2020학년도 1학기 과제물

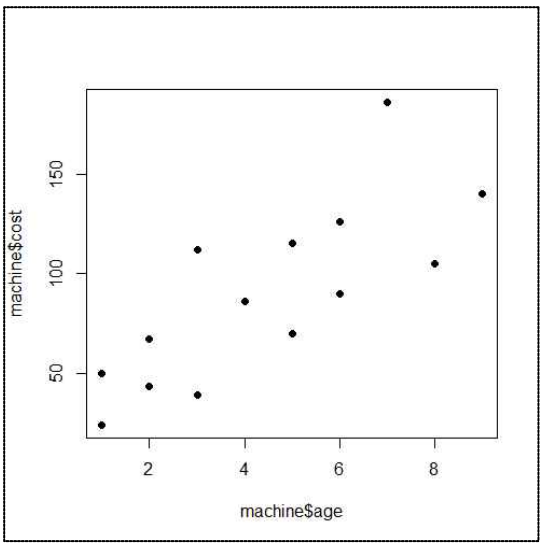
교과목명 : 회귀모형
학　　번 : 202035-368086
성　　명 : 김동현
연 락 처 : 010-9555-9834, barkle@knou.ac.kr

○ 과 제 명 : 회귀모형 출석수업대체과제물

[1장 : 1번]

(1) 이 데이터의 산점도를 그려라.

```
machine = read.table("machine.txt", header=T)
plot(machine$age, machine$cost, pch=19)
```



(2) 최소제곱법에 의한 회귀직선을 적합시켜라.

```
machine.lm = lm(cost ~ age, data=machine)
```

```
> machine.lm

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

Coefficients:
(Intercept)          age
      29.11        13.64
```

(3) 추정치의 표준오차 $S_{y \cdot x}$를 구하라

$S_{y \cdot x}$ = 29.11

(4) 결정계수와 상관계수를 구하라

결정계수 = 0.6098

상관계수 = sqrt(0.6098) = 0.7809

```
> summary(machine.lm)

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

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
```

(5) 분산분석표를 작성하고 회귀직선의 유의여부를 검정하라(유의수준 α=0.05)

```
> anova(machine.lm)
Analysis of Variance Table

Response: cost
          Df Sum Sq Mean Sq F value    Pr(>F)
age        1  15887 15887.2  18.753 0.0009779 ***
Residuals 12  10166   847.2
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

p값이 0.0009779로 0.05보다 작으므로 회귀직선은 유의하다.

(6) 사용연도가 4년인 기계의 평균정비비용은 어느 정도인가를 추정하라

Y = 29.107 + 13.637X

X=4 이면 Y=83.655

(7) 잔차 $e_i = y_i - \hat{y_i}$ 를 구하여 잔차의 합이 영임을 확인하라.

```
> sum(machine.lm$resid)
[1] 0
```

(8) 잔차들의 x_i에 대한 가중합, $\sum x_i e_i$를 구하라.

```
> sum(machine$age * machine.lm$resid)
[1] -9.769963e-15
```

(9) 잔차들의 $\hat{y_i}$에 대한 가중합, $\sum y_i e_i$를 구하라.

```
> sum((29.107 + 13.637*machine$age) * machine.lm$resid)
[1] 4.547474e-13
```

(10) 두 변수 x, y를 표준화된 변수로 고친 후 회귀직선을 적합시키고, 그 회귀계수가 두 변수 x, y간의 상관계수와 같음을 밝혀라.

```
> age_std = (machine[1] - mean(unlist(machine[1]))) / sd(unlist(machine[1]))
> cost_std = (machine[2] - mean(unlist(machine[2]))) / sd(unlist(machine[2]))
> machine2 <- cbind( machine, age_std, cost_std )
> names(machine2)[3] <- c("age_std")
> names(machine2)[4] <- c("cost_std")
> machine2.lm = lm(cost_std ~ age_std , data=machine2)
> summary(machine2.lm)

Call:
lm(formula = cost_std ~ age_std, data = machine2)

