lab01-birthwt-homework.knit

자료읽기

```
library(tidyverse)
```

```
## -- Attaching packages --------------------------------------- tidyverse 1.3.1 --
```

```
## v ggplot2 3.3.5     v purrr   0.3.4
## v tibble  3.1.4     v dplyr   1.0.7
## v tidyr   1.1.3     v stringr 1.4.0
## v readr   2.0.1     v forcats 0.5.1
```

```
## -- Conflicts ------------------------------------------ tidyverse_conflicts() --
## x dplyr::filter() masks stats::filter()
## x dplyr::lag()    masks stats::lag()
```

```
library(caret)
```

```
## Warning: 패키지 'caret'는 R 버전 4.1.2에서 작성되었습니다
```

```
## 필요한 패키지를 로딩중입니다: lattice
```

```
##
## 다음의 패키지를 부착합니다: 'caret'
```

```
## The following object is masked from 'package:purrr':
##
##     lift
```

```
DF <- read.csv('C:₩₩WORK₩₩data₩₩birthwt4times.csv')
DF <-
DF %>%
mutate(
 low = factor(low),
 lwtkg = round(lwt*0.453592,1),
 race = factor(race))
DF$lwt <- NULL
```

자료분할

```
TR <- DF[seq(1, nrow(DF), 2),]
dim(TR)
```

```
## [1] 378  11
```

```
TS <- DF[seq(2, nrow(DF), 2),]
dim(TS)
```

