데이터 마이닝

2018605059 정수연

1. 일상생활에서 볼 수 있는 Odds 사례

팀장 선정 룰렛

● 5명에서 조별 과제 시 팀장 선정 룰렛 돌리기

- 내가 팀장으로 선정 될 확률 $= \frac{1}{5}$

- 내가 팀장으로 선정 될 Odds $= \frac{\frac{1}{5}}{1-\frac{1}{5}} = \frac{1}{4}$

2. 선형회귀분석 (득점예측)

- 2.1 **(데이터 수집 및 전처리)**
- 데이터 수집 : kbo.csv (1982~2017 프로야구 팀 타격 기록)

- 변수(15개 이상) : 연도, 팀, 경기, 타석, 타수, 득점, 안타, x2루타, x3루타, 홈런, 총루타, 타점, 도루, 도루실패, 볼넷, 몸에 맞는 공, 고의사구, 삼진, 병살, 희생번트, 희생플라이

```
kbo <- read.csv("C:/RStudio/datamining/kbo.csv",header=TRUE)
kbo
```

	병살	희생번트	희생플라이
1	56	32	27
2	50	36	18
3	35	46	18
4	59	28	21
5	61	41	27
6	44	33	17
7	62	66	27
8	75	37	23
9	58	70	30
10	63	89	18
11	71	58	24
12	76	73	18
13	62	62	18
14	58	56	28
15	66	62	34

	연도	팀	경기	타석	타수	득점	안타	x2루타	x3루타	홈런
1	1982	MBC	80	3061	2686	419	757	124	12	65
2	1982	삼성	80	3043	2647	429	705	126	18	57
3	1982	OB	80	3098	2745	399	778	137	23	57
4	1982	해태	80	2990	2665	374	696	110	14	84
5	1982	롯데	80	3062	2628	353	674	112	8	59
6	1982	삼미	80	2954	2653	302	637	117	20	40
7	1983	삼성	100	3847	3383	448	889	143	14	90
8	1983	해태	100	3734	3340	423	892	130	15	78
9	1983	MBC	100	3715	3273	405	837	145	21	45
10	1983	OB	100	3766	3330	418	863	142	26	50
11	1983	롯데	100	3740	3308	370	807	133	18	78
12	1983	삼미	100	3738	3317	345	814	113	14	62
13	1984	삼성	100	3756	3298	435	889	147	18	78
14	1984	롯데	100	3729	3267	405	840	139	20	71
15	1984	OB	100	3660	3181	382	816	133	18	53

	총루타	타점	도루	도루실패	볼넷	몸에.맞는.공	고의사구	삼진
1	1100	381	134	60	268	47	20	316
2	1038	374	147	42	307	30	2	349
3	1132	362	106	61	247	41	22	254
4	1086	332	155	52	235	41	12	296
5	979	325	83	53	326	40	8	315
6	914	272	74	43	221	29	3	369
7	1330	421	70	42	314	55	9	427
8	1286	388	131	91	294	40	16	382
9	1159	367	128	67	281	60	8	368
10	1207	378	60	45	279	50	11	352
11	1210	342	76	61	308	39	2	423
12	1141	315	35	41	282	47	16	404
13	1306	413	75	61	313	63	18	422
14	1232	374	108	65	326	48	17	446
15	1144	362	111	62	339	43	17	367

2. 선형회귀분석 (득점예측)

- 2.1 **(데이터 수집 및 전처리)**
- 범주형 변수 (팀) 변환처리

```
kbo$팀 <- as.factor(kbo$팀)
str(kbo)
```

```
'data.frame':   283 obs. of  21 variables:
 $ 연도          : int  1982 1982 1982 1982 1982 1982 1983 1983 1983 1983 ...
 $ 팀            : Factor w/ 22 levels "KIA","kt","LG",..: 4 14 6 19 10 12 14 19 4 6 ...
 $ 경기          : int  80 80 80 80 80 80 100 100 100 100 ...
 $ 타석          : int  3061 3043 3098 2990 3062 2954 3847 3734 3715 3766 ...
 $ 타수          : int  2686 2647 2745 2665 2628 2653 3383 3340 3273 3330 ...
 $ 득점          : int  419 429 399 374 353 302 448 423 405 418 ...
 $ 안타          : int  757 705 778 696 674 637 889 892 837 863 ...
 $ X2루타        : int  124 126 137 110 112 117 143 130 145 142 ...
 $ X3루타        : int  12 18 23 14 8 20 14 15 21 26 ...
 $ 홈런          : int  65 57 57 84 59 40 90 78 45 50 ...
 $ 총루타        : int  1100 1038 1132 1086 979 914 1330 1286 1159 1207 ...
 $ 타점          : int  381 374 362 332 325 272 421 388 367 378 ...
 $ 도루          : int  134 147 106 155 83 74 70 131 128 60 ...
 $ 도루실패      : int  60 42 61 52 53 43 42 91 67 45 ...
 $ 볼넷          : int  268 307 247 235 326 221 314 294 281 279 ...
 $ 몸에.맞는.공: int  47 30 41 41 40 29 55 40 60 50 ...
 $ 고의사구      : int  20 2 22 12 8 3 9 16 8 11 ...
 $ 삼진          : int  316 349 254 296 315 369 427 382 368 352 ...
 $ 병살          : int  56 50 35 59 61 44 62 75 58 63 ...
 $ 희생번트      : int  32 36 46 28 41 33 66 37 70 89 ...
 $ 희생플라이    : int  27 18 18 21 27 17 27 23 30 18 ...
```

2. 선형회귀분석 (득점예측)

- 2.1 **(데이터 수집 및 전처리)**
- 데이터확인

summary(kbo)

```
      연도            팀            경기            타석      
 Min.   :1982   롯데   : 36   Min.   : 80.0   Min.   :2954  
 1st Qu.:1992   삼성   : 36   1st Qu.:126.0   1st Qu.:4680  
 Median :2001   LG     : 28   Median :126.0   Median :4895  
 Mean   :2001   한화   : 24   Mean   :126.1   Mean   :4852  
 3rd Qu.:2010   두산   : 19   3rd Qu.:133.0   3rd Qu.:5180  
 Max.   :2017   해태   : 19   Max.   :144.0   Max.   :5863  
                (Other):121                                 
      타수           득점            안타          X2루타      
 Min.   :2628   Min.   :302.0   Min.   : 637   Min.   :110.0  
 1st Qu.:4092   1st Qu.:489.5   1st Qu.:1028   1st Qu.:169.0  
 Median :4260   Median :578.0   Median :1120   Median :192.0  
 Mean   :4229   Mean   :580.3   Mean   :1125   Mean   :193.5  
 3rd Qu.:4486   3rd Qu.:668.0   3rd Qu.:1223   3rd Qu.:216.0  
 Max.   :5142   Max.   :935.0   Max.   :1554   Max.   :304.0  

     X3루타           홈런          총루타      
 Min.   : 3.00   Min.   : 29.0   Min.   : 914  
 1st Qu.:16.00   1st Qu.: 73.5   1st Qu.:1488  
 Median :21.00   Median : 97.0   Median :1638  
 Mean   :21.56   Mean   :103.5   Mean   :1672  
 3rd Qu.:26.00   3rd Qu.:130.5   3rd Qu.:1880  
 Max.   :62.00   Max.   :234.0   Max.   :2465  
```

```
      타점            도루          도루실패      
 Min.   :272.0   Min.   : 35.0   Min.   : 20.00  
 1st Qu.:458.0   1st Qu.: 86.0   1st Qu.: 47.00  
 Median :539.0   Median :107.0   Median : 55.00  
 Mean   :544.1   Mean   :109.5   Mean   : 56.12  
 3rd Qu.:629.5   3rd Qu.:131.0   3rd Qu.: 65.00  
 Max.   :877.0   Max.   :220.0   Max.   :101.00  

      볼넷          몸에.맞는.공      고의사구      
 Min.   :221.0   Min.   : 23.00   Min.   : 2.00  
 1st Qu.:389.0   1st Qu.: 49.50   1st Qu.:13.00  
 Median :446.0   Median : 63.00   Median :17.00  
 Mean   :441.2   Mean   : 64.97   Mean   :17.99  
 3rd Qu.:496.5   3rd Qu.: 78.50   3rd Qu.:22.00  
 Max.   :621.0   Max.   :130.00   Max.   :48.00  

      삼진            병살          희생번트      
 Min.   : 254.0   Min.   : 35.00   Min.   : 21.00  
 1st Qu.: 608.5   1st Qu.: 84.00   1st Qu.: 62.50  
 Median : 748.0   Median : 94.00   Median : 77.00  
 Mean   : 729.4   Mean   : 95.66   Mean   : 79.61  
 3rd Qu.: 865.5   3rd Qu.:107.00   3rd Qu.: 93.50  
 Max.   :1186.0   Max.   :146.00   Max.   :153.00  

   희생플라이    
 Min.   :12.00  
 1st Qu.:29.00  
 Median :35.00  
 Mean   :36.12  
 3rd Qu.:43.00  
 Max.   :68.00  
```

2. 선형회귀분석 (득점예측)

- 2.1 **(데이터 수집 및 전처리)**
- 스케일링&정규화

```
kbo_n<-data.frame(kbo)

경기_min<-min(kbo_n$경기)
경기_max<-max(kbo_n$경기)
kbo_n$경기<-scale(kbo_n$경기,center=경기_min,scale=경기_max-경기_min)

타석_min<-min(kbo_n$타석)
타석_max<-max(kbo_n$타석)
kbo_n$타석<-scale(kbo_n$타석,center=타석_min,scale=타석_max-타석_min)

타수_min<-min(kbo_n$타수)
타수_max<-max(kbo_n$타수)
kbo_n$타수<-scale(kbo_n$타수,center=타수_min,scale=타수_max-타수_min)

안타_min<-min(kbo_n$안타)
안타_max<-max(kbo_n$안타)
kbo_n$안타<-scale(kbo_n$안타,center=안타_min,scale=안타_max-안타_min)

X2루타_min<-min(kbo_n$X2루타)
X2루타_max<-max(kbo_n$X2루타)
kbo_n$X2루타<-scale(kbo_n$X2루타,center=X2루타_min,scale=X2루타_max-X2루타_min)

X3루타_min<-min(kbo_n$X3루타)
X3루타_max<-max(kbo_n$X3루타)
kbo_n$X3루타<-scale(kbo_n$X3루타,center=X3루타_min,scale=X3루타_max-X3루타_min)
```

2. 선형회귀분석 (득점예측)

- 2.1 **(데이터 수집 및 전처리)**
- 스케일링&정규화

```
홈런_min<-min(kbo_n$홈런)
홈런_max<-max(kbo_n$홈런)
kbo_n$홈런<-scale(kbo_n$홈런,center=홈런_min,scale=홈런_max-홈런_min)

총루타_min<-min(kbo_n$총루타)
총루타_max<-max(kbo_n$총루타)
kbo_n$총루타<-scale(kbo_n$총루타,center=총루타_min,scale=총루타_max-총루타_min)

타점_min<-min(kbo_n$타점)
타점_max<-max(kbo_n$타점)
kbo_n$타점<-scale(kbo_n$타점,center=타점_min,scale=타점_max-타점_min)

도루_min<-min(kbo_n$도루)
도루_max<-max(kbo_n$도루)
kbo_n$도루<-scale(kbo_n$도루,center=도루_min,scale=도루_max-도루_min)

도루실패_min<-min(kbo_n$도루실패)
도루실패_max<-max(kbo_n$도루실패)
kbo_n$도루실패<-scale(kbo_n$도루실패,center=도루실패_min,scale=도루실패_max-도루실패_min)

볼넷_min<-min(kbo_n$볼넷)
볼넷_max<-max(kbo_n$볼넷)
kbo_n$볼넷<-scale(kbo_n$볼넷,center=볼넷_min,scale=볼넷_max-볼넷_min)
```

2. 선형회귀분석 (득점예측)

- 2.1 **(데이터 수집 및 전처리)**
- 스케일링&정규화

```
몸에.맞는.공_min<-min(kbo_n$몸에.맞는.공)
몸에.맞는.공_max<-max(kbo_n$몸에.맞는.공)
kbo_n$몸에.맞는.공<-scale(kbo_n$몸에.맞는.공,center=몸에.맞는.공_min,scale=몸에.맞는.공_max-몸에.맞는.공_min)

고의사구_min<-min(kbo_n$고의사구)
고의사구_max<-max(kbo_n$고의사구)
kbo_n$고의사구<-scale(kbo_n$고의사구,center=고의사구_min,scale=고의사구_max-고의사구_min)

삼진_min<-min(kbo_n$삼진)
삼진_max<-max(kbo_n$삼진)
kbo_n$삼진<-scale(kbo_n$삼진,center=삼진_min,scale=삼진_max-삼진_min)

병살_min<-min(kbo_n$병살)
병살_max<-max(kbo_n$병살)
kbo_n$병살<-scale(kbo_n$병살,center=병살_min,scale=병살_max-병살_min)

희생번트_min<-min(kbo_n$희생번트)
희생번트_max<-max(kbo_n$희생번트)
kbo_n$희생번트<-scale(kbo_n$희생번트,center=희생번트_min,scale=희생번트_max-희생번트_min)

희생플라이_min<-min(kbo_n$희생플라이)
희생플라이_max<-max(kbo_n$희생플라이)
kbo_n$희생플라이<-scale(kbo_n$희생플라이,center=희생플라이_min,scale=희생플라이_max-희생플라이_min)

summary(kbo_n)
```

2. 선형회귀분석 (득점예측)

