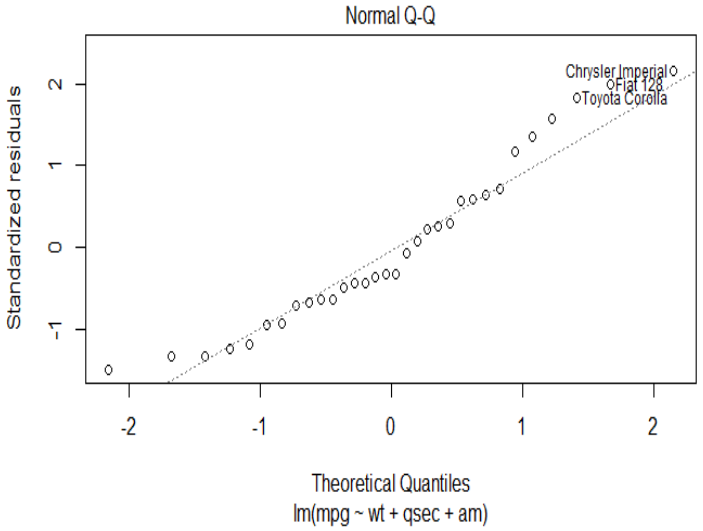
mpg ~ wt + qsec + am

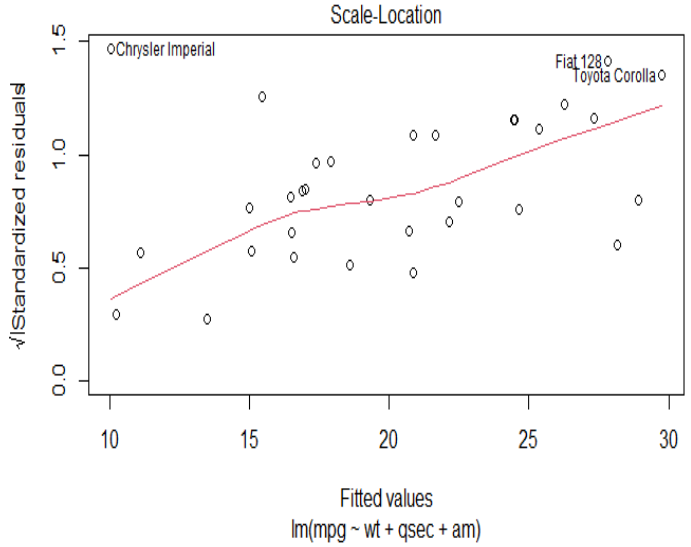
	Df	Sum of Sq	RSS	AIC
<none>			169.29	61.307
+ hp	1	9.219	160.07	61.515
+ carb	1	8.036	161.25	61.751
+ disp	1	3.276	166.01	62.682
+ cyl	1	1.501	167.78	63.022
+ drat	1	1.400	167.89	63.042
+ gear	1	0.123	169.16	63.284
+ vs	1	0.000	169.29	63.307
- am	1	26.178	195.46	63.908
- qsec	1	109.034	278.32	75.217
- wt	1	183.347	352.63	82.790

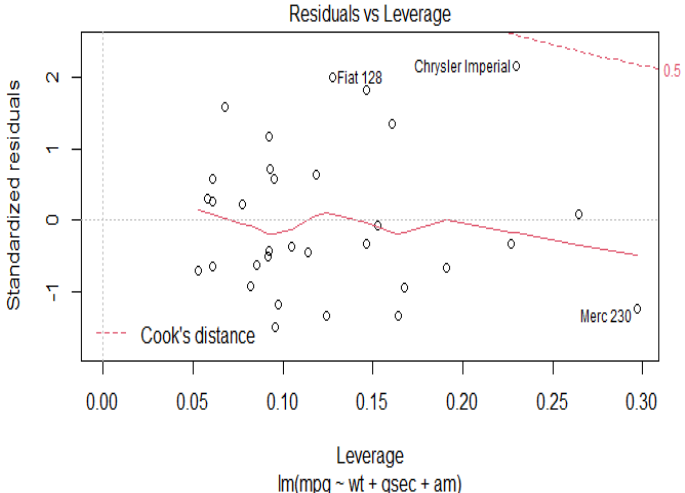
- 모델 진단 그래프

선형회귀 모델의 평가를 여러 그래프로 시각화할 수 있다.