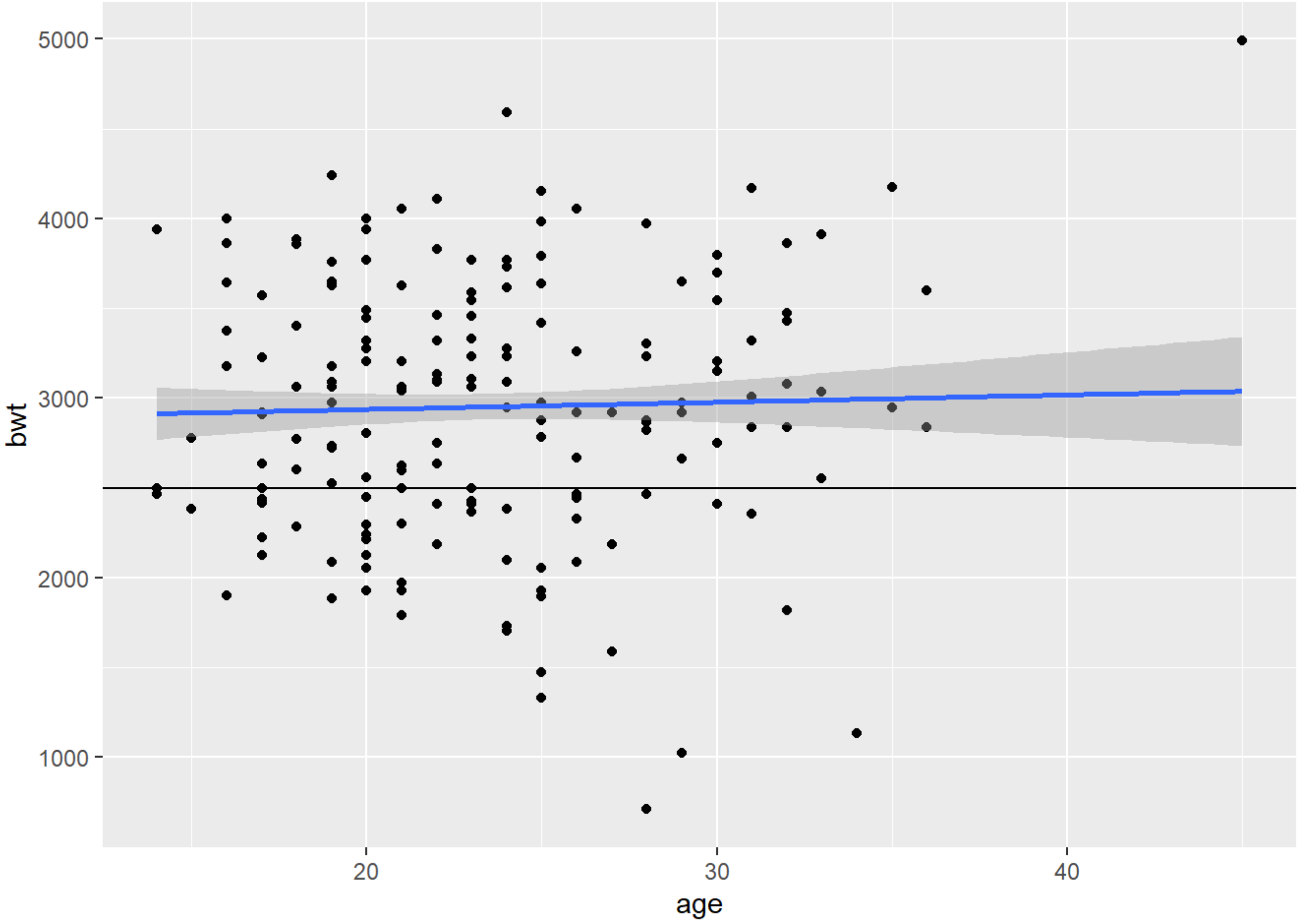
```
## [1] 378  11
```

1. 선형회귀 모형

age vs bwt

```
ggplot(TR, aes(x=age, y=bwt)) +
geom_point() +
 geom_hline(yintercept=2500) +
 geom_smooth(method='lm')
```

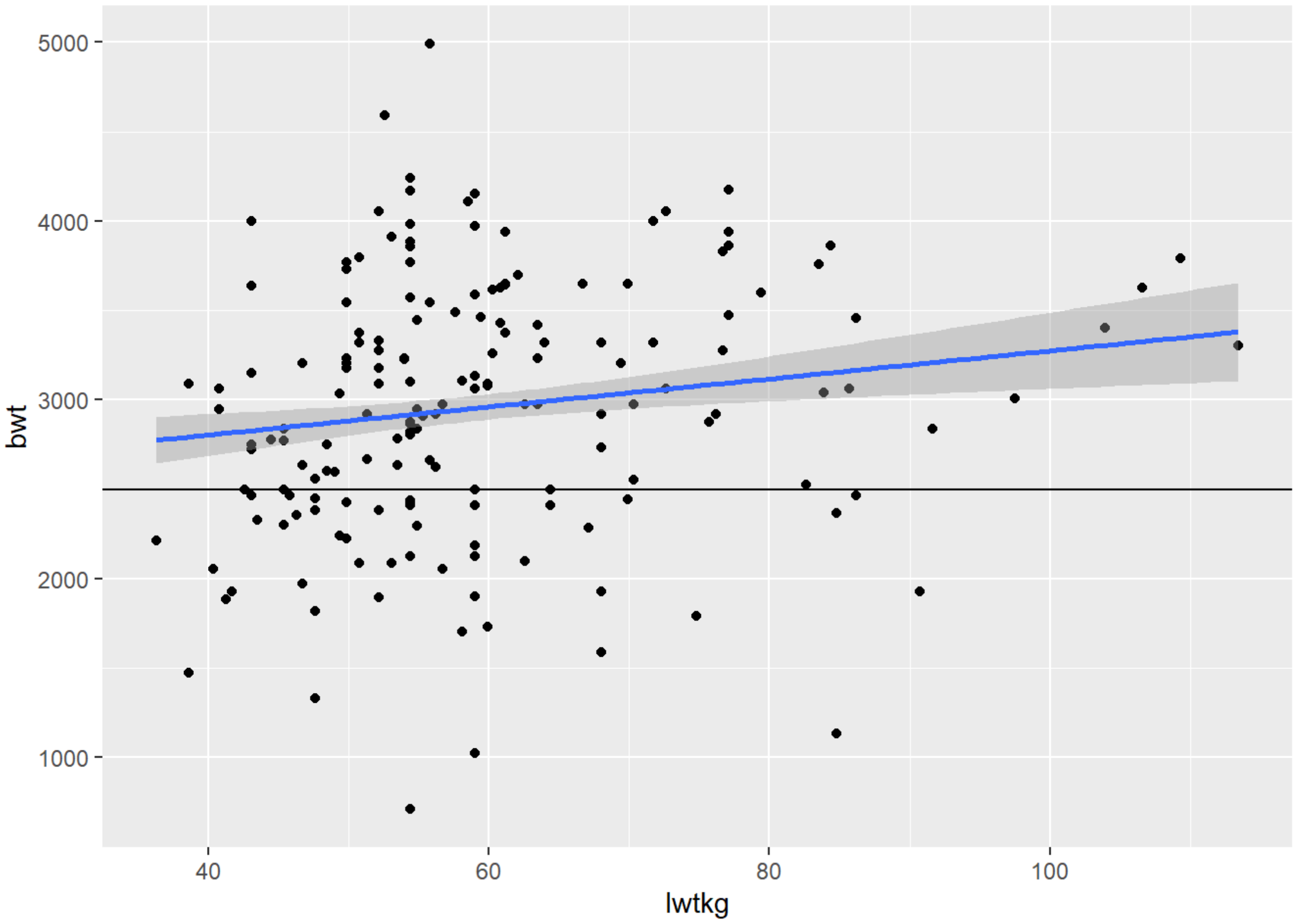
```
## `geom_smooth()` using formula 'y ~ x'
```