- 2.1 **(데이터 수집 및 전처리)**
- 스케일링&정규화

```
      연도            팀          경기.V1           타석.V1           타수.V1
 Min.   :1982   롯데   : 36   Min.   :0.0000000   Min.   :0.0000000   Min.   :0.0000000
 1st Qu.:1992   삼성   : 36   1st Qu.:0.7187500   1st Qu.:0.5933310   1st Qu.:0.5821400
 Median :2001   LG     : 28   Median :0.7187500   Median :0.6672396   Median :0.6491647
 Mean   :2001   한화   : 24   Mean   :0.7208481   Mean   :0.6523899   Mean   :0.6368984
 3rd Qu.:2010   두산   : 19   3rd Qu.:0.8281250   3rd Qu.:0.7652114   3rd Qu.:0.7392601
 Max.   :2017   해태   : 19   Max.   :1.0000000   Max.   :1.0000000   Max.   :1.0000000
                (Other):121
      득점            안타.V1          X2루타.V1          X3루타.V1          홈런.V1
 Min.   :302.0   Min.   :0.0000000   Min.   :0.0000000   Min.   :0.0000000   Min.   :0.0000000
 1st Qu.:489.5   1st Qu.:0.4269357   1st Qu.:0.3041237   1st Qu.:0.2203390   1st Qu.:0.2170732
 Median :578.0   Median :0.5267176   Median :0.4226804   Median :0.3050847   Median :0.3317073
 Mean   :580.3   Mean   :0.5324244   Mean   :0.4304579   Mean   :0.3145475   Mean   :0.3634750
 3rd Qu.:668.0   3rd Qu.:0.6390403   3rd Qu.:0.5463918   3rd Qu.:0.3898305   3rd Qu.:0.4951220
 Max.   :935.0   Max.   :1.0000000   Max.   :1.0000000   Max.   :1.0000000   Max.   :1.0000000
     총루타.V1          타점.V1           도루.V1          도루실패.V1          볼넷.V1
 Min.   :0.0000000   Min.   :0.0000000   Min.   :0.0000000   Min.   :0.0000000   Min.   :0.0000000
 1st Qu.:0.3697614   1st Qu.:0.3074380   1st Qu.:0.2756757   1st Qu.:0.3333333   1st Qu.:0.4200000
 Median :0.4667956   Median :0.4413223   Median :0.3891892   Median :0.4320988   Median :0.5625000
 Mean   :0.4889721   Mean   :0.4497270   Mean   :0.4026932   Mean   :0.4459713   Mean   :0.5503887
 3rd Qu.:0.6228240   3rd Qu.:0.5909091   3rd Qu.:0.5189189   3rd Qu.:0.5555556   3rd Qu.:0.6887500
 Max.   :1.0000000   Max.   :1.0000000   Max.   :1.0000000   Max.   :1.0000000   Max.   :1.0000000
   몸에.맞는.공.V1      고의사구.V1          삼진.V1          병살.V1          희생번트.V1
 Min.   :0.0000000   Min.   :0.0000000   Min.   :0.0000000   Min.   :0.0000000   Min.   :0.0000000
 1st Qu.:0.2476636   1st Qu.:0.2391304   1st Qu.:0.3803648   1st Qu.:0.4414414   1st Qu.:0.3143939
 Median :0.3738318   Median :0.3260870   Median :0.5300429   Median :0.5315315   Median :0.4242424
 Mean   :0.3922261   Mean   :0.3476725   Mean   :0.5101306   Mean   :0.5464935   Mean   :0.4440518
 3rd Qu.:0.5186916   3rd Qu.:0.4347826   3rd Qu.:0.6561159   3rd Qu.:0.6486486   3rd Qu.:0.5492424
 Max.   :1.0000000   Max.   :1.0000000   Max.   :1.0000000   Max.   :1.0000000   Max.   :1.0000000
   희생플라이.V1
 Min.   :0.0000000
 1st Qu.:0.3035714
 Median :0.4107143
 Mean   :0.4306537
 3rd Qu.:0.5535714
 Max.   :1.0000000
```

2. 선형회귀분석 (득점예측)

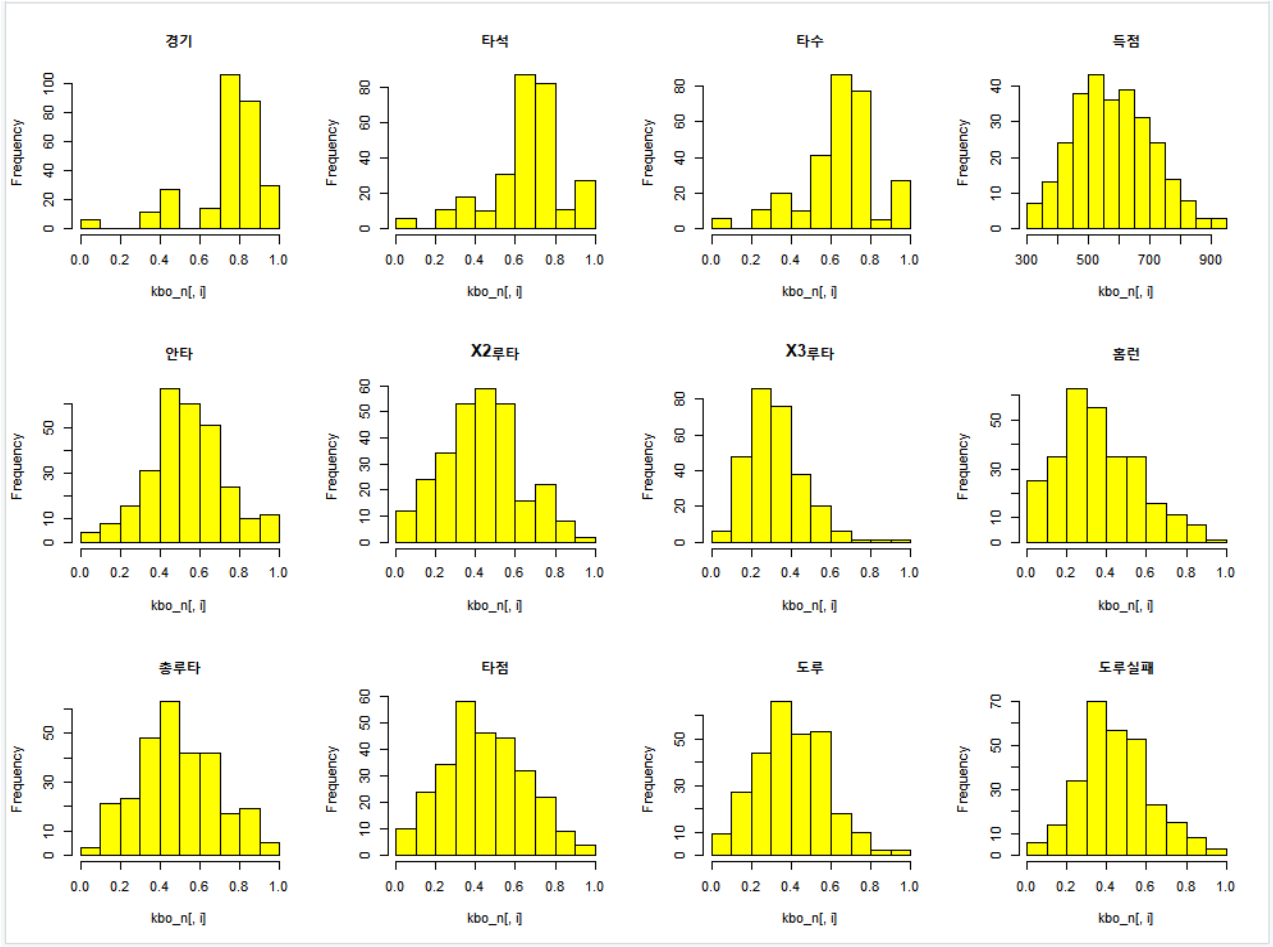
- 2.2 **(탐색적 데이터 분석) - 가시화**
- Histogram

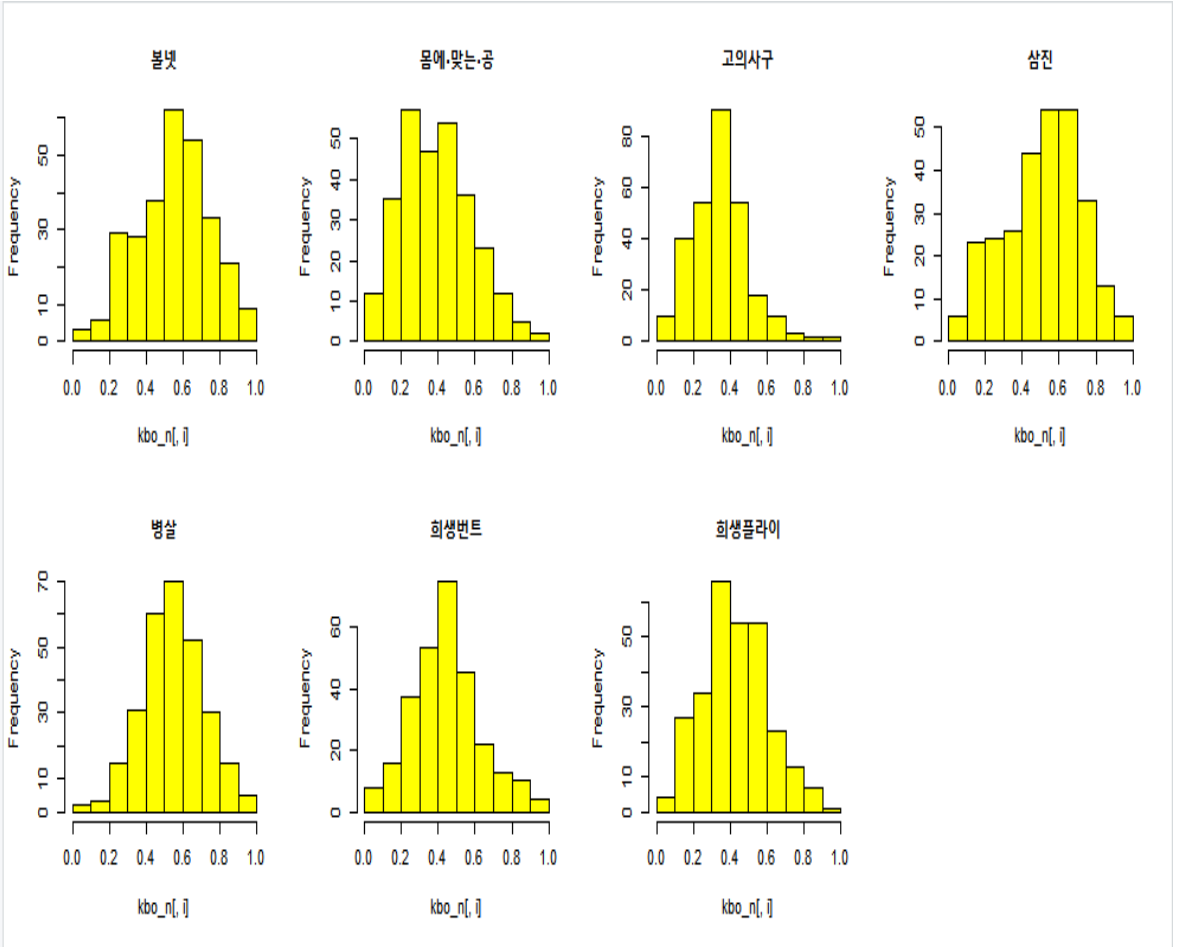
```
par(mfrow=c(3,4))
for(i in 3:14) {
  hist(kbo_n[,i],main=colnames(kbo_n)[i],col="yellow")
}
```

```
par(mfrow=c(3,4))
for(i in 15:21) {
  hist(kbo_n[,i],main=colnames(kbo_n)[i],col="yellow")
}
```

2. 선형회귀분석 (득점예측)

- **2.2 (탐색적 데이터 분석) - 가시화**
- Histogram 가시화





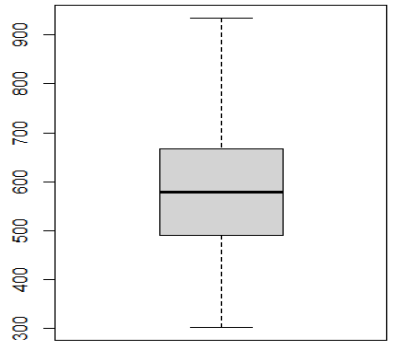
2. 선형회귀분석 (득점예측)

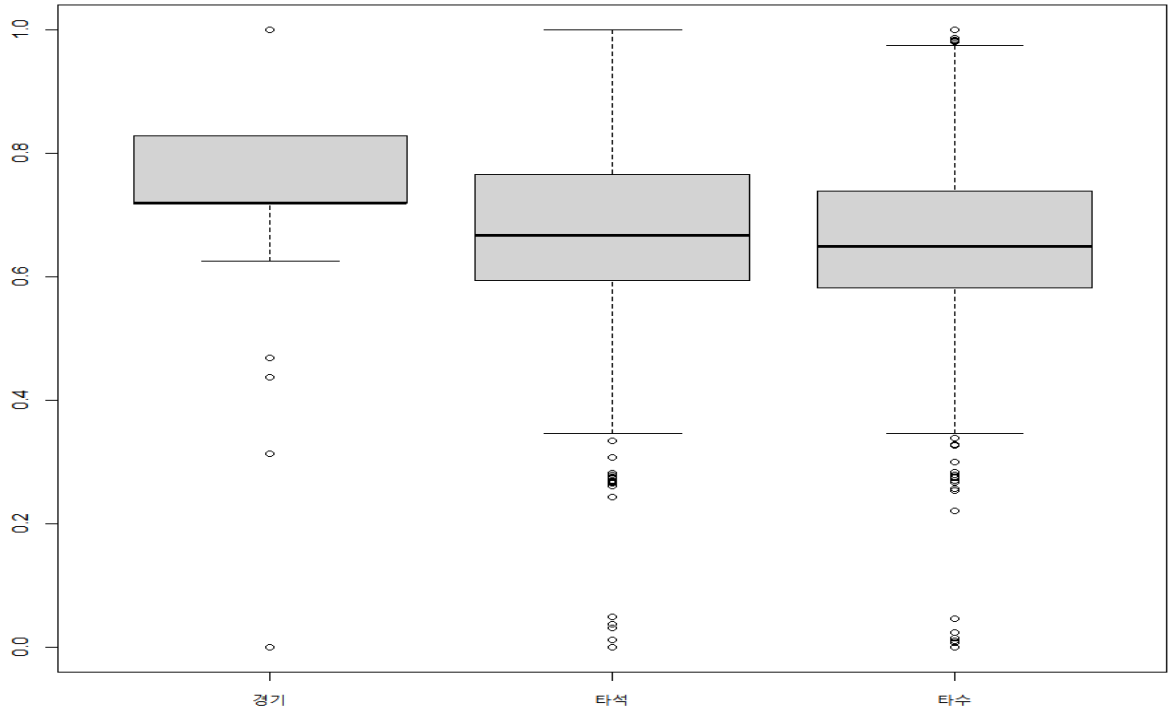
- **2.2 (탐색적 데이터 분석) - 가시화**
- Boxplot 가시화