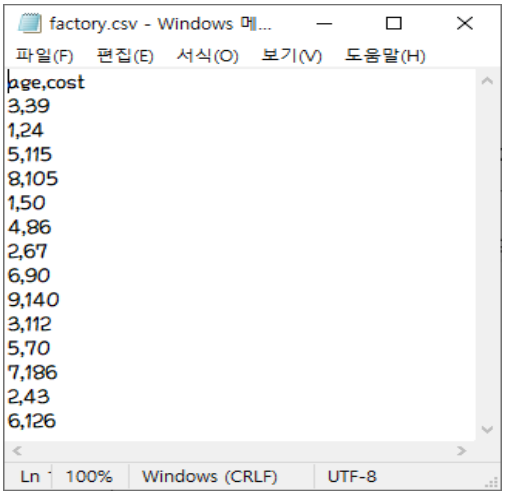


[회귀분석 예(1)]

큰 공장에서 동일한 기계들의 정비기록에 관한 표본자료

fdata <- read.csv("model/factory.csv")

attach(fdata)

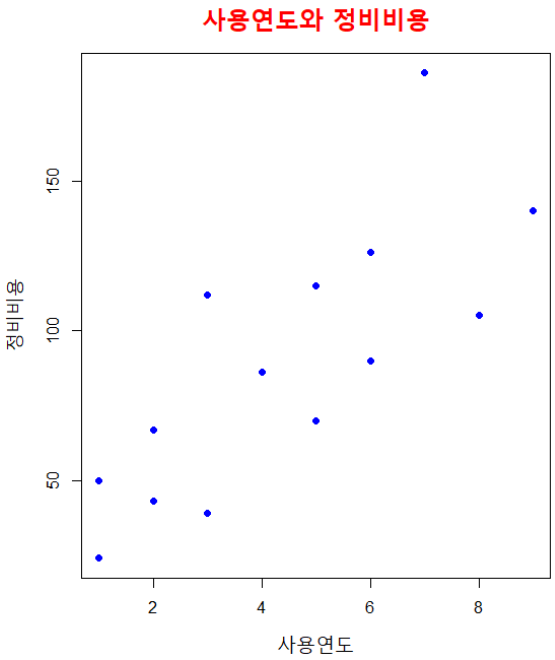


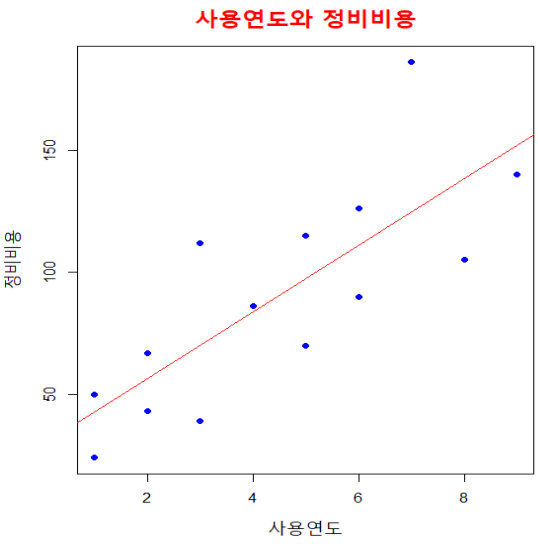
(1) 이 데이터의 산점도를 그려라.

plot(age, cost, xlab="사용연도", ylab="정비비용", pch=19,

col="blue", cex.lab=1.5)

title("사용연도와 정비비용", cex.main=2, col.main="red")





(2) 최소제곱법에 의한 회귀직선을 적합시켜라.

factory.lm <- lm(cost ~ age, data=fdata)

abline(factory.lm, col="red")

(3) 추정치의 표준오차를 구하라. : 29.11

summary(factory.lm)

```
Call:
lm(formula = cost ~ age, data = fdata)

Residuals:
    Min      1Q  Median      3Q     Max
-33.204 -20.383  -4.748  13.957  61.433

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)   29.107     15.969   1.823 0.093341 .
age           13.637      3.149   4.330 0.000978 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 29.11 on 12 degrees of freedom
Multiple R-squared:  0.6098,    Adjusted R-squared:  0.5773
F-statistic: 18.75 on 1 and 12 DF,  p-value: 0.0009779
```

추정치의 표준오차

(4) 결정계수와 상관계수를 구한다.

결정계수 : 0.6098

상관계수 : 기울기가 양의 값(13.637)이므로 양의 상관관계를 갖는다. -> sqrt(0.6098) → 0.7808969

```
Call:
lm(formula = cost ~ age, data = fdata)

Residuals:
    Min      1Q  Median      3Q     Max
-33.204 -20.383  -4.748  13.957  61.433

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)   29.107     15.969   1.823 0.093341 .
age           13.637      3.149   4.330 0.000978 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 29.11 on 12 degrees of freedom
Multiple R-squared:  0.6098,    Adjusted R-squared:  0.5773
F-statistic: 18.75 on 1 and 12 DF,  p-value: 0.0009779
```

결정계수

R 분석 결과에서는 검정통계량 $\mathbf{F_0}$에 대한 유의확률 p-값이 제공된다. 0.0009779 <0.005 이므로 귀무가설을 기각한다. 따라서 구해진 회귀직선은 유의미하다.