lwtkg vs bwt

```
ggplot(TR, aes(x=lwtkg, y=bwt)) +
geom_point() +
 geom_hline(yintercept=2500) +
 geom_smooth(method='lm')
```

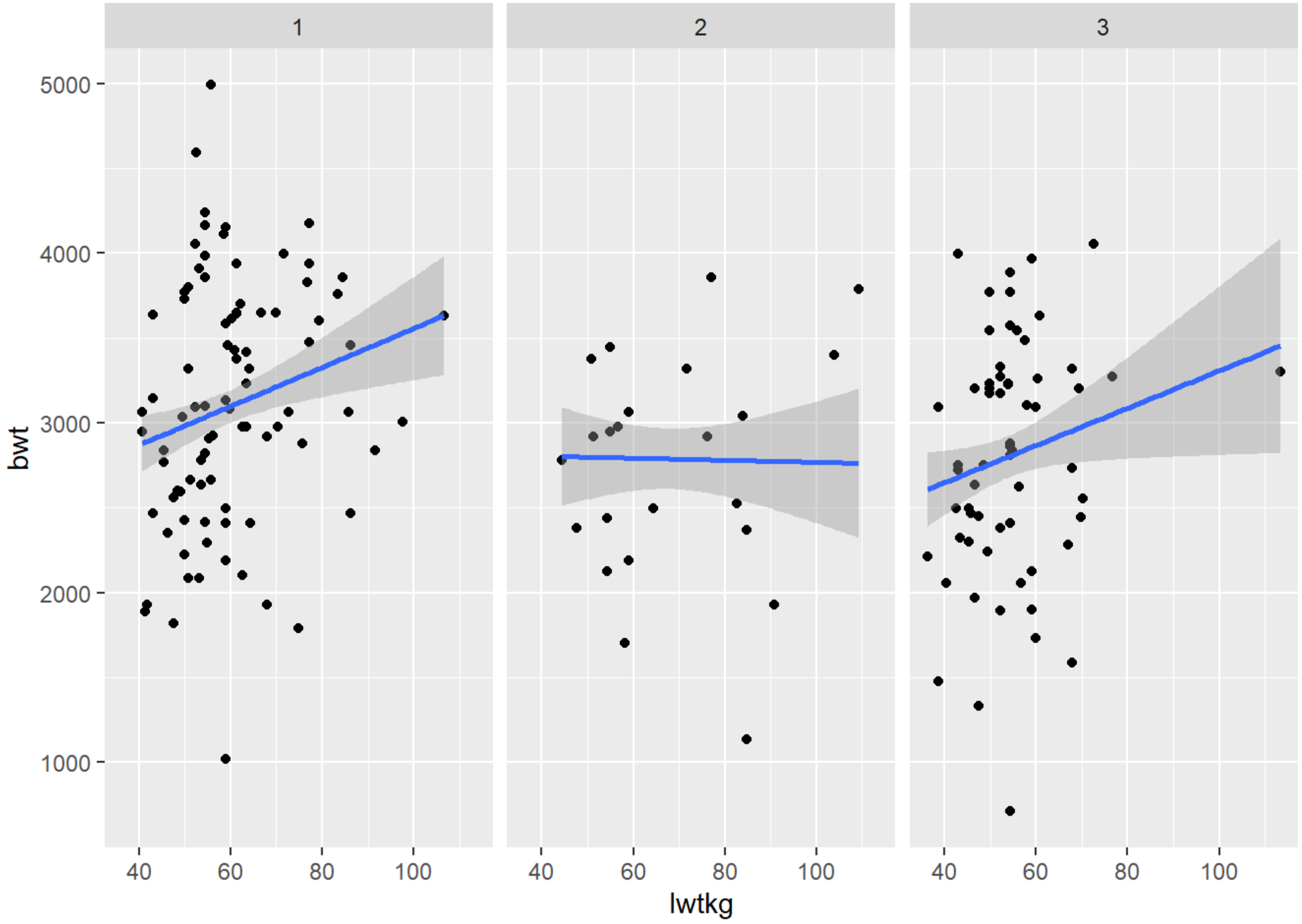
```
## `geom_smooth()` using formula 'y ~ x'
```