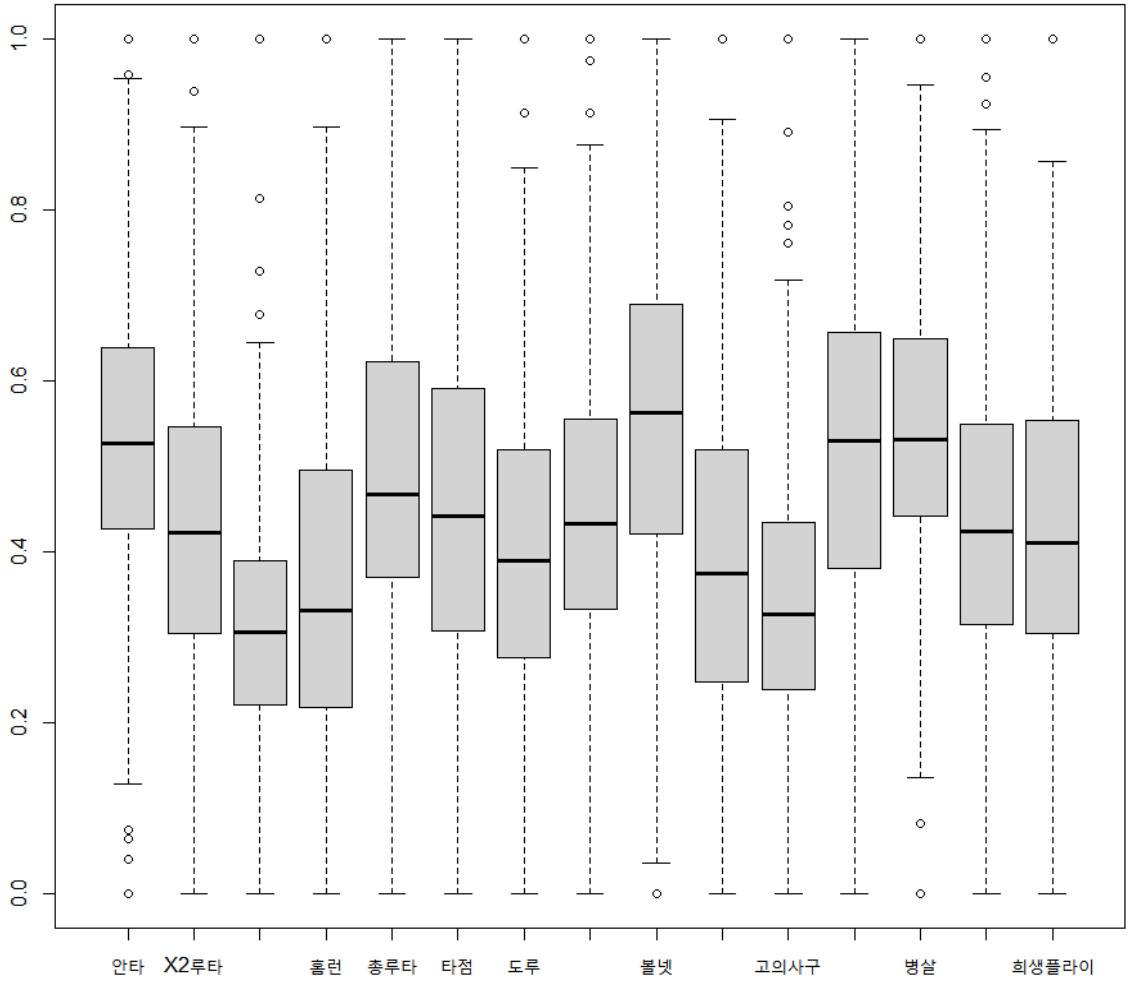
```
boxplot(kbo_n[,c(3:5)])
```

```
boxplot(kbo_n[,c(7:21)])
```

```
boxplot(kbo_n[,c(6)])
```



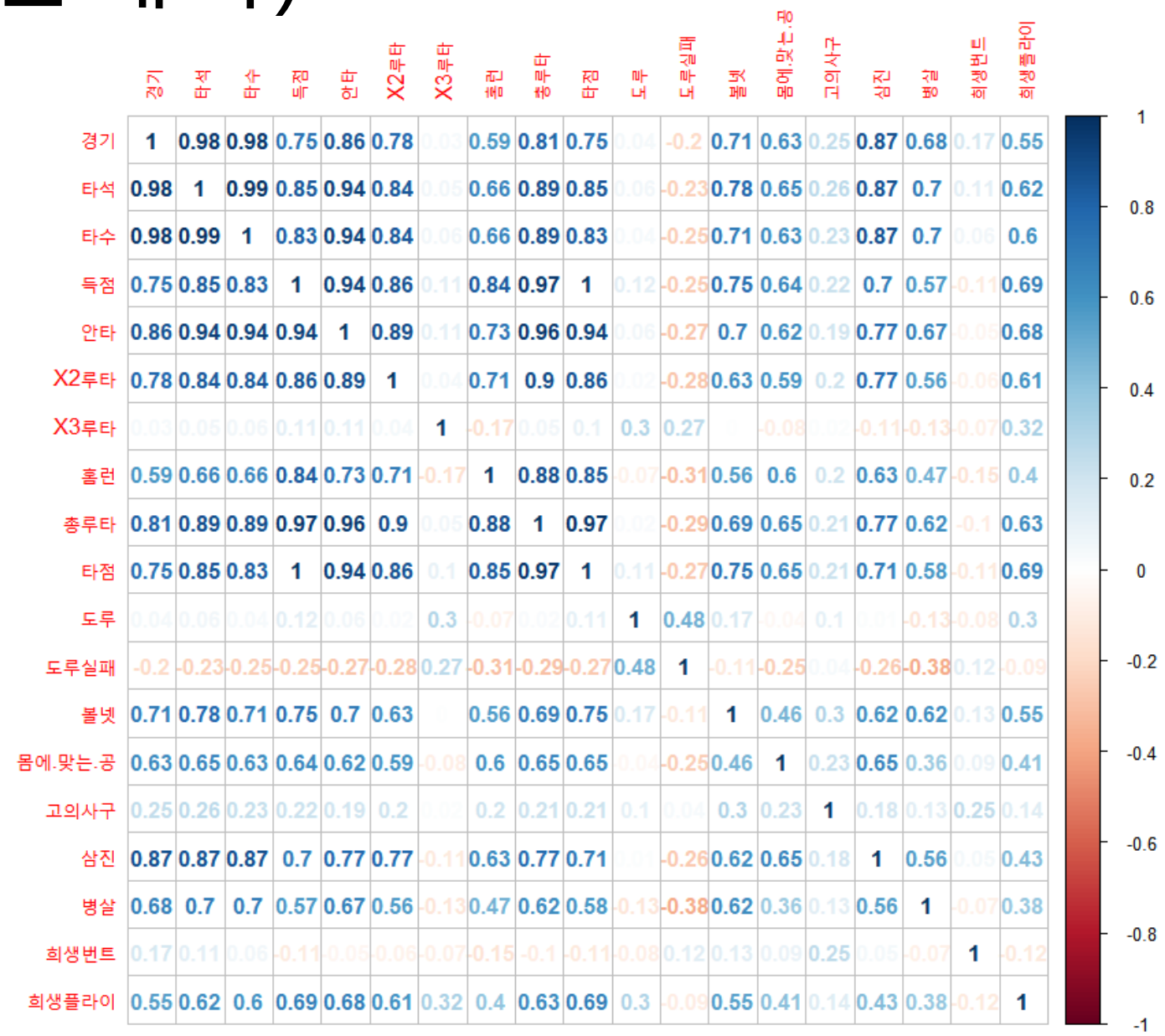




2. 선형회귀분석 (득점예측)

- 2.2 **(탐색적 데이터 분석)**
- correlation matrix

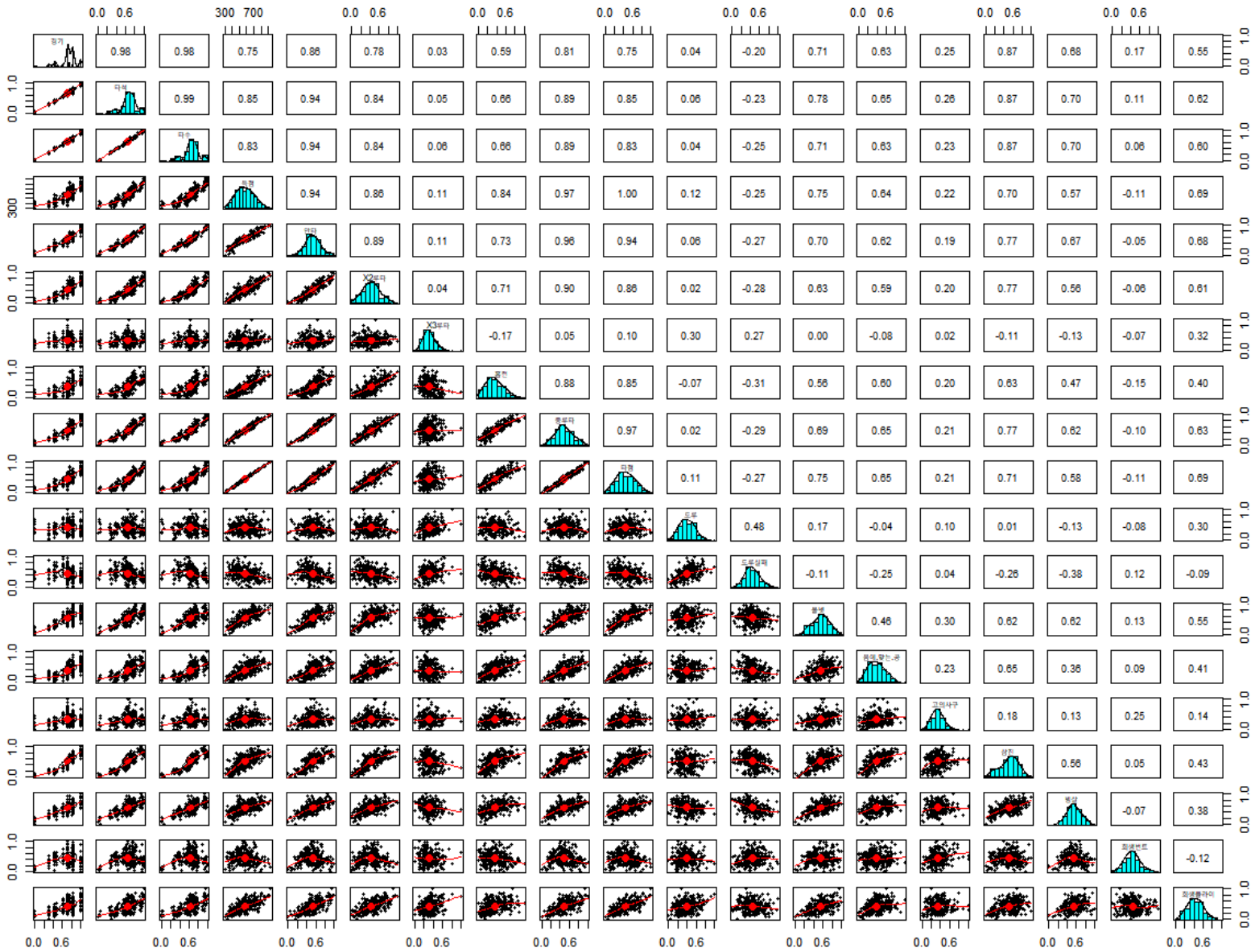
```
install.packages("corrplot")
library(corrplot)
vkbo_n <- kbo_n[,c(3:21)]
par(mfrow=c(1,1))
cor_matrix=cor(vkbo_n)
corrplot(cor_matrix,method="num")
```



2. 선형회귀분석 (득점예측)

- 2.2 **(탐색적 데이터 분석)**
- Multi plots

```
install.packages('psych')
library(psych)
subset <- cbind
pairs.panels(vkbo_n)
```



2. 선형회귀분석 (득점예측)

유의수준 : 0.05

- 2.3 **(학습모델 구축) - 해석**

```
multi_model <- lm(득점 ~ ., data = vkbo_n)
summary(multi_model)
```

```
Call:
lm(formula = 득점 ~ ., data = vkbo_n)

Residuals:
     Min       1Q   Median       3Q      Max
-29.8474  -4.4607  -0.3295   4.9748  19.4427

Coefficients: (1 not defined because of singularities)
             Estimate Std. Error t value Pr(>|t|)
(Intercept)   357.2421    17.2124  20.755  < 2e-16 ***
경기            9.4191    19.8204   0.475 0.635022
타석         3809.6933  1018.3452   3.741 0.000225 ***
타수        -3346.6371   882.9969  -3.790 0.000186 ***
안타           56.1759    19.8082   2.836 0.004921 **
X2루타         15.5404     5.6312   2.760 0.006189 **
X3루타         12.9851     4.3910   2.957 0.003384 **
홈런           -1.9806     6.7301  -0.294 0.768768
총루타             NA         NA      NA       NA
타점          593.5283    15.9894  37.120  < 2e-16 ***
도루            6.2792     3.6134   1.738 0.083417 .
도루실패        0.5071     3.5101   0.144 0.885245
볼넷         -509.2732   140.0353  -3.637 0.000332 ***
몸에.맞는.공 -133.5784    37.7508  -3.538 0.000475 ***
고의사구        0.6540     3.2827   0.199 0.842238
삼진           -1.8733     5.7189  -0.328 0.743506
병살           -4.6132     4.9886  -0.925 0.355931
희생번트     -173.0666    46.6442  -3.710 0.000252 ***
희생플라이    -80.1778    20.0827  -3.992 8.48e-05 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 7.75 on 265 degrees of freedom
Multiple R-squared:  0.9965,    Adjusted R-squared:  0.9963
F-statistic:  4464 on 17 and 265 DF,  p-value: < 2.2e-16
```

· 잔차

- 최솟값 : -29.8474
- 1사분위(25%위치) : -4.4607
- 중앙값 : -0.3295
- 3사분위(75%위치) : 4.9748
- 최댓값 : 19.4427

· 모형의 설명력 : 결정계수 = 0.9965, 수정결정계수 = 0.9963로 약 99%정도 설명 가능해 통계적으로 유의미함.