(6) 사용연도가 4년인 기계의 평균정비비용은 어느 정도인가를 추정한다.

$\hat{Y}$ = 13.637 * X + 29.107

```
> 13.637 * 4 + 29.107
[1] 83.655
```

사용연도가 4년인 기계의 평균정비비용은 83.655 이다.

R 함수로 구하면 다음과 같다.

```
> predict(factory.lm, newdata=data.frame(age=4) )
       1
83.65552
```

(7) 잔차를 구하여 잔차의 합이 0임을 확인한다.

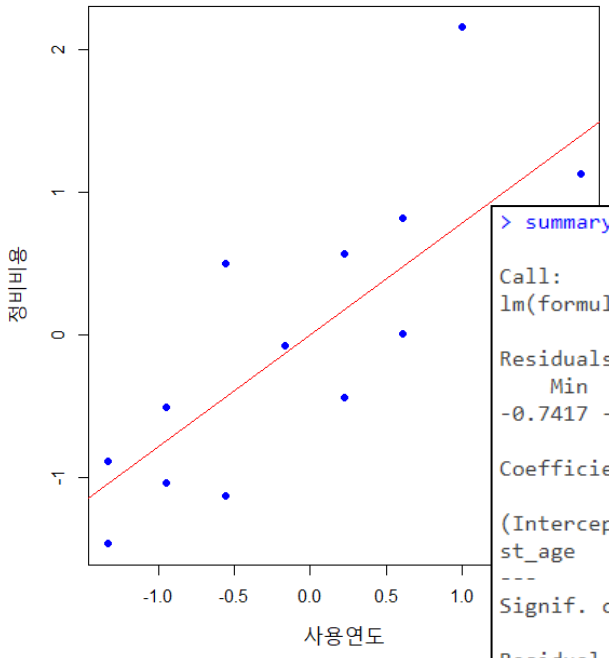
```
> sum(factory.lm$residuals)
[1] 0
```

(8) 두 변수 x와 y를 표준화된 변수로 고친 후 회귀직선을 적합시키고, 그 회귀계수가 두 변수 x, y 간의 상관계수와 같음을 밝힌다.

표준화(standardization) : 원래의 변수측정치를 평균이 0이고 분산이 1인 새로운 변수로 변화시키는 과정을 의미한다.
변수에서 그 변수의 평균을 차감한 값을 그 변수의 표준편차로 나누면 된다.

```
st_fdata <- cbind(fdata, st_age=(age-mean(age))/sd(age), st_cost=(cost-mean(cost))/sd(cost))
attach(st_fdata)
st_factory.lm <- lm(st_cost ~ st_age, data=st_fdata)
plot(st_age, st_cost, xlab="사용연도", ylab="정비비용", pch=19, col="blue", cex.lab=1.5)
title("변수 표준화 후의 사용연도와 정비비용", cex.main=2, col.main="red")
abline(st_factory.lm, col="red")
summary(st_factory.lm)
```

변수 표준화 후의 사용연도와 정비비용




```
> summary(st_factory.lm)

Call:
lm(formula = st_cost ~ st_age, data = st_fdata)

Residuals:
    Min      1Q  Median      3Q     Max
-0.7417 -0.4553 -0.1061  0.3118  1.3723

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) 8.251e-17  1.738e-01    0.00 1.000000
st_age      7.809e-01  1.803e-01    4.33 0.000978 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.6502 on 12 degrees of freedom
Multiple R-squared:  0.6098,    Adjusted R-squared:  0.5773
F-statistic: 18.75 on 1 and 12 DF,  p-value: 0.0009779
```

양의 상관관계 동일

결정계수 동일

상관계수 동일

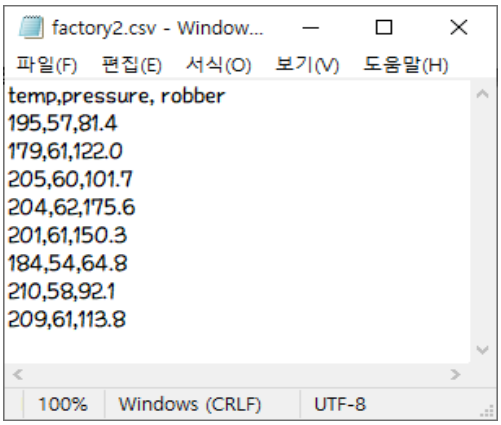
[변수 표준화 전의 회귀 계수들]

결정계수 : 0.6098

상관계수 : 기울기가 양의 값(13.637)이므로 양의 상관관계를 갖는다. -> sqrt(0.6098) -> 0.7808969

[회귀분석 예(2)]