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
age_std     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
```

표준화된 x, y의 회귀계수는 0.7809이고, x, y의 상관계수는 sqrt(0.6098) = 0.7809 이다.

[1장 : 5번]

(1) 앞의 <연습문제 1번>에 대하여 β_1, β_0, $\mu_{y\cdot x}(x=8)$의 90% 신뢰구간을 구하라.

β_1 의 신뢰계수 100(1-α)% 신뢰구간은

$$\left[b_1 - t(n-2;\alpha/2)\sqrt{\frac{MSE}{S_{XX}}} \sim b_1 + t(n-2;\alpha/2)\sqrt{\frac{MSE}{S_{XX}}} \right]$$

β_0 의 신뢰계수 100(1-α)% 신뢰구간은

$$\left[b_0 - t(n-2;\alpha/2)\sqrt{MSE\left(\frac{1}{n}+\frac{\overline{X}^2}{S_{XX}}\right)} \sim b_0 + t(n-2;\alpha/2)\sqrt{MSE\left(\frac{1}{n}+\frac{\overline{X}^2}{S_{XX}}\right)} \right]$$

$\mu_{y\cdot x}$ 의 신뢰계수 100(1-α)% 신뢰구간은

$$\left[\hat{Y}_{new} - t(n-2;\alpha/2)\sqrt{MSE\left(1+\frac{1}{n}+\frac{(X-\overline{X})^2}{S_{XX}}\right)} \sim \hat{Y}_{new} + t(n-2;\alpha/2)\sqrt{MSE\left(1+\frac{1}{n}+\frac{(X-\overline{X})^2}{S_{XX}}\right)} \right]$$

n=14, t(12,0.5)=1.782, MSE=847.2, S_{XX}=85.429, $\overline{X}$=4.428

이 값을 대입하면

β_1 의 신뢰계수 100(1-α)% 신뢰구간은 [23.495 ~ 34.719]

β_0 의 신뢰계수 100(1-α)% 신뢰구간은 [-14.820 ~ 42.094]

x=8일 때 $\mu_{y\cdot x}$의 신뢰계수 100(1-α)% 신뢰구간은 [80.895 ~ 195.511]

```
> n=14
> qt(0.95, n-2)
[1] 1.782288
> MSE=847.2
> S_XX= sum( (machine$age - mean(machine$age))^2)
> bar_X = mean(machine$age)
> beta1_min = 29.107 - 1.782 * sqrt(MSE/S_XX)
> beta1_max = 29.107 + 1.782 * sqrt(MSE/S_XX)
> beta1_min; beta1_max
[1] 23.49524
[1] 34.71876
> beta0_min = 13.637 - 1.782 * sqrt(MSE*((1/n) + (bar_X^2)/S_XX))
> beta0_max = 13.637 + 1.782 * sqrt(MSE*((1/n) + (bar_X^2)/S_XX))
> beta0_min; beta0_max
[1] -14.81982
[1] 42.09382
> hat_Y_min = 138.203 - 1.782 * sqrt(MSE * (1 + 1/n + (8-bar_X)^2 / S_XX))
> hat_Y_max = 138.203 + 1.782 * sqrt(MSE * (1 + 1/n + (8-bar_X)^2 / S_XX))
> hat_Y_min; hat_Y_max
[1] 80.89549
[1] 195.5105
```

(2) 또한 다음의 가설검정을 α=0.01에서 실시하라.

H_0: $\beta_1=10$ / H_1: $\beta_1 \neq 10$

검정통계량 $t_0 = \dfrac{b_1 - \beta_{10}}{\sqrt{\left(\dfrac{MSE}{S_{XX}}\right)}}$ = 1.155 이고, t(12, 0.005) = 3.054 이다

$|t_0| > t(12,0.05)$ 즉, 1.155 > 3.054 는 참이 아니므로 귀무가설을 기각할 수 없다.

```
> t_0 = (13.637 - 10) / sqrt(MSE/S_XX)
> t = qt(0.005, n-2, lower.tail=FALSE)
> t_0; t
[1] 1.15492
[1] 3.05454
```

[2장 : 2번]

(1) 데이터로부터 회귀모형을 추정하라.

Y = −554.5267 − 0.1743 × X_1 + 11.8449 × X_2 + ε

```
> factory = read.table("factory.txt", header=T)
> factory.lm = lm(str ~ temp + press, data=factory)
> summary(factory.lm)

Call:
lm(formula = str ~ temp + press, data = factory)

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
press         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
```

(2) 오차분산 σ^2을 MSE로 추정하고, $Var(b_0)$, $Var(b_1)$, $Var(b_2)$의 추정치를 구하라

오차분산 σ^2 의 추정치 MSE는 469.4 이다.