· 모형의 유의성 : F-statistic = 4464,
p-value(< 2.2e-16)< 0.05 유의미O

2. 선형회귀분석 (득점예측)

```
> #결측치확인
> sum(is.na(vkbo_n$총루타))
[1] 0
```

유의수준 : 0.05

- **2.3 (학습모델 구축) - 해석**

```
Call:
lm(formula = 득점 ~ ., data = vkbo_n)

Residuals:
     Min       1Q   Median       3Q      Max
-29.8474  -4.4607  -0.3295   4.9748  19.4427

Coefficients: (1 not defined because of singularities)
              Estimate Std. Error t value Pr(>|t|)
(Intercept)    357.2421    17.2124  20.755  < 2e-16 ***
경기             9.4191    19.8204   0.475 0.635022
타석          3809.6933  1018.3452   3.741 0.000225 ***
타수         -3346.6371   882.9969  -3.790 0.000186 ***
안타            56.1759    19.8082   2.836 0.004921 **
X2루타          15.5404     5.6312   2.760 0.006189 **
X3루타          12.9851     4.3910   2.957 0.003384 **
홈런            -1.9806     6.7301  -0.294 0.768768
총루타              NA         NA      NA       NA
타점           593.5283    15.9894  37.120  < 2e-16 ***
도루             6.2792     3.6134   1.738 0.083417 .
도루실패         0.5071     3.5101   0.144 0.885245
볼넷          -509.2732   140.0353  -3.637 0.000332 ***
몸에.맞는.공  -133.5784    37.7508  -3.538 0.000475 ***
고의사구         0.6540     3.2827   0.199 0.842238
삼진            -1.8733     5.7189  -0.328 0.743506
병살            -4.6132     4.9886  -0.925 0.355931
희생번트      -173.0666    46.6442  -3.710 0.000252 ***
희생플라이     -80.1778    20.0827  -3.992 8.48e-05 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 7.75 on 265 degrees of freedom
Multiple R-squared:  0.9965,    Adjusted R-squared:  0.9963
F-statistic:  4464 on 17 and 265 DF,  p-value: < 2.2e-16
```

· 총루타는 상관 관계가 높은 변수로 NA로 나타남.
· 경기의 계수 = 9.4191, p-value > 0.05 유의미X
· 타석의 계수 = 3809.6933, p-value < 0.05 유의미O
· 타수의 계수 = -3346.6371, p-value < 0.05 유의미O
· 안타의 계수 = 56.1759, p-value < 0.05 유의미O
· X2루타의 계수 = 15.5404, p-value < 0.05 유의미O
· X3루타의 계수 = 12.9851, p-value < 0.05 유의미O
· 홈런의 계수 = -1.9806, p-value > 0.05 유의미X
· 타점의 계수 = 593.5283, p-value < 0.05 유의미O
· 도루의 계수 = 6.27929, p-value > 0.05 유의미X
· 도루실패의 계수 = 0.5071, p-value > 0.05 유의미X
· 볼넷의 계수 = -509.2732, p-value < 0.05 유의미O
· 몸에.맞는.공의 계수 = -133.5784, p-value < 0.05 유의미O
· 고의사구의 계수 = 0.6540, p-value > 0.05 유의미X
· 삼진의 계수 = -1.8733, p-value > 0.05 유의미X
· 병살의 계수 = -4.6132, p-value > 0.05 유의미X
· 희생번트의 계수 = -173.0666, p-value < 0.05 유의미O
· 희생플라이의 계수 = -80.1778, p-value < 0.05 유의미O

따라서, 초록색 계수들로 득점을 추정할 수 있다.

2. 선형회귀분석 (득점예측)

- 2.4 **(선형회귀 결과해석)**

· MSE(Mean squared error), MAE(Mean absolute error) 측정

```
pred = predict(multi_model, data=vkbo_n)
library(Metrics)
mse(vkbo_n$득점, pred)
mae(vkbo_n$득점, pred)
```

```
> mse(vkbo_n$득점, pred)
[1] 56.24698
> mae(vkbo_n$득점, pred)
[1] 5.777964
```

· 항상 MAE < MSE이다.

· MSE는 잔차 값이 음수됨을 방지해주고, 오차의 민감도를 높인다.
따라서, 에러값이 큰 데이터에 영향을 많이 받는다.

· MAE 값을 통해 전반적인 에러 값이 5.7이라는 것을 알 수 있다.
그러나, 큰 데이터의 영향을 반영하지 못한다.

2. 선형회귀분석 (득점예측)

- 2.5 **(변수선택법 사용해 정확도 향상)**

· p-value가 유의하지 않은 변수 차례로 제거

```
multi_model2 <- lm(득점 ~ 타석 + 타수 + 안타 + X2루타 + X3루타 + 타점 + 도루 + 볼넷 + 몸에.맞는.공 + 병살 + 희생번트 + 희생플라이, data = vkbo_n)
summary(multi_model2)
```

```
Call:
lm(formula = 득점 ~ 타석 + 타수 + 안타 + X2루타 + X3루타 + 타점 +
    도루 + 볼넷 + 몸에.맞는.공 + 병살 + 희생번트 + 희생플라이,
    data = vkbo_n)

Residuals:
     Min       1Q   Median       3Q      Max
-29.5796  -4.3354  -0.3271   4.8305  19.3883

Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept)    356.907     16.507  21.621  < 2e-16 ***
타석          3804.656    993.340   3.830 0.000159 ***
타수         -3333.231    858.493  -3.883 0.000130 ***
안타            55.445     14.953   3.708 0.000254 ***
X2루타          15.248      5.344   2.853 0.004667 **
X3루타          14.023      3.800   3.690 0.000271 ***
타점           589.707      9.860  59.807  < 2e-16 ***
도루             6.886      3.095   2.225 0.026938 *
볼넷          -507.696    136.372  -3.723 0.000240 ***
몸에.맞는.공  -133.133     36.596  -3.638 0.000329 ***
병살            -4.069      4.630  -0.879 0.380299
희생번트      -171.687     45.361  -3.785 0.000189 ***
희생플라이     -79.341     19.459  -4.077    6e-05 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 7.685 on 270 degrees of freedom
Multiple R-squared:  0.9965,    Adjusted R-squared:  0.9964
F-statistic:  6432 on 12 and 270 DF,  p-value: < 2.2e-16
```

경기, 홈런, 총루타,
도루실패, 고의사구, 삼진 제거

=> 수정결정계수 증가.

2. 선형회귀분석 (득점예측)

- 2.5 **(변수선택법 사용해 정확도 향상)**

· p-value가 유의하지 않은 변수 차례로 제거

```
multi_model <- lm(득점 ~ 타석 + 타수 + 안타 + X2루타 + X3루타 + 타점 +  볼넷 + 몸에.맞는.공 + 희생번트 + 희생플라이, data = vkbo_n)
summary(multi_model)
```

```
Call:
lm(formula = 득점 ~ 타석 + 타수 + 안타 + X2루타 + X3루타 + 타점 +
    볼넷 + 몸에.맞는.공 + 희생번트 + 희생플라이, data = vkbo_n)

Residuals:
     Min       1Q   Median       3Q      Max
-26.9610  -4.5530   0.1309   5.0175  19.0472

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)   358.811     16.450  21.812  < 2e-16 ***
타석         3858.213   1000.328   3.857 0.000143 ***
타수        -3378.747    864.614  -3.908 0.000118 ***
안타           47.181     14.449   3.265 0.001233 **
X2루타         14.863      5.356   2.775 0.005898 **
X3루타         16.622      3.672   4.527 8.96e-06 ***
타점          594.002      9.602  61.865  < 2e-16 ***
볼넷         -515.524    137.447  -3.751 0.000215 ***
몸에.맞는.공 -135.499     36.823  -3.680 0.000281 ***
희생번트     -173.806     45.617  -3.810 0.000172 ***
희생플라이    -77.741     19.610  -3.964 9.42e-05 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 7.759 on 272 degrees of freedom
Multiple R-squared:  0.9964,    Adjusted R-squared:  0.9963
F-statistic:  7572 on 10 and 272 DF,  p-value: < 2.2e-16
```

도루, 병살 추가로 제거

=> 정확도 떨어짐.

2. 선형회귀분석 (득점예측)

- 2.5 **(변수선택법 사용해 정확도 향상)**

· Forward selection, Backward elimination, 외 Stepwise selection 선택적 사용

```
vkbo_n.forward <- step(multi_model, direction = "forward")
summary(vkbo_n.forward)
```

```
Start:  AIC=1176.42
득점 ~ 경기 + 타석 + 타수 + 안타 + X2루타 + X3루타 + 홈런 + 총루타 +
    타점 + 도루 + 도루실패 + 볼넷 + 몸에.맞는.공 + 고의사구 +
    삼진 + 병살 + 희생번트 + 희생플라이

Call:
lm(formula = 득점 ~ 경기 + 타석 + 타수 + 안타 + X2루타 + X3루타 +
    홈런 + 총루타 + 타점 + 도루 + 도루실패 + 볼넷 + 몸에.맞는.공 +
    고의사구 + 삼진 + 병살 + 희생번트 + 희생플라이, data = vkbo_n)

Residuals:
     Min       1Q   Median       3Q      Max
-29.8474  -4.4607  -0.3295   4.9748  19.4427

Coefficients: (1 not defined because of singularities)
              Estimate Std. Error t value Pr(>|t|)
(Intercept)    357.2421    17.2124  20.755  < 2e-16 ***
경기             9.4191    19.8204   0.475 0.635022
타석          3809.6933  1018.3452   3.741 0.000225 ***
타수         -3346.6371   882.9969  -3.790 0.000186 ***
안타            56.1759    19.8082   2.836 0.004921 **
X2루타          15.5404     5.6312   2.760 0.006189 **
X3루타          12.9851     4.3910   2.957 0.003384 **
홈런            -1.9806     6.7301  -0.294 0.768768
총루타              NA         NA      NA       NA
타점           593.5283    15.9894  37.120  < 2e-16 ***
도루             6.2792     3.6134   1.738 0.083417 .
도루실패         0.5071     3.5101   0.144 0.885245
볼넷          -509.2732   140.0353  -3.637 0.000332 ***
몸에.맞는.공  -133.5784    37.7508  -3.538 0.000475 ***
고의사구         0.6540     3.2827   0.199 0.842238
삼진            -1.8733     5.7189  -0.328 0.743506
병살            -4.6132     4.9886  -0.925 0.355931
희생번트      -173.0666    46.6442  -3.710 0.000252 ***
희생플라이     -80.1778    20.0827  -3.992 8.48e-05 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 7.75 on 265 degrees of freedom
Multiple R-squared:  0.9965,    Adjusted R-squared:  0.9963
F-statistic:  4464 on 17 and 265 DF,  p-value: < 2.2e-16
```

Forward selection

=> 기존 모델과 별차이 없음.

2. 선형회귀분석 (득점예측)

- 2.5 **(변수선택법 사용해 정확도 향상)**

· Forward selection, Backward elimination, 외 Stepwise selection 선택적 사용

```
vkbo_n.backward <- step(multi_model, direction = "backward")
summary(vkbo_n.backward)
```

```
Step:  AIC=1165.75
득점 ~ 타석 + 타수 + 안타 + X2루타 + X3루타 + 타점 + 도루 + 볼넷 +
    몸에.맞는.공 + 희생번트 + 희생플라이

Call:
lm(formula = 득점 ~ 타석 + 타수 + 안타 + X2루타 + X3루타 + 타점 +
    도루 + 볼넷 + 몸에.맞는.공 + 희생번트 + 희생플라이, data = vkbo_n)

Residuals:
     Min       1Q   Median       3Q      Max
-29.4681  -4.5591  -0.0305   4.9175  19.3663

Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept)    354.995     16.356  21.704  < 2e-16 ***
타석          3754.336    991.271   3.787 0.000188 ***
타수         -3290.584    856.760  -3.841 0.000153 ***
안타            52.029     14.432   3.605 0.000371 ***
X2루타          15.734      5.314   2.961 0.003338 **
X3루타          14.741      3.710   3.973 9.10e-05 ***
타점           591.871      9.543  62.019  < 2e-16 ***
도루             7.595      2.987   2.542 0.011565 *
볼넷          -502.460    136.184  -3.690 0.000271 ***
몸에.맞는.공  -131.023     36.501  -3.590 0.000393 ***
희생번트      -168.661     45.211  -3.731 0.000233 ***
희생플라이     -78.348     19.418  -4.035 7.11e-05 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 7.682 on 271 degrees of freedom
Multiple R-squared:  0.9965,    Adjusted R-squared:  0.9964
F-statistic:  7022 on 11 and 271 DF,  p-value: < 2.2e-16
```

Backward elimination

⇒ **기존 모델 보다 수정결정계수 높아짐.**

⇒ **Forward selection보다 AIC 작음.**

2. 선형회귀분석 (득점예측)

- 2.5 **(변수선택법 사용해 정확도 향상)**

· Forward selection, Backward elimination, 외 Stepwise selection 선택적 사용

```
vkbo_n.stepwise <- step(multi_model, direction = "both")
summary(vkbo_n.stepwise)
```

```
Step:  AIC=1165.75
득점 ~ 타석 + 타수 + 안타 + X2루타 + X3루타 + 타점 + 도루 + 볼넷 +
    몸에.맞는.공 + 희생번트 + 희생플라이

Call:
lm(formula = 득점 ~ 타석 + 타수 + 안타 + X2루타 + X3루타 + 타점 +
    도루 + 볼넷 + 몸에.맞는.공 + 희생번트 + 희생플라이, data = vkbo_n)

Residuals:
     Min       1Q   Median       3Q      Max
-29.4681  -4.5591  -0.0305   4.9175  19.3663

Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept)    354.995     16.356  21.704  < 2e-16 ***
타석          3754.336    991.271   3.787 0.000188 ***
타수         -3290.584    856.760  -3.841 0.000153 ***
안타            52.029     14.432   3.605 0.000371 ***
X2루타          15.734      5.314   2.961 0.003338 **
X3루타          14.741      3.710   3.973 9.10e-05 ***
타점           591.871      9.543  62.019  < 2e-16 ***
도루             7.595      2.987   2.542 0.011565 *
볼넷          -502.460    136.184  -3.690 0.000271 ***
몸에.맞는.공  -131.023     36.501  -3.590 0.000393 ***
희생번트      -168.661     45.211  -3.731 0.000233 ***
희생플라이     -78.348     19.418  -4.035 7.11e-05 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 7.682 on 271 degrees of freedom
Multiple R-squared:  0.9965,    Adjusted R-squared:  0.9964
F-statistic:  7022 on 11 and 271 DF,  p-value: < 2.2e-16
```