lwtkg vs bwt | race

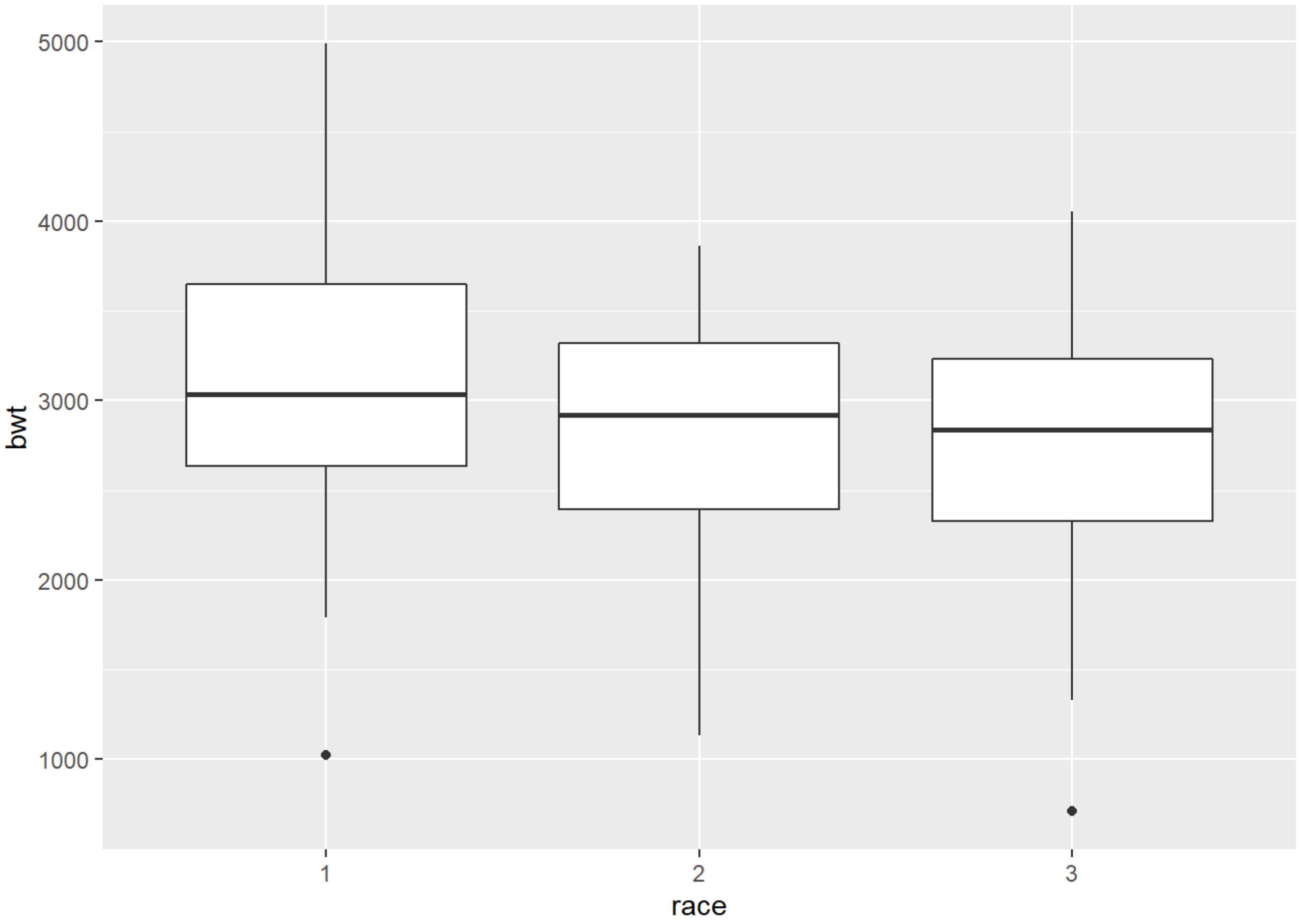
```
ggplot(TR, aes(x=lwtkg, y=bwt)) +
geom_point() +
 geom_smooth(method='lm') +
 facet_wrap(~race)
```

```
## `geom_smooth()` using formula 'y ~ x'
```