```
> anova(factory.lm)
Analysis of Variance Table

Response: str
          Df Sum Sq Mean Sq F value  Pr(>F)
temp       1  634.9   634.9  1.3526 0.29730
press      1 6295.7  6295.7 13.4133 0.01456 *
Residuals  5 2346.8   469.4
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

$Var(b_0)$, $Var(b_1)$, $Var(b_2)$의 추정치는 $(X'X)^{-1}$ 행렬의 구성원을 c_{ij}라 하면 $c_{ii}\sigma^2$ 이 된다

$Var(b_0)$ = 82.87499 × 469.4 = 38901.52

$Var(b_1)$ = 0.001242 × 469.4 = 0.5831197

$Var(b_2)$ = 0.022285 × 469.4 = 10.46077

```
> X = factory[, c(1:2)]
> X = cbind(1, X)
> X = as.matrix(X)
> XTX = t(X) %*% X
> XTXI = solve(XTX)
> 469.4*XTXI[1,1];469.4*XTXI[2,2];469.4*XTXI[3,3];
[1] 38901.52
[1] 0.5831197
[1] 10.46077
```

(3) X1 = 200, X2 = 59인 경우 Y의 추정치는 얼마인가? 이 Y의 분산을 추정하라.

회귀직선이 $Y = -554.5267 - 0.1743 \times X_1 + 11.8449 \times X_2$ 이므로

추정치 Y = 109.4624

추정치 Y의 분산 $Var(\hat{Y}) = x'(X'X)^{-1}x\sigma^2$

x' = (1 200 59) 을 대입하면 $Var(\hat{y})$ = 0.1312068 × 469.4 = 61.58847

(4) 추정된 회귀계수 b_1, b_2의 의미

회귀계수는 각각의 독립변수와 반응변수의 관계를 나타낸다. b_1=-0.1743 이라는 것은 온도가 1도 상승할 때 제품의 강도가 0.1743만큼 약해지는 관계라는 것을 의미하며, b_2=11.8449 인 것은 압력이 1psi 상승할 때 제품의 강도가 11.8449만큼 강해진다는 의미이다.

(5) 분산분석표를 작성하고, α=0.05로 F-검정을 행하라

요인	자유도	제곱합	평균제곱	F_0
회귀	2	6930.6	3465.3	7.38
잔차	5	2346.8	469.4	
계	7	9277.4		

F값=7.38에 대한 유의확률은 0.0322로 0.05보다 작으므로 중회귀모형이 유의하다고 할 수 있다.

(6) 결정계수 R^2을 구하라

결정계수 R^2 = 0.747이다.

(7) X_1, X_2, Y를 모두 표준화시키고, 표준화된 중회귀방정식을 구하라.

Y = −0.05499 X_1 + 0.8825 X_2 + ε

(절편 값은 0으로 봄)

```
> temp_std = (factory[1] - mean(unlist(factory[1]))) / sd(unlist(factory[1]))
> press_std = (factory[2] - mean(unlist(factory[2]))) / sd(unlist(factory[2]))
> str_std = (factory[3] - mean(unlist(factory[3]))) / sd(unlist(factory[3]))
> factory2 <- cbind(temp_std, press_std, str_std)
> factory2.lm = lm(str ~ temp + press, data=factory2)
> summary(factory2.lm)

Call:
lm(formula = str ~ temp + press, data = factory2)

Residuals:
      1       2       3       4       5       6       7       8
-0.1442 -0.4070 -0.5148  0.8596  0.4757  0.3233 -0.1038 -0.4887

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) -1.202e-16  2.104e-01   0.000   1.0000
temp        -5.499e-02  2.410e-01  -0.228   0.8285
press        8.825e-01  2.410e-01   3.662   0.0146 *
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.5951 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
```

[2장 : 3번]

(1) 회귀방정식을 구하라

$\hat{Y} = 2.409 + 0.0698X_1 - 0.0248X_2 + 0.0059X_3$

```
> water = read.table("water.txt", header=T)
> water.lm = lm(water ~ temp+day+ton, data=water)
> summary(water.lm)

Call:
lm(formula = water ~ temp + day + ton, data = water)

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 .
temp         0.069788   0.012640   5.521  0.00149 **
day         -0.024767   0.044830  -0.552  0.60060
ton          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
```

(2) b_1, b_2, b_3의 의미는 무엇인가?

평균온도와 작업일수, 작업량이 물소비량에 미치는 영향을 나타낸 값이다. 물 소비량은 온도와 작업량과는 양의 상관관계가 있고, 작업일과는 음의 상관관계가 있다.