Stepwise selection

⇒ **기존 모델 보다 수정결정계수 높아짐.**

⇒ **Forward selection보다 AIC 작음.**

2. 로지스틱회귀분석 (TenYearCHD 예측)

- 2.1 **(데이터 수집 및 전처리)**

- 데이터 수집 : framingham.csv (심장병)

- 변수(15개 이상) : male(성별), age(나이), education(최고학력), currentSmoker(최근 담배 핀 여부), cigsPerDay(하루에 핀 담배 수), BPMeds(혈압약 복용 여부), prevalentStroke(뇌졸중 기록), prevalentHyp(고혈압 기록), diabetes(당뇨병 기록), totChol(콜레스트롤 단계), sysBP(혈압수준), diaBP(확장기 혈압), BMI(체질량 지수), heartrate(심박수 판독값), glucose(포도당수치), TenYearCHD(향후 10년 관상동맥 질환 걸릴 여부)

```
framingham <- read.csv("C:/RStudio/datamining/framingham.csv",header=TRUE)
framingham
```

2. 로지스틱회귀분석 (TenYearCHD 예측)

- 2.1 **(데이터 수집 및 전처리)**
- 데이터 수집 : framingham.csv (심장병)

	male	age	education	currentSmoker	cigsPerDay	BPMeds	prevalentStroke	prevalentHyp	diabetes
1	1	39	4	0	0	0	0	0	0
2	0	46	2	0	0	0	0	0	0
3	1	48	1	1	20	0	0	0	0
4	0	61	3	1	30	0	0	1	0
5	0	46	3	1	23	0	0	0	0
6	0	43	2	0	0	0	0	1	0
7	0	63	1	0	0	0	0	0	0
8	0	45	2	1	20	0	0	0	0
9	1	52	1	0	0	0	0	1	0
10	1	43	1	1	30	0	0	1	0
11	0	50	1	0	0	0	0	0	0
12	0	43	2	0	0	0	0	0	0
13	1	46	1	1	15	0	0	1	0
14	0	41	3	0	0	1	0	1	0
15	0	39	2	1	9	0	0	0	0
16	0	38	2	1	20	0	0	1	0
17	1	48	3	1	10	0	0	1	0
18	0	46	2	1	20	0	0	0	0
19	0	38	2	1	5	0	0	0	0
20	1	41	2	0	0	0	0	0	0
21	0	42	2	1	30	0	0	0	0
22	0	43	1	0	0	0	0	0	0
23	0	52	1	0	0	0	0	0	0
24	0	52	3	1	20	0	0	0	0
25	1	44	2	1	30	0	0	1	0
26	1	47	4	1	20	0	0	0	0
27	0	60	1	0	0	0	0	0	0
28	1	35	2	1	20	0	0	1	0
29	0	61	3	0	0	0	0	1	0
30	0	60	1	0	0	0	0	0	0
31	1	36	4	1	35	0	0	0	0
32	1	43	4	1	43	0	0	0	0
33	0	59	1	0	0	0	0	1	0
34	1	61	NA	1	5	0	0	0	0
35	1	54	1	1	20	0	0	1	0
36	1	37	2	0	0	0	0	1	0
37	1	56	NA	0	0	0	0	0	0
38	1	52	1	0	0	0	0	1	1

	totChol	sysBP	diaBP	BMI	heartRate	glucose	TenYearCHD
1	195	106.0	70.0	26.97	80	77	0
2	250	121.0	81.0	28.73	95	76	0
3	245	127.5	80.0	25.34	75	70	0
4	225	150.0	95.0	28.58	65	103	1
5	285	130.0	84.0	23.10	85	85	0
6	228	180.0	110.0	30.30	77	99	0
7	205	138.0	71.0	33.11	60	85	1
8	313	100.0	71.0	21.68	79	78	0
9	260	141.5	89.0	26.36	76	79	0
10	225	162.0	107.0	23.61	93	88	0
11	254	133.0	76.0	22.91	75	76	0
12	247	131.0	88.0	27.64	72	61	0
13	294	142.0	94.0	26.31	98	64	0
14	332	124.0	88.0	31.31	65	84	0
15	226	114.0	64.0	22.35	85	NA	0
16	221	140.0	90.0	21.35	95	70	1
17	232	138.0	90.0	22.37	64	72	0
18	291	112.0	78.0	23.38	80	89	1
19	195	122.0	84.5	23.24	75	78	0
20	195	139.0	88.0	26.88	85	65	0
21	190	108.0	70.5	21.59	72	85	0
22	185	123.5	77.5	29.89	70	NA	0
23	234	148.0	78.0	34.17	70	113	0
24	215	132.0	82.0	25.11	71	75	0
25	270	137.5	90.0	21.96	75	83	0
26	294	102.0	68.0	24.18	62	66	1
27	260	110.0	72.5	26.59	65	NA	0
28	225	132.0	91.0	26.09	73	83	0
29	272	182.0	121.0	32.80	85	65	1
30	247	130.0	88.0	30.36	72	74	0
31	295	102.0	68.0	28.15	60	63	0
32	226	115.0	85.5	27.57	75	75	0
33	209	150.0	85.0	20.77	90	88	1
34	175	134.0	82.5	18.59	72	75	1
35	214	147.0	74.0	24.71	96	87	0
36	225	124.5	92.5	38.53	95	83	0
37	257	153.5	102.0	28.09	72	75	0
38	178	160.0	98.0	40.11	75	225	0

2. 로지스틱회귀분석 (TenYearCHD 예측)

- 2.1 **(데이터 수집 및 전처리)**
- 데이터타입 확인

```
str(framingham)
```

```
'data.frame':   4238 obs. of  16 variables:
 $ male            : int  1 0 1 0 0 0 0 0 1 1 ...
 $ age             : int  39 46 48 61 46 43 63 45 52 43 ...
 $ education       : int  4 2 1 3 3 2 1 2 1 1 ...
 $ currentSmoker   : int  0 0 1 1 1 0 0 1 0 1 ...
 $ cigsPerDay      : int  0 0 20 30 23 0 0 20 0 30 ...
 $ BPMeds          : int  0 0 0 0 0 0 0 0 0 0 ...
 $ prevalentStroke: int  0 0 0 0 0 0 0 0 0 0 ...
 $ prevalentHyp    : int  0 0 0 1 0 1 0 0 1 1 ...
 $ diabetes        : int  0 0 0 0 0 0 0 0 0 0 ...
 $ totChol         : int  195 250 245 225 285 228 205 313 260 225 ...
 $ sysBP           : num  106 121 128 150 130 ...
 $ diaBP           : num  70 81 80 95 84 110 71 71 89 107 ...
 $ BMI             : num  27 28.7 25.3 28.6 23.1 ...
 $ heartRate       : int  80 95 75 65 85 77 60 79 76 93 ...
 $ glucose         : int  77 76 70 103 85 99 85 78 79 88 ...
 $ TenYearCHD      : int  0 0 0 1 0 0 1 0 0 0 ...
```

- 결측치제거

```
framingham <- na.omit(framingham)
View(framingham)
```

```
'data.frame':   3656 obs. of  16 variables:
 $ male            : int  1 0 1 0 0 0 0 0 1 1 ...
 $ age             : int  39 46 48 61 46 43 63 45 52 43 ...
 $ education       : int  4 2 1 3 3 2 1 2 1 1 ...
 $ currentSmoker   : int  0 0 1 1 1 0 0 1 0 1 ...
 $ cigsPerDay      : int  0 0 20 30 23 0 0 20 0 30 ...
 $ BPMeds          : int  0 0 0 0 0 0 0 0 0 0 ...
 $ prevalentStroke: int  0 0 0 0 0 0 0 0 0 0 ...
 $ prevalentHyp    : int  0 0 0 1 0 1 0 0 1 1 ...
 $ diabetes        : int  0 0 0 0 0 0 0 0 0 0 ...
 $ totChol         : int  195 250 245 225 285 228 205 313 260 225 ...
 $ sysBP           : num  106 121 128 150 130 ...
 $ diaBP           : num  70 81 80 95 84 110 71 71 89 107 ...
 $ BMI             : num  27 28.7 25.3 28.6 23.1 ...
 $ heartRate       : int  80 95 75 65 85 77 60 79 76 93 ...
 $ glucose         : int  77 76 70 103 85 99 85 78 79 88 ...
 $ TenYearCHD      : int  0 0 0 1 0 0 1 0 0 0 ...
 - attr(*, "na.action")= 'omit' Named int [1:582] 15 22 27 34 37 43 50 55 71 73 ...
  ..- attr(*, "names")= chr [1:582] "15" "22" "27" "34" ...
```

2. 로지스틱회귀분석 (TenYearCHD 예측)

- 2.1 **(데이터 수집 및 전처리)**
- 범주형데이터 -> 더미변수로 변환

```
library(dummies)
framingham <- dummy.data.frame(framingham, names=c("education"))
```

```
'data.frame':   3656 obs. of  19 variables:
 $ male           : int  1 0 1 0 0 0 0 0 1 1 ...
 $ age            : int  39 46 48 61 46 43 63 45 52 43 ...
 $ education1     : int  0 0 1 0 0 0 1 0 1 1 ...
 $ education2     : int  0 1 0 0 0 1 0 1 0 0 ...
 $ education3     : int  0 0 0 1 1 0 0 0 0 0 ...
 $ education4     : int  1 0 0 0 0 0 0 0 0 0 ...
 $ currentSmoker  : int  0 0 1 1 1 0 0 1 0 1 ...
 $ cigsPerDay     : int  0 0 20 30 23 0 0 20 0 30 ...
 $ BPMeds         : int  0 0 0 0 0 0 0 0 0 0 ...
 $ prevalentStroke: int  0 0 0 0 0 0 0 0 0 0 ...
 $ prevalentHyp   : int  0 0 0 1 0 1 0 0 1 1 ...
 $ diabetes       : int  0 0 0 0 0 0 0 0 0 0 ...
 $ totChol        : int  195 250 245 225 285 228 205 313 260 225 ...
 $ sysBP          : num  106 121 128 150 130 ...
 $ diaBP          : num  70 81 80 95 84 110 71 71 89 107 ...
 $ BMI            : num  27 28.7 25.3 28.6 23.1 ...
 $ heartRate      : int  80 95 75 65 85 77 60 79 76 93 ...
 $ glucose        : int  77 76 70 103 85 99 85 78 79 88 ...
 $ TenYearCHD     : int  0 0 0 1 0 0 1 0 0 0 ...
 - attr(*, "dummies")=List of 1
  ..$ education: int [1:4] 3 4 5 6
```

2. 로지스틱회귀분석 (TenYearCHD 예측)

- 2.1 **(데이터 수집 및 전처리)**
- 데이터확인

summary(framingham)

```
      male              age          education1       education2       education3       education4     
 Min.   :0.0000   Min.   :32.00   Min.   :0.0000   Min.   :0.0000   Min.   :0.0000   Min.   :0.0000  
 1st Qu.:0.0000   1st Qu.:42.00   1st Qu.:0.0000   1st Qu.:0.0000   1st Qu.:0.0000   1st Qu.:0.0000  
 Median :0.0000   Median :49.00   Median :0.0000   Median :0.0000   Median :0.0000   Median :0.0000  
 Mean   :0.4437   Mean   :49.56   Mean   :0.4174   Mean   :0.3011   Mean   :0.1658   Mean   :0.1157  
 3rd Qu.:1.0000   3rd Qu.:56.00   3rd Qu.:1.0000   3rd Qu.:1.0000   3rd Qu.:0.0000   3rd Qu.:0.0000  
 Max.   :1.0000   Max.   :70.00   Max.   :1.0000   Max.   :1.0000   Max.   :1.0000   Max.   :1.0000  
 currentSmoker      cigsPerDay        BPMeds        prevalentStroke     prevalentHyp       diabetes       
 Min.   :0.0000   Min.   : 0.000   Min.   :0.00000   Min.   :0.000000   Min.   :0.0000   Min.   :0.00000  
 1st Qu.:0.0000   1st Qu.: 0.000   1st Qu.:0.00000   1st Qu.:0.000000   1st Qu.:0.0000   1st Qu.:0.00000  
 Median :0.0000   Median : 0.000   Median :0.00000   Median :0.000000   Median :0.0000   Median :0.00000  
 Mean   :0.4891   Mean   : 9.022   Mean   :0.03036   Mean   :0.005744   Mean   :0.3115   Mean   :0.02708  
 3rd Qu.:1.0000   3rd Qu.:20.000   3rd Qu.:0.00000   3rd Qu.:0.000000   3rd Qu.:1.0000   3rd Qu.:0.00000  
 Max.   :1.0000   Max.   :70.000   Max.   :1.00000   Max.   :1.000000   Max.   :1.0000   Max.   :1.00000  
    totChol         sysBP           diaBP             BMI          heartRate         glucose        TenYearCHD     
 Min.   :113.0   Min.   : 83.5   Min.   : 48.00   Min.   :15.54   Min.   : 44.00   Min.   : 40.00   Min.   :0.0000  
 1st Qu.:206.0   1st Qu.:117.0   1st Qu.: 75.00   1st Qu.:23.08   1st Qu.: 68.00   1st Qu.: 71.00   1st Qu.:0.0000  
 Median :234.0   Median :128.0   Median : 82.00   Median :25.38   Median : 75.00   Median : 78.00   Median :0.0000  
 Mean   :236.9   Mean   :132.4   Mean   : 82.91   Mean   :25.78   Mean   : 75.73   Mean   : 81.86   Mean   :0.1524  
 3rd Qu.:263.2   3rd Qu.:144.0   3rd Qu.: 90.00   3rd Qu.:28.04   3rd Qu.: 82.00   3rd Qu.: 87.00   3rd Qu.:0.0000  
 Max.   :600.0   Max.   :295.0   Max.   :142.50   Max.   :56.80   Max.   :143.00   Max.   :394.00   Max.   :1.0000  
```