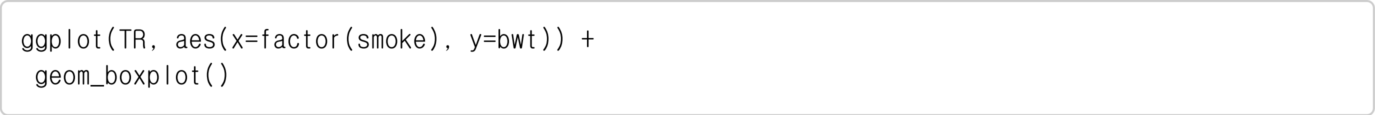


race vs bwt

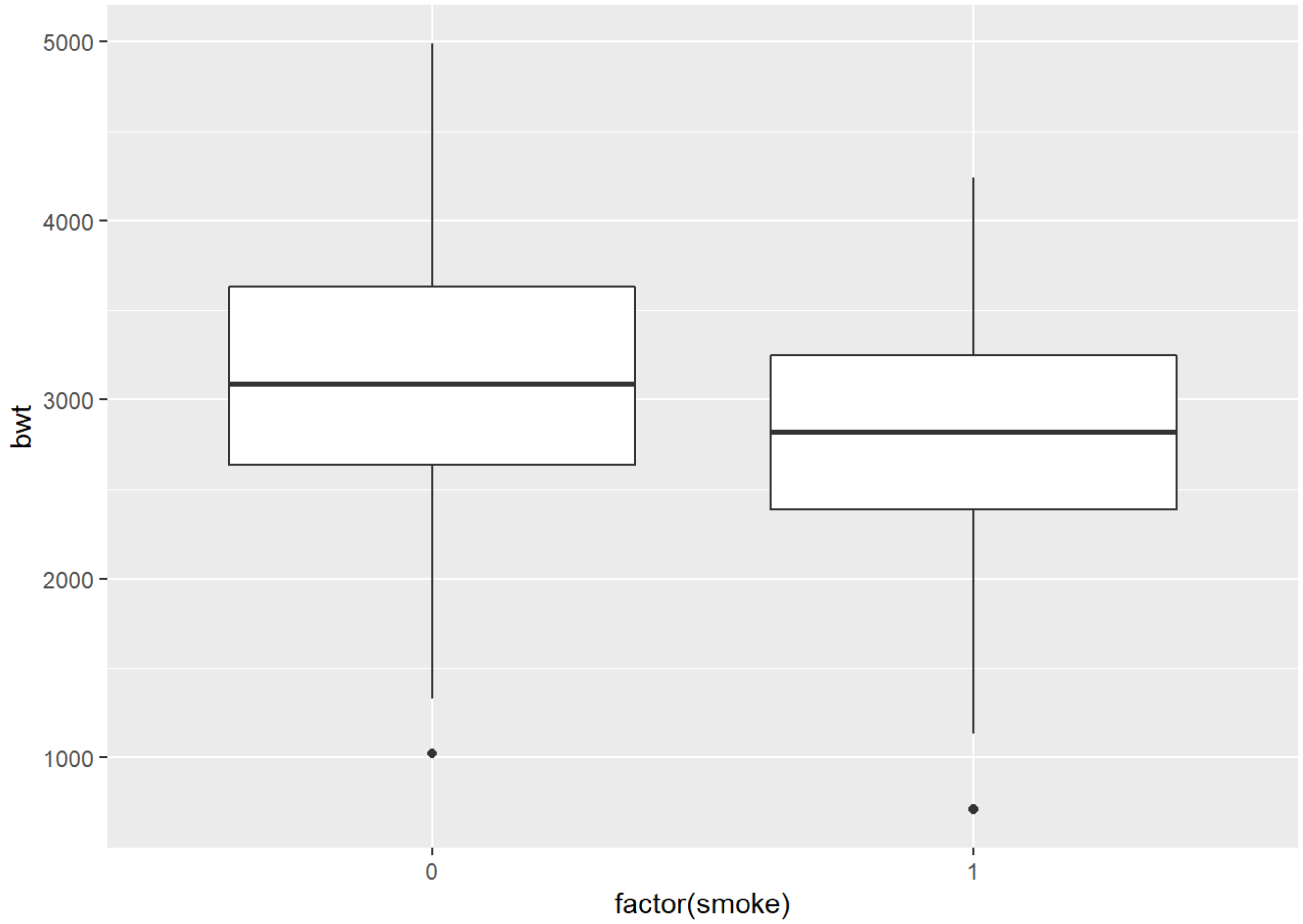
```
ggplot(TR, aes(x=race, y=bwt)) +
 geom_boxplot()
```



smoke vs bwt