(3) 분산분석표를 작성하고, 결정계수 R^2를 구하라

```
> anova(water.lm)
Analysis of Variance Table

Response: water
          Df  Sum Sq Mean Sq F value    Pr(>F)
temp       1 2.00432 2.00432 67.7386 0.0001738 ***
day        1 0.00227 0.00227  0.0768 0.7909535
ton        1 0.03988 0.03988  1.3477 0.2897756
Residuals  6 0.17753 0.02959
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

요인	자유도	제곱합	평균제곱	F_0
회귀	3	2.04647	0.6822	2.306
잔차	6	0.17753	0.2959	
계	9	2.224		

결정계수 R^2는 0.9202 이다.

(4) X_1=20, X_2=27, X_3=60에서 평균 물소비량을 추정하라

$Y = 2.409213 + 0.069788*20 - 0.024767*27 + 0.005864*60 = 3.488104$

[3장 : 3번]

```
library(leaps)
TA = read.csv("p116.csv", header=T)
# 앞으로부터 선택법
start.lm = lm(Y ~ 1, data=TA)
full.lm = lm(Y ~. , data=TA)
step(start.lm, scope=list(lower=start.lm, upper=full.lm), direction="forward")
# 뒤로부터 제거법
step(full.lm, data=TA, direction="backward")
# 단계별 회귀방법
step(start.lm, scope=list(upper=full.lm), data=TA, direction="both")
# 모든 가능한 회귀방법
TA.lm = regsubsets(Y ~. , data=TA)
summary(TA.lm)
```

(1) 앞으로부터 선택법

$Y = 9.944 + 0.0643X_9 - 0.0741X_1 - 0.1051X_4 + 0.7974X_8 - 0.7744X_{12}$

```
Call:
lm(formula = Y ~ X9 + X1 + X4 + X8 + X12, data = TA)

Coefficients:
(Intercept)           X9           X1           X4           X8          X12
    9.94408      0.06428     -0.07405     -0.10510      0.79736     -0.77443
```

(2) 뒤로부터 제거법

$Y = 9.944 - 0.0741X_1 - 0.1051X_4 + 0.7974X_8 + 0.0643X_9 - 0.7744X_{12}$

```
Call:
lm(formula = Y ~ X1 + X4 + X8 + X9 + X12, data = TA)

Coefficients:
(Intercept)           X1           X4           X8           X9          X12
    9.94408      -0.07405     -0.10510      0.79736      0.06428     -0.77443
```

(3) 단계별 회귀방법

$Y = 9.944 + 0.0643X_9 - 0.0741X_1 - 0.1051X_4 + 0.7974X_8 - 0.7744X_{12}$

```
Call:
lm(formula = Y ~ X9 + X1 + X4 + X8 + X12, data = TA)

Coefficients:
(Intercept)           X9           X1           X4           X8          X12
    9.94408      0.06428     -0.07405     -0.10510      0.79736     -0.77443
```

(4) 모든 회귀방법

```
> TA.lm = regsubsets(Y ~. , data=TA)
> summary(TA.lm)
Subset selection object
Call: regsubsets.formula(Y ~ ., data = TA)
13 Variables  (and intercept)
    Forced in Forced out
X1      FALSE      FALSE
X2      FALSE      FALSE
X3      FALSE      FALSE
X4      FALSE      FALSE
X5      FALSE      FALSE
X6      FALSE      FALSE
X7      FALSE      FALSE
X8      FALSE      FALSE
X9      FALSE      FALSE
X10     FALSE      FALSE
X11     FALSE      FALSE
X12     FALSE      FALSE
X13     FALSE      FALSE
1 subsets of each size up to 8
Selection Algorithm: exhaustive
         X1  X2  X3  X4  X5  X6  X7  X8  X9  X10 X11 X12 X13
1  ( 1 ) " " " " " " " " " " " " " " " " "*" " " " " " " " "
2  ( 1 ) "*" " " " " " " " " " " " " " " " " " " " " " " " "
3  ( 1 ) "*" " " " " " "*" " " " " " " " " " " " " " " " " "
4  ( 1 ) "*" " " " " " "*" " " " " " " " "*" "*" " " " " " "
5  ( 1 ) "*" " " " " " "*" " " " " " " " "*" "*" " " " " "*" " "
6  ( 1 ) "*" " " "*" "*" " " " " " " " " "*" "*" " " " " "*" " "
7  ( 1 ) "*" " " "*" "*" " " " " " " " " "*" "*" " " " " "*" "*"
8  ( 1 ) "*" " " "*" "*" " " " " " " "*" "*" "*" " " " " "*" "*"
```

(5) 비교

모두 동일한 변수를 선택한 것으로 나타남

(모든 가능한 회귀방법에서는 수정결정계수를 기준으로 선택)

선택방법	선택한 변수
앞으로부터 선택법	X1, X4, X8, X9, X12
뒤로부터 제거법	X1, X4, X8, X9, X12
단계별 회귀방법	X1, X4, X8, X9, X12
모든 가능한 회귀방법	X1, X4, X8, X9, X12

[3장 : 4번]