2. 로지스틱회귀분석 (TenYearCHD 예측)

- 2.1 **(데이터 수집 및 전처리)**
- 스케일링&정규화

```
framingham_n<-data.frame(framingham)

age_min<-min(framingham_n$age)
age_max<-max(framingham_n$age)
framingham_n$age<-scale(framingham_n$age,center=age_min,scale=age_max-age_min)

currentSmoker_min<-min(framingham_n$currentSmoker)
currentSmoker_max<-max(framingham_n$currentSmoker)
framingham_n$currentSmoker<-scale(framingham_n$currentSmoker,center=currentSmoker_min,scale=currentSmoker_max-currentSmoker_min)

cigsPerDay_min<-min(framingham_n$cigsPerDay)
cigsPerDay_max<-max(framingham_n$cigsPerDay)
framingham_n$cigsPerDay<-scale(framingham_n$cigsPerDay,center=cigsPerDay_min,scale=cigsPerDay_max-cigsPerDay_min)

BPMeds_min<-min(framingham_n$BPMeds)
BPMeds_max<-max(framingham_n$BPMeds)
framingham_n$BPMeds<-scale(framingham_n$BPMeds,center=BPMeds_min,scale=BPMeds_max-BPMeds_min)

prevalentStroke_min<-min(framingham_n$prevalentStroke)
prevalentStroke_max<-max(framingham_n$prevalentStroke)
framingham_n$prevalentStroke<-scale(framingham_n$prevalentStroke,center=prevalentStroke_min,scale=prevalentStroke_max-prevalentStroke_min)

prevalentHyp_min<-min(framingham_n$prevalentHyp)
prevalentHyp_max<-max(framingham_n$prevalentHyp)
framingham_n$prevalentHyp<-scale(framingham_n$prevalentHyp,center=prevalentHyp_min,scale=prevalentHyp_max-prevalentHyp_min)
```

2. 로지스틱회귀분석 (TenYearCHD 예측)

- 2.1 **(데이터 수집 및 전처리)**
- 스케일링&정규화

```
diabetes_min<-min(framingham_n$diabetes)
diabetes_max<-max(framingham_n$diabetes)
framingham_n$diabetes<-scale(framingham_n$diabetes,center=diabetes_min,scale=diabetes_max-diabetes_min)

totChol_min<-min(framingham_n$totChol)
totChol_max<-max(framingham_n$totChol)
framingham_n$totChol<-scale(framingham_n$totChol,center=totChol_min,scale=totChol_max-totChol_min)

sysBP_min<-min(framingham_n$sysBP)
sysBP_max<-max(framingham_n$sysBP)
framingham_n$sysBP<-scale(framingham_n$sysBP,center=sysBP_min,scale=sysBP_max-sysBP_min)

diaBP_min<-min(framingham_n$diaBP)
diaBP_max<-max(framingham_n$diaBP)
framingham_n$diaBP<-scale(framingham_n$diaBP,center=diaBP_min,scale=diaBP_max-diaBP_min)

BMI_min<-min(framingham_n$BMI)
BMI_max<-max(framingham_n$BMI)
framingham_n$BMI<-scale(framingham_n$BMI,center=BMI_min,scale=BMI_max-BMI_min)

heartRate_min<-min(framingham_n$heartRate)
heartRate_max<-max(framingham_n$heartRate)
framingham_n$heartRate<-scale(framingham_n$heartRate,center=heartRate_min,scale=heartRate_max-heartRate_min)
```

2. 로지스틱회귀분석 (TenYearCHD 예측)

- 2.1 **(데이터 수집 및 전처리)**
- 스케일링&정규화

```
glucose_min<-min(framingham_n$glucose)
glucose_max<-max(framingham_n$glucose)
framingham_n$glucose<-scale(framingham_n$glucose,center=glucose_min,scale=glucose_max-glucose_min)

summary(framingham_n)
```

```
      male             age.V1          education1       education2       education3       education4
 Min.   :0.0000   Min.   :0.0000000   Min.   :0.0000   Min.   :0.0000   Min.   :0.0000   Min.   :0.0000
 1st Qu.:0.0000   1st Qu.:0.2631579   1st Qu.:0.0000   1st Qu.:0.0000   1st Qu.:0.0000   1st Qu.:0.0000
 Median :0.0000   Median :0.4473684   Median :0.0000   Median :0.0000   Median :0.0000   Median :0.0000
 Mean   :0.4437   Mean   :0.4620379   Mean   :0.4174   Mean   :0.3011   Mean   :0.1658   Mean   :0.1157
 3rd Qu.:1.0000   3rd Qu.:0.6315789   3rd Qu.:1.0000   3rd Qu.:1.0000   3rd Qu.:0.0000   3rd Qu.:0.0000
 Max.   :1.0000   Max.   :1.0000000   Max.   :1.0000   Max.   :1.0000   Max.   :1.0000   Max.   :1.0000
  currentSmoker.V1    cigsPerDay.V1        BPMeds.V1       prevalentStroke.V1   prevalentHyp.V1
 Min.   :0.0000000   Min.   :0.0000000   Min.   :0.0000000   Min.   :0.000000   Min.   :0.0000000
 1st Qu.:0.0000000   1st Qu.:0.0000000   1st Qu.:0.0000000   1st Qu.:0.000000   1st Qu.:0.0000000
 Median :0.0000000   Median :0.0000000   Median :0.0000000   Median :0.000000   Median :0.0000000
 Mean   :0.4890591   Mean   :0.1288879   Mean   :0.0303611   Mean   :0.005744   Mean   :0.3115427
 3rd Qu.:1.0000000   3rd Qu.:0.2857143   3rd Qu.:0.0000000   3rd Qu.:0.000000   3rd Qu.:1.0000000
 Max.   :1.0000000   Max.   :1.0000000   Max.   :1.0000000   Max.   :1.000000   Max.   :1.0000000
    diabetes.V1         totChol.V1          sysBP.V1            diaBP.V1             BMI.V1
 Min.   :0.0000000   Min.   :0.0000000   Min.   :0.0000000   Min.   :0.0000000   Min.   :0.0000000
 1st Qu.:0.0000000   1st Qu.:0.1909651   1st Qu.:0.1583924   1st Qu.:0.2857143   1st Qu.:0.1827436
 Median :0.0000000   Median :0.2484600   Median :0.2104019   Median :0.3597884   Median :0.2384876
 Mean   :0.0270788   Mean   :0.2543595   Mean   :0.2310545   Mean   :0.3694398   Mean   :0.2482837
 3rd Qu.:0.0000000   3rd Qu.:0.3085216   3rd Qu.:0.2860520   3rd Qu.:0.4444444   3rd Qu.:0.3029569
 Max.   :1.0000000   Max.   :1.0000000   Max.   :1.0000000   Max.   :1.0000000   Max.   :1.0000000
   heartRate.V1        glucose.V1        TenYearCHD
 Min.   :0.0000000   Min.   :0.0000000   Min.   :0.0000
 1st Qu.:0.2424242   1st Qu.:0.0875706   1st Qu.:0.0000
 Median :0.3131313   Median :0.1073446   Median :0.0000
 Mean   :0.3205109   Mean   :0.1182376   Mean   :0.1524
 3rd Qu.:0.3838384   3rd Qu.:0.1327684   3rd Qu.:0.0000
 Max.   :1.0000000   Max.   :1.0000000   Max.   :1.0000
```

2. 로지스틱회귀분석 (TenYearCHD 예측)

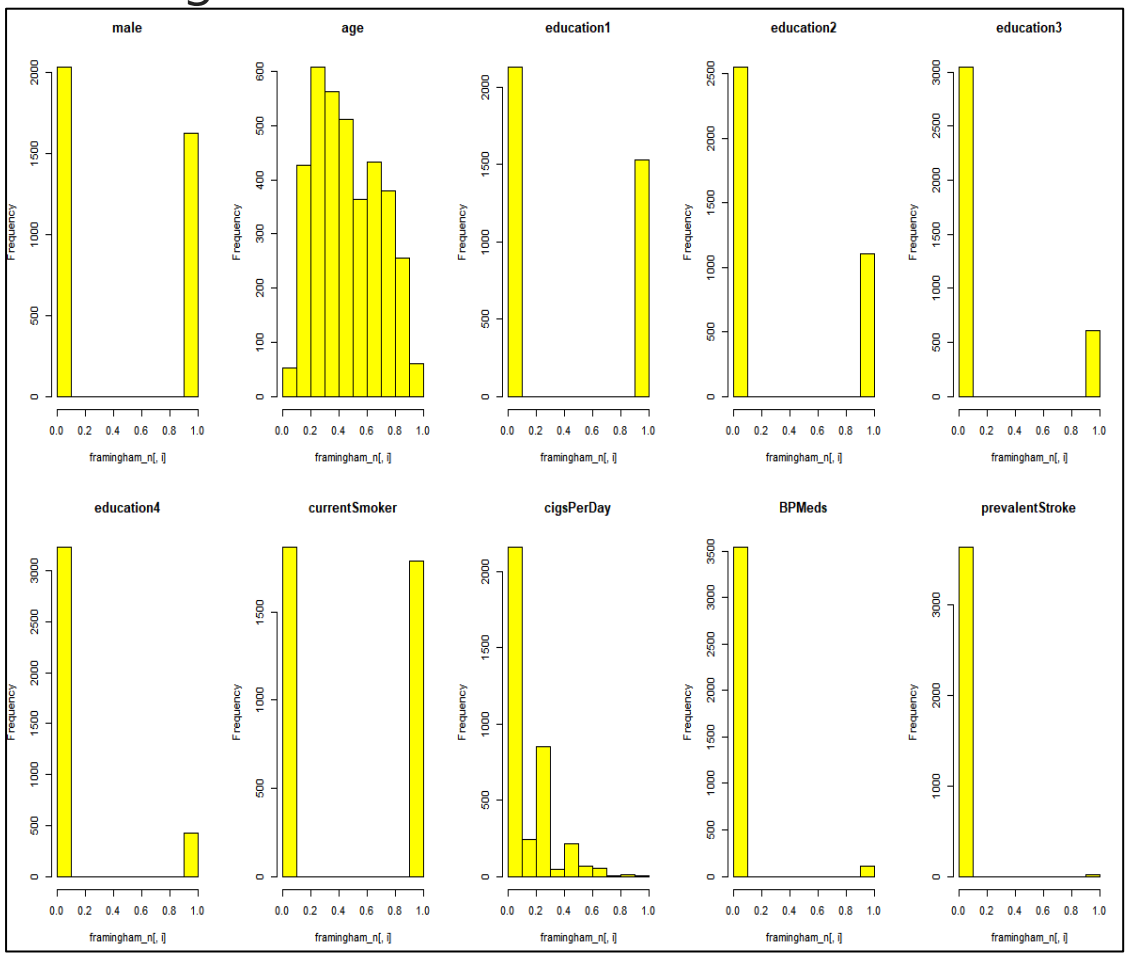
- 2.2 **(탐색적 데이터 분석) - 가시화**
- Histogram

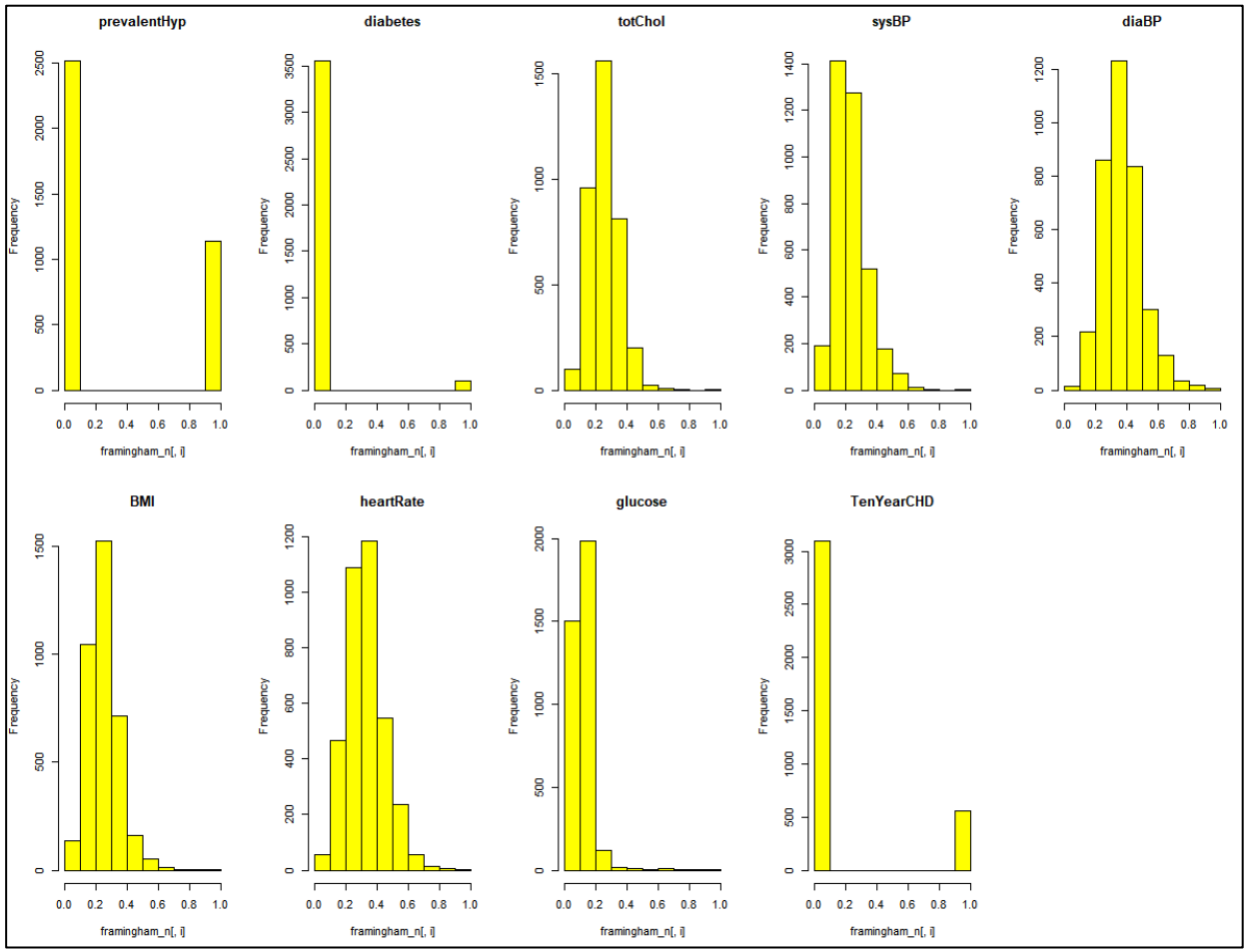
```
par(mfrow=c(2,5))
for(i in 1:10) {
  hist(framingham_n[,i],main=colnames(framingham_n)[i],col="yellow")
}
```

```
par(mfrow=c(2,5))
for(i in 11:19) {
  hist(framingham_n[,i],main=colnames(framingham_n)[i],col="yellow")
}
```