2번 어떤 공장에서 나오는 제품의 강도가 그 공정의 온도와 압력에 어떤 영향을 받는가를 조사하기 위하여 얻은 데이터



(1) 회귀모형 추정

$\hat{Y}$=-554.527-0.174X1+11.845X2

```
> fdata2.lm <- lm(robber~temp+pressure, data=fdata2)
> summary(fdata2.lm)

Call:
lm(formula = robber ~ temp + pressure, data = fdata2)

Residuals:
      1       2       3       4       5       6       7       8
 -5.250 -14.817 -18.742  31.294  17.316  11.768  -3.781 -17.789

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) -554.5267   197.2264  -2.812   0.0375 *
temp          -0.1743     0.7636  -0.228   0.8285
pressure      11.8449     3.2342   3.662   0.0146 *
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 21.66 on 5 degrees of freedom
Multiple R-squared:  0.747,     Adjusted R-squared:  0.6459
F-statistic: 7.383 on 2 and 5 DF,  p-value: 0.03218
```

[회귀분석 예(3)]

```
> fdata3 <- read.csv("model/factory3.csv")
> fdata3.lm <- lm(Y~X1+X2+X3, data=fdata3)
> summary(fdata3.lm)

Call:
lm(formula = Y ~ X1 + X2 + X3, data = fdata3)

Residuals:
     Min       1Q   Median       3Q      Max
-0.23490 -0.07744 -0.02166  0.08840  0.23442

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)  2.409213   1.125954   2.140  0.07618 .
X1           0.069788   0.012640   5.521  0.00149 **
X2          -0.024767   0.044830  -0.552  0.60060
X3           0.005864   0.005052   1.161  0.28978
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.172 on 6 degrees of freedom
Multiple R-squared:  0.9202,    Adjusted R-squared:  0.8803
F-statistic: 23.05 on 3 and 6 DF,  p-value: 0.001079
```

회귀방정식

$\hat{Y} = 2.409213+0.069788X1-0.024767X2+0.005864X3$

X1 = 20, X2=27, X3=60에서의 평균 물 소비량 추정

```
> 2.409213+0.069788*20-0.024767*27+0.005864*60
[1] 3.488104
```

[변동 계수]

두 개 이상의 데이터에 대하여 퍼짐 정도를 비교하기 위해서 두 데이터의 표준편차를 구하여 비교하는 것은 측정단위가 서로 다르거나 데이터 값의 차이가 커서 무의미한 경우가 많다. 이러한 경우에 사용하는 측도가 표준 편차를 평균으로 나눈 변이계수(변동계수:coefficient of variation)를 사용한다. 변동 계수는 표준 편차를 산술 평균으로 나눈 것이다. 상대 표준 편차라고도 한다. 측정단위가 서로 다른 자료를 비교하고자 할 때 쓰인다. 즉, 범위나 분산과 같은 산포도를 계산하는 것만으로는 충분하지 않아 상대적인 산포도를 비교해야 한다.

어떠한 백분율 값을 측정한 것으로 보이는 두 그룹이 있다.

group1 <- c(86, 85, 92, 89, 83, 90, 88, 91, 79, 83)

group2 <- c(0.88, 0.91, 0.94, 0.84, 0.97, 0.89, 0.99, 0.88, 0.89, 0.96)

group1 은 100을 곱한 백분율의 상태 group2 는 0~1 범위의 백분율 상태이다. 실제로 위와 같은 예는 조사자의 취향에 따라 단위가 통일되지 못하는 사례로 많다. 당연히 이 둘의 데이터 산포도를 측정하고자 표준편차를 구하게 될 때 group1 의 표준편차가 작을 것이다.

```
sd(group1)
```

```
## [1] 4.141927
```

```
sd(group2)
```

```
## [1] 0.04790036
```

산포도의 공평한 비교가 될 수 없으므로 group2 자료에 100을 곱한 후 표준편차를 다시 구하거나 반대로 group1 자료에 100을 나누어 표준편차를 구하여 처리할 수도 있겠지만 이것은 원 데이터를 보존하지 못하는 방법이므로 상황에 따라 위험성이 있을 것이다. 바로 이런 경우 변동계수를 이용하는 것이 좋다.

```
sd(group1) / mean(group1)
```

```
## [1] 0.04782825
```

```
sd(group2) / mean(group2)
```

```
## [1] 0.05235012
```

- 회귀모델의 체크사항

최적의 모델을 찾은 후, 모델이 적합한지 다음 사항을 체크한다.

(1) 해당 회귀모델이 통계적으로 유의미한가?

(2) 해당 회귀 모델의 회귀계수가 유의미한가?

(3) 해당 회귀모델이 얼마나 설명력을 갖는가?

(4) 모델이 데이터를 잘 적합하고 있는가?