```
car = read.csv("p118.csv", header=T)
# 앞으로부터 선택법
start.lm = lm(Y ~ 1, data=car)
full.lm = lm(Y ~. , data=car)
step(start.lm, scope=list(lower=start.lm, upper=full.lm), direction="forward")
# 뒤로부터 제거법
step(full.lm, data=car, direction="backward")
# 단계별 회귀방법
step(start.lm, scope=list(upper=full.lm), data=car, direction="both")
# 모든 가능한 회귀방법
car.lm = regsubsets(Y ~. , data=car)
rs = summary(car.lm)
rs
rs$adjr2
```

(1) 앞으로부터 선택법

$Y = 15.904 + -3.153X_5 - 0.400X_1 + 4.001X_8 + 0.875X_6$

```
Call:
lm(formula = Y ~ X5 + X1 + X8 + X6, data = car)

Coefficients:
(Intercept)           X5           X1           X8           X6
    15.9042      -3.1534      -0.3999       4.0007       0.8748
```

(2) 뒤로부터 제거법

$Y = 3.432 + 3.931X_4 + 2.479X_8 + 2.410X_9 - 2.536X_{10}$

```
Call:
lm(formula = Y ~ X4 + X8 + X9 + X10, data = car)

Coefficients:
(Intercept)           X4           X8           X9          X10
      3.432        3.931        2.479        2.410       -2.536
```

(3) 단계별 회귀방법

$Y = 10.167 - 3.459X_5 + 4.527X_8 + 1.101X_6$

```
Call:
lm(formula = Y ~ X5 + X8 + X6, data = car)

Coefficients:
(Intercept)           X5           X8           X6
     10.167       -3.459        4.527        1.101
```

(4) 모든 회귀방법

```
> car.lm = regsubsets(Y ~. , data=car)
> rs = summary(car.lm)
> rs
Subset selection object
Call: regsubsets.formula(Y ~ ., data = car)
10 variables  (and intercept)
    Forced in Forced out
X1      FALSE      FALSE
X2      FALSE      FALSE
X3      FALSE      FALSE
X4      FALSE      FALSE
X5      FALSE      FALSE
X6      FALSE      FALSE
X7      FALSE      FALSE
X8      FALSE      FALSE
X9      FALSE      FALSE
X10     FALSE      FALSE
1 subsets of each size up to 8
Selection Algorithm: exhaustive
         X1  X2  X3  X4  X5  X6  X7  X8  X9  X10
1  ( 1 ) " " " " " " " " "*" " " " " " " " " "
2  ( 1 ) "*" " " " " " " " "*" " " " " " " " " "
3  ( 1 ) " " " " " " " " "*" " " " " " " "*" "*"
4  ( 1 ) " " " " " " " " "*" " " " " " " "*" "*"
5  ( 1 ) " " " " " " " " "*" " " " " "*" "*" "*"
6  ( 1 ) " " " " " " " " "*" "*" "*" " " "*" "*"
7  ( 1 ) "*" " " " " " " "*" "*" "*" " " "*" "*"
8  ( 1 ) "*" " " " " " " "*" "*" "*" "*" "*" "*"
> rs$adjr2
[1] 0.7746166 0.8314518 0.8531437 0.8614018 0.8634642 0.8634920 0.8610993 0.8602488
```

(5) 비교

선택방법에 따라 모두 다른 변수를 선택하게 되었음

(모든 가능한 회귀방법에서는 수정결정계수를 기준으로 선택)

선택방법	선택한 변수
앞으로부터 선택법	X1, X5, X6, X8
뒤로부터 제거법	X4, X8, X9, X10
단계별 회귀방법	X5, X6, X8
모든 가능한 회귀방법	X4, X5, X6, X8, X9, X12