2. 로지스틱회귀분석 (TenYearCHD 예측)

- 2.2 **(탐색적 데이터 분석) - 가시화**
- Histogram



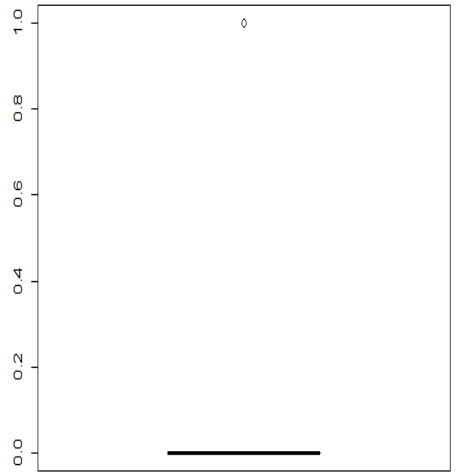


2. 로지스틱회귀분석 (TenYearCHD 예측)

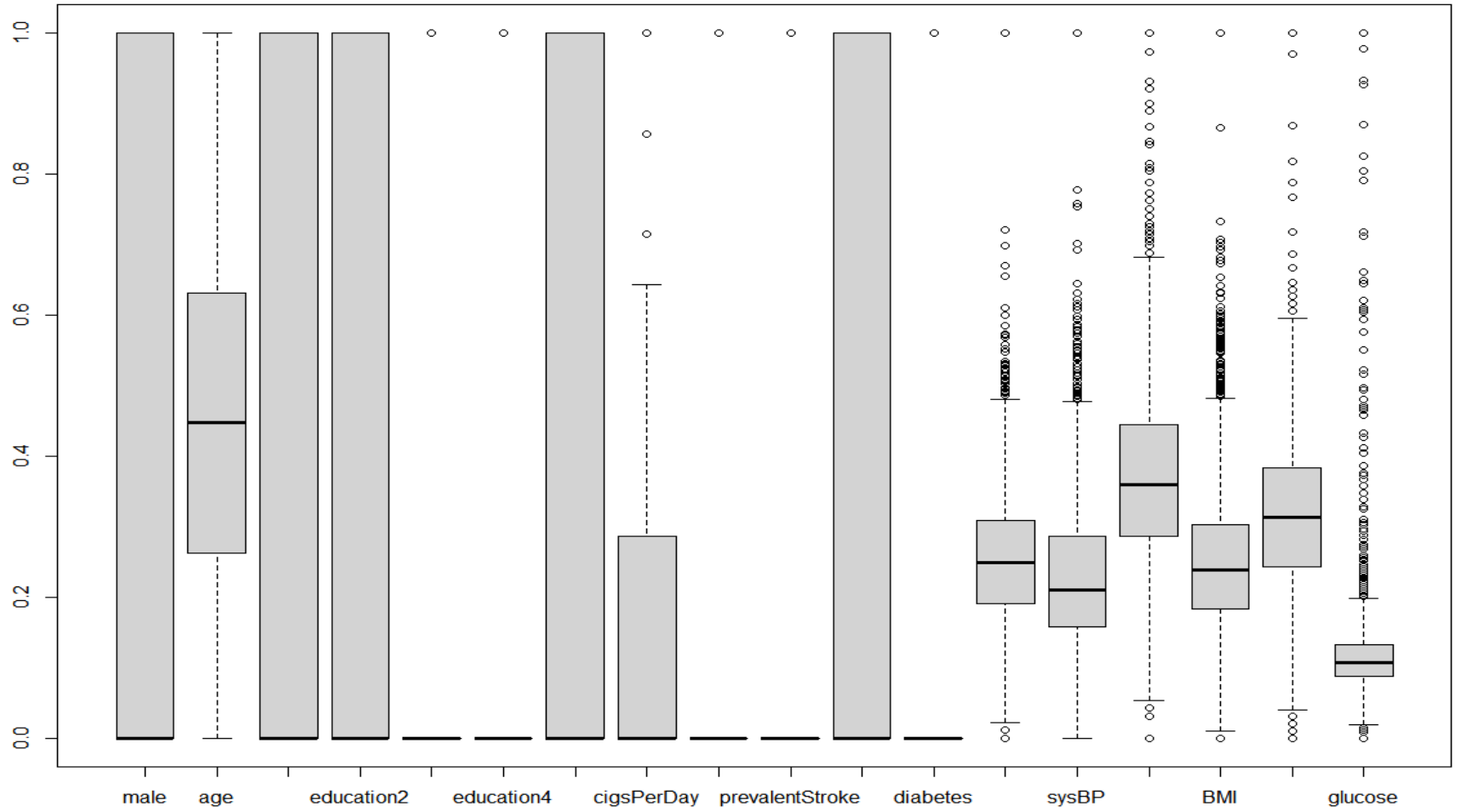
- **2.2 (탐색적 데이터 분석) - 가시화**
- Boxplot 가시화

```
par(mfrow=c(1,1))
boxplot(framingham_n[,c(1:18)])
```

```
boxplot(framingham_n[,c(19)])
```



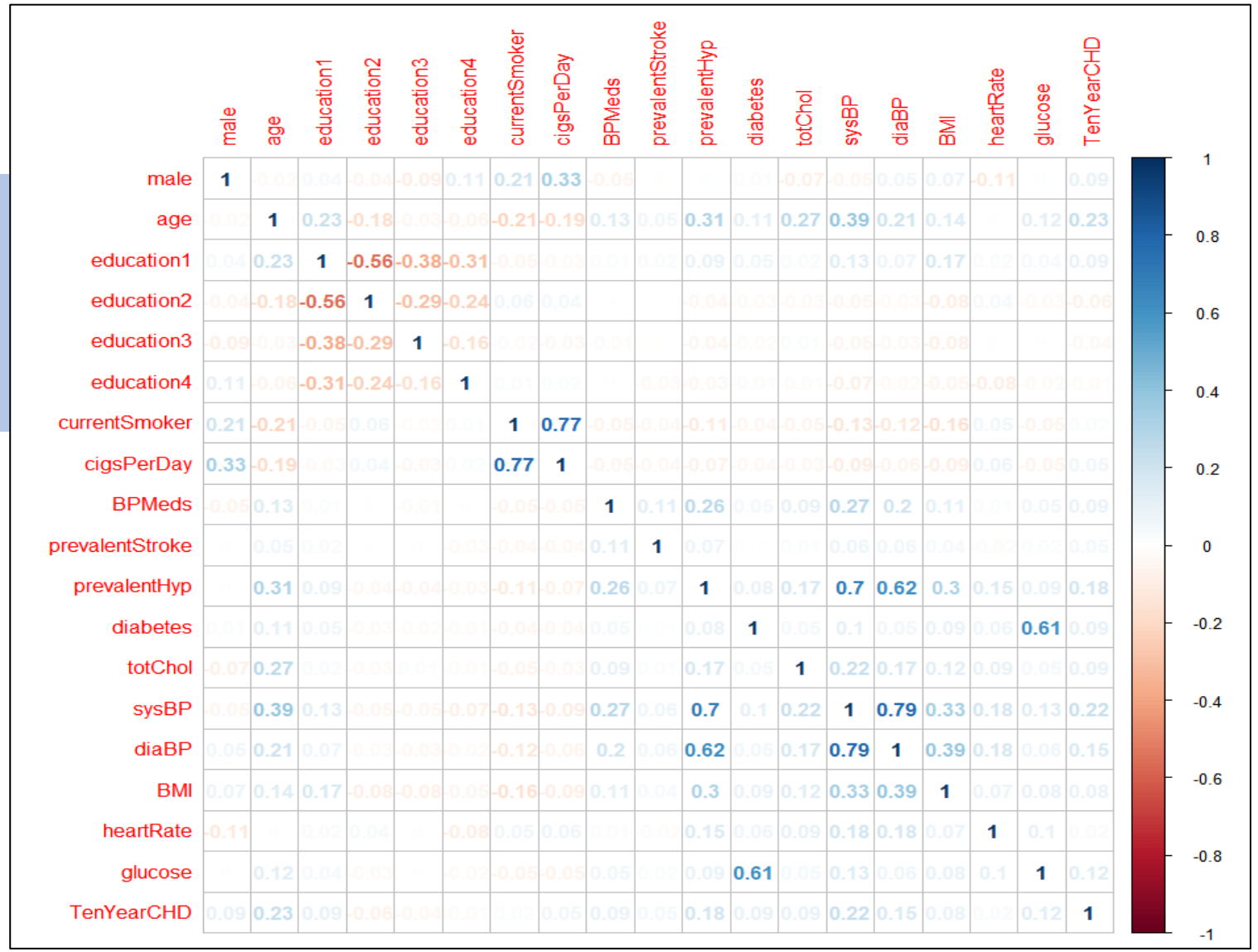
TenYearCHD



2. 로지스틱회귀분석 (TenYearCHD 예측)

- 2.2 **(탐색적 데이터 분석)**
- correlation matrix

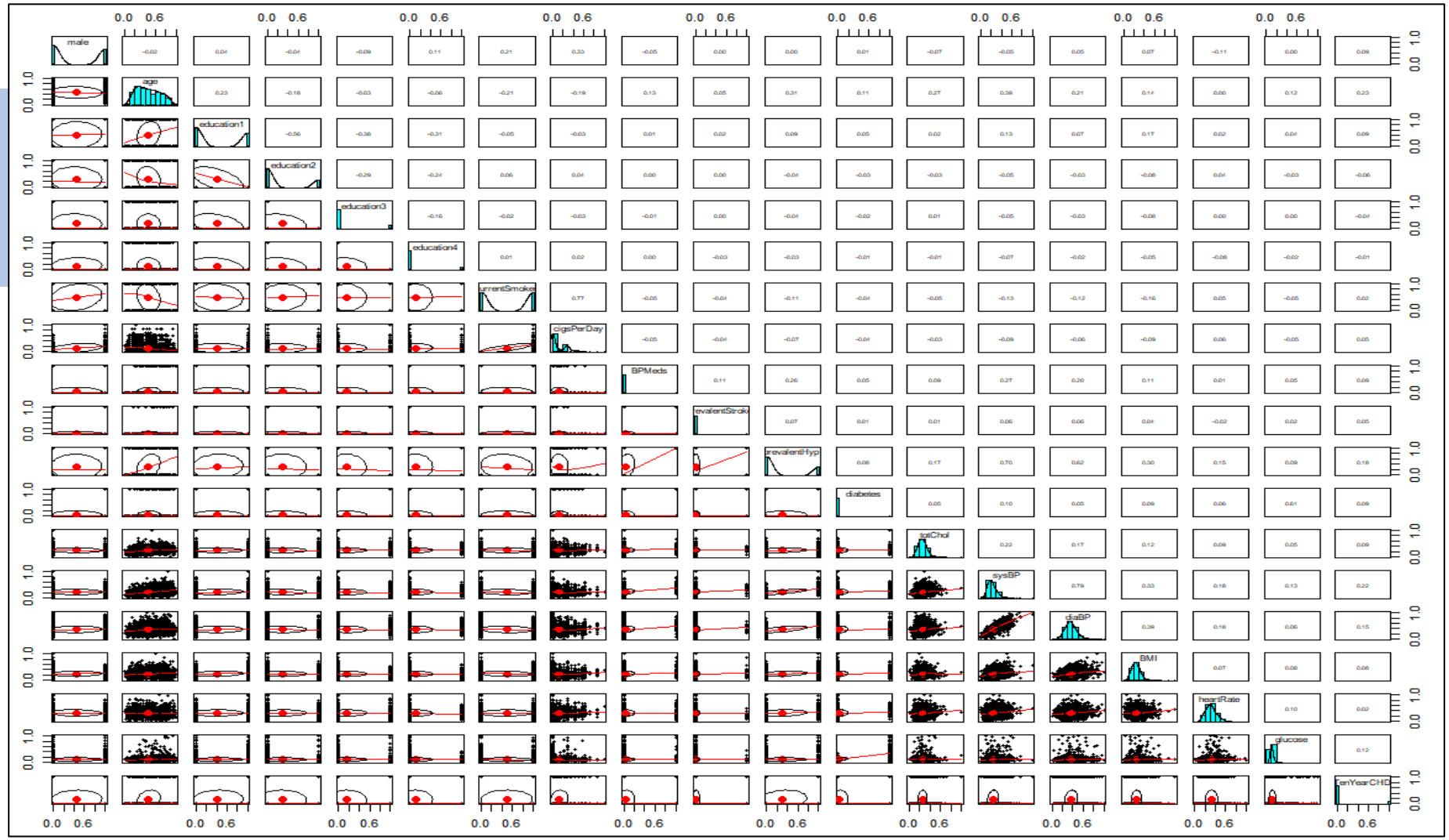
```
install.packages("corrplot")
library(corrplot)
par(mfrow=c(1,1))
cor_matrix=cor(framingham_n)
corrplot(cor_matrix,method="num")
```



2. 로지스틱회귀분석 (TenYearCHD 예측)

- 2.2 **(탐색적 데이터 분석)**
- Multi plots

```
install.packages('psych')
library(psych)
subset <- cbind
pairs.panels(framingham_n)
```



2. 로지스틱회귀분석 (TenYearCHD 예측)

유의수준 : 0.05

- 2.3 **(학습모델 구축) – 학습/테스트 구분**

```
set.seed(2022)
test_id <- sample(1:nrow(framingham_n),
round(nrow(framingham_n)*0.96))
framingham_n_train <- framingham_n[-test_id, ]
framingham_n_test <- framingham_n[test_id, ]

print("Training: ", str(nrow(framingham_n_train)))
print("Test: ", str(nrow(framingham_n_test)))
```

```
> print("Training: ", str(nrow(framingham_n_train)))
 int 146
[1] "Training: "
> print("Test: ", str(nrow(framingham_n_test)))
 int 3510
[1] "Test: "
```

2. 로지스틱회귀분석 (TenYearCHD 예측)

유의수준 : 0.05

- 2.3 **(학습모델 구축) - 해석**

```
logistic_model <- glm(TenYearCHD~.,
framingham_n_train, family = binomial())
summary(logistic_model)
```

```
Call:
glm(formula = TenYearCHD ~ ., family = binomial(), data =
 framingham_n_train)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-1.8152  -0.5008  -0.2518  -0.1311   3.1363

Coefficients: (1 not defined because of singularities)
                 Estimate Std. Error z value Pr(>|z|)
(Intercept)      -6.23342    2.41873  -2.577  0.00996 **
male              0.81823    0.70031   1.168  0.24265
age               4.29434    1.53111   2.805  0.00504 **
education1        0.92213    1.24013   0.744  0.45713
education2        1.07068    1.25645   0.852  0.39413
education3        2.08217    1.32299   1.574  0.11552
education4             NA         NA      NA       NA
currentSmoker     0.08889    1.09025   0.082  0.93502
cigsPerDay        2.57790    3.25610   0.792  0.42853
BPMeds            0.38804    2.22931   0.174  0.86182
prevalentStroke   2.78996    1.56347   1.784  0.07435 .
prevalentHyp     -1.66846    1.18771  -1.405  0.16009
diabetes         -6.10544    3.61731  -1.688  0.09144 .
totChol          -5.11588    3.94068  -1.298  0.19421
sysBP            11.69553    5.60404   2.087  0.03689 *
diaBP            -9.22307    3.81610  -2.417  0.01565 *
BMI               2.45345    3.56930   0.687  0.49185
heartRate        -2.08144    3.27182  -0.636  0.52466
glucose          20.43528    7.54315   2.709  0.00675 **
---
Signif. codes:
0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 120.265  on 145  degrees of freedom
Residual deviance:  83.441  on 128  degrees of freedom
AIC: 119.44

Number of Fisher Scoring iterations: 6
```

· 잔차

- 최솟값 : -1.8152 - 1사분위(25%위치) : -0.5008
  중앙값 : -0.2518 - 3사분위(75%위치) : -0.1311
  최댓값 : 3.1363

- 잔차 이탈도는 작을수록 좋고, 카이제곱분포를 따른다.

```
> #적합도 검정
> qchisq(0.95,df=128)
[1] 155.4047
```

- 잔차 이탈도가 83.441로 임계치보다 작으므로 모형은 적합하다.

2. 로지스틱회귀분석 (TenYearCHD 예측)

- 2.3 **(학습모델 구축) - 해석**

```
Call:
glm(formula = TenYearCHD ~ ., family = binomial(), data =
 framingham_n_train)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-1.8152  -0.5008  -0.2518  -0.1311   3.1363

Coefficients: (1 not defined because of singularities)
             Estimate Std. Error z value Pr(>|z|)
(Intercept)  -6.23342    2.41873  -2.577  0.00996 **
male          0.81823    0.70031   1.168  0.24265
age           4.29434    1.53111   2.805  0.00504 **
education1    0.92213    1.24013   0.744  0.45713
education2    1.07068    1.25645   0.852  0.39413
education3    2.08217    1.32299   1.574  0.11552
education4         NA         NA      NA       NA
currentSmoker 0.08889    1.09025   0.082  0.93502
cigsPerDay    2.57790    3.25610   0.792  0.42853
BPMeds        0.38804    2.22931   0.174  0.86182
prevalentStroke 2.78996  1.56347   1.784  0.07435 .
prevalentHyp -1.66846    1.18771  -1.405  0.16009
diabetes     -6.10544    3.61731  -1.688  0.09144 .
totChol      -5.11588    3.94068  -1.298  0.19421
sysBP        11.69553    5.60404   2.087  0.03689 *
diaBP        -9.22307    3.81610  -2.417  0.01565 *
BMI           2.45345    3.56930   0.687  0.49185
heartRate    -2.08144    3.27182  -0.636  0.52466
glucose      20.43528    7.54315   2.709  0.00675 **
---
Signif. codes:
0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 120.265  on 145  degrees of freedom
Residual deviance:  83.441  on 128  degrees of freedom
AIC: 119.44

Number of Fisher Scoring iterations: 6
```

유의수준 : 0.05

```
#결측치판단
sum(is.na(framingham_n$education4))
1] 0
```

· education4는 singularities때문에 NA로 나타남.
· male의 계수 = 0.81823, p-value > 0.05 유의미X
· age의 계수 = 4.29434, p-value < 0.05 유의미O
· education1의 계수 = 0.92213, p-value > 0.05 유의미X
· education2의 계수 = 1.07068, p-value > 0.05 유의미X
· education3의 계수 = 2.08217, p-value > 0.05 유의미X
· currentSmoker의 계수 = 0.08889, p-value > 0.05 유의미X
· cigsPerDay의 계수 = 2.57790, p-value > 0.05 유의미X
· BPMeds = 0.38804, p-value > 0.05 유의미X
· prevalentStroke의 계수 = 2.78996, p-value > 0.05 유의미X
· prevalentHyp의 계수 = -1.66846, p-value > 0.05 유의미X
· diabetes의 계수 = -6.10544, p-value > 0.05 유의미X
· totChol의 계수 = -5.11588, p-value > 0.05 유의미X
· sysBP의 계수 = 11.69553, p-value < 0.05 유의미O
· diaBP의 계수 = -9.22307, p-value < 0.05 유의미O
· BMI의 계수 = 2.45345, p-value > 0.05 유의미X
· heartRate의 계수 = -2.08144, p-value > 0.05 유의미X
· glucose의 계수 = 20.43528, p-value < 0.05 유의미O

따라서, 초록색 계수들로 TenYearCHD을 추정할 수 있다.

2. 로지스틱회귀분석 (TenYearCHD 예측)

- 2.4 **(로지스틱회귀 결과해석)**

· confusion matrix와 recall, precision, F1 measure측정

```
perf_eval <- function(cm){
  TPR = Recall = cm[2,2]/sum(cm[2,])
  Precision = cm[2,2]/sum(cm[,2])
  TNR = cm[1,1]/sum(cm[1,])
  ACC = sum(diag(cm)) / sum(cm)
  BCR = sqrt(TPR*TNR)
  F1 = 2 * Recall * Precision / (Recall + Precision)

  re <- data.frame(TPR = TPR,
                   Precision = Precision,
                   TNR = TNR,
                   ACC = ACC,
                   BCR = BCR,
                   F1 = F1)
  return(re)
}
```

```
pred_prob <- predict(logistic_model,
type="response", newdata = framingham_n_test)
pred_class <- rep(0, nrow(framingham_n_test))
pred_class[pred_prob > 0.5] <- 1
cm <- table(pred=pred_class,
actual=framingham_n_test$TenYearCHD)
cm
perf_eval(cm)
```

```
> perf_eval(cm)
        TPR Precision       TNR       ACC       BCR        F1
1 0.3174603 0.1492537 0.8600368 0.8210826 0.5225204 0.2030457
```

· 혼동행렬

```
          Reference
Prediction    0    1
         0 2802  456
         1  172   80
```

· recall = 0.3174603
· precision = 0.1492537
· F1 measure = 0.2030457

2. 로지스틱회귀분석 (TenYearCHD 예측)

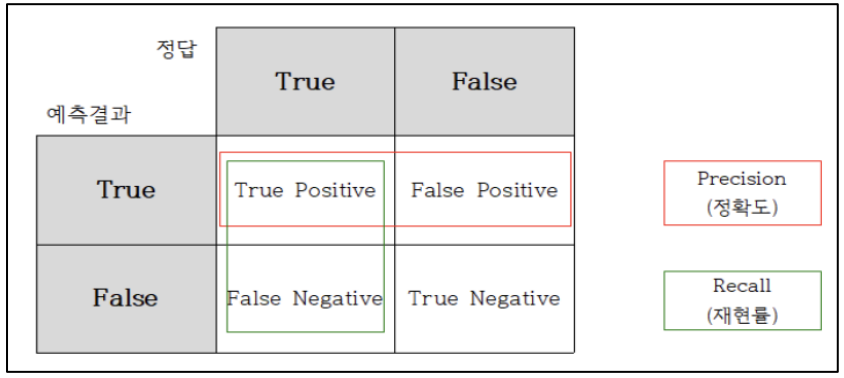
- 2.4 **(로지스틱회귀 결과해석)**

· confusion matrix와 recall, precision, F1 measure측정

```
install.packages("e1071")
library(e1071)
confusionMatrix(as.factor(pred_class),as.factor(framingham_n_test$TenYearCHD))
```

```
Confusion Matrix and Statistics

          Reference
Prediction    0    1
         0 2802  456
         1  172   80

               Accuracy : 0.8211
                 95% CI : (0.808, 0.8336)
    No Information Rate : 0.8473
    P-Value [Acc > NIR] : 1

                  Kappa : 0.1168

 Mcnemar's Test P-Value : <2e-16

            Sensitivity : 0.9422
            Specificity : 0.1493
         Pos Pred Value : 0.8600
         Neg Pred Value : 0.3175
             Prevalence : 0.8473
         Detection Rate : 0.7983
   Detection Prevalence : 0.9282
      Balanced Accuracy : 0.5457

       'Positive' Class : 0
```

· 혼동행렬



· 재현률

$$\frac{TP}{TP + FN}$$

· 정밀도

$$\frac{TP}{TP + FP}$$

· F1 measure

- 정밀도와 재현율의 조화평균

F1= s* Precision*Recall/(Precision + recall)

2. 로지스틱회귀분석 (TenYearCHD 예측)

- 2.5 **(변수선택법 사용해 정확도 향상)**

· p-value가 유의하지 않은 변수 차례로 제거

```
logistic_model <- glm(TenYearCHD ~ age + sysBP + diaBP + glucose, framingham_n_train, family = binomial())
summary(logistic_model)
```

```
Call:
glm(formula = TenYearCHD ~ age + sysBP + diaBP + glucose, family = binomial
(),
    data = framingham_n_train)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-1.1812  -0.5630  -0.3815  -0.2686   2.5764

Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept)  -3.3043     0.9562  -3.456 0.000549 ***
age           2.9707     1.1311   2.626 0.008630 **
sysBP         5.2319     3.1247   1.674 0.094063 .
diaBP        -6.3262     2.7473  -2.303 0.021296 *
glucose       7.7965     4.1639   1.872 0.061151 .
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 120.26  on 145  degrees of freedom
Residual deviance: 101.91  on 141  degrees of freedom
AIC: 111.91

Number of Fisher Scoring iterations: 5
```

```
> perf_eval(cm)
        TPR  Precision       TNR       ACC       BCR        F1
1 0.4528302 0.08955224 0.8566392 0.8444444 0.6228259 0.1495327
```

유의미한 변수(age, sysBP, diaBP, glucose)로만

⇒ **기존 모델 보다 F1값 낮아짐.**

2. 로지스틱회귀분석 (TenYearCHD 예측)

- 2.5 **(변수선택법 사용해 정확도 향상)**

· Forward selection, Backward elimination, 외 Stepwise selection 선택적 사용

```
model_fwd <- step(glm(TenYearCHD ~ 1, framingham_n_train,
                      family = binomial()),
                  direction = "forward", trace = 0,
                  scope = formula(logistic_model))

pred_prob <- predict(model_fwd, framingham_n_test, type="response")
pred_class <- rep(0, nrow(framingham_n_test))
pred_class[pred_prob > 0.5] <- 1
cm <- table(pred=pred_class, actual=framingham_n_test$TenYearCHD)
perf_eval(cm)
```

	TPR	Precision	TNR	ACC	BCR	F1
1	0.3470588	0.1100746	0.8571856	0.8324786	0.54543	0.1671388

Forward selection

⇒ **기존 모델 보다 F1값 낮아짐.**

2. 로지스틱회귀분석 (TenYearCHD 예측)

- 2.5 **(변수선택법 사용해 정확도 향상)**

· Forward selection, Backward elimination, 외 Stepwise selection 선택적 사용

```
model_bwd <- step(glm(TenYearCHD ~ ., framingham_n_train,
                      family = binomial()),
                  direction = "backward", trace = 0,
                  scope = list(lower=TenYearCHD ~ 1, upper = formula(logistic_model)))

pred_prob <- predict(model_bwd, framingham_n_test, type="response")
pred_class <- rep(0, nrow(framingham_n_test))
pred_class[pred_prob > 0.5] <- 1
cm <- table(pred=pred_class, actual=framingham_n_test$TenYearCHD)
perf_eval(cm)
```

```
        TPR Precision       TNR       ACC       BCR        F1
1 0.3319149 0.1455224 0.8601527 0.8247863 0.5343196 0.2023346
```

Backward elimination

⇒ **기존 모델 보다 F1 높아짐.**

2. 로지스틱회귀분석 (TenYearCHD 예측)

- 2.5 **(변수선택법 사용해 정확도 향상)**

· Forward selection, Backward elimination, 외 Stepwise selection 선택적 사용

```
model_step <- step(glm(TenYearCHD ~ ., framingham_n_train,
                       family = binomial()), direction = "both", trace = 0,
                   scope = list(lower=TenYearCHD ~ 1, upper = formula(logistic_model)))

pred_prob <- predict(model_step, framingham_n_test, type="response")
pred_class <- rep(0, nrow(framingham_n_test))
pred_class[pred_prob > 0.5] <- 1
cm <- table(pred=pred_class, actual=framingham_n_test$TenYearCHD)
perf_eval(cm)
```

```
       TPR Precision       TNR       ACC       BCR        F1
1 0.3319149 0.1455224 0.8601527 0.8247863 0.5343196 0.2023346
```

Stepwise selection

⇒ **기존 모델 보다 F1값 높아짐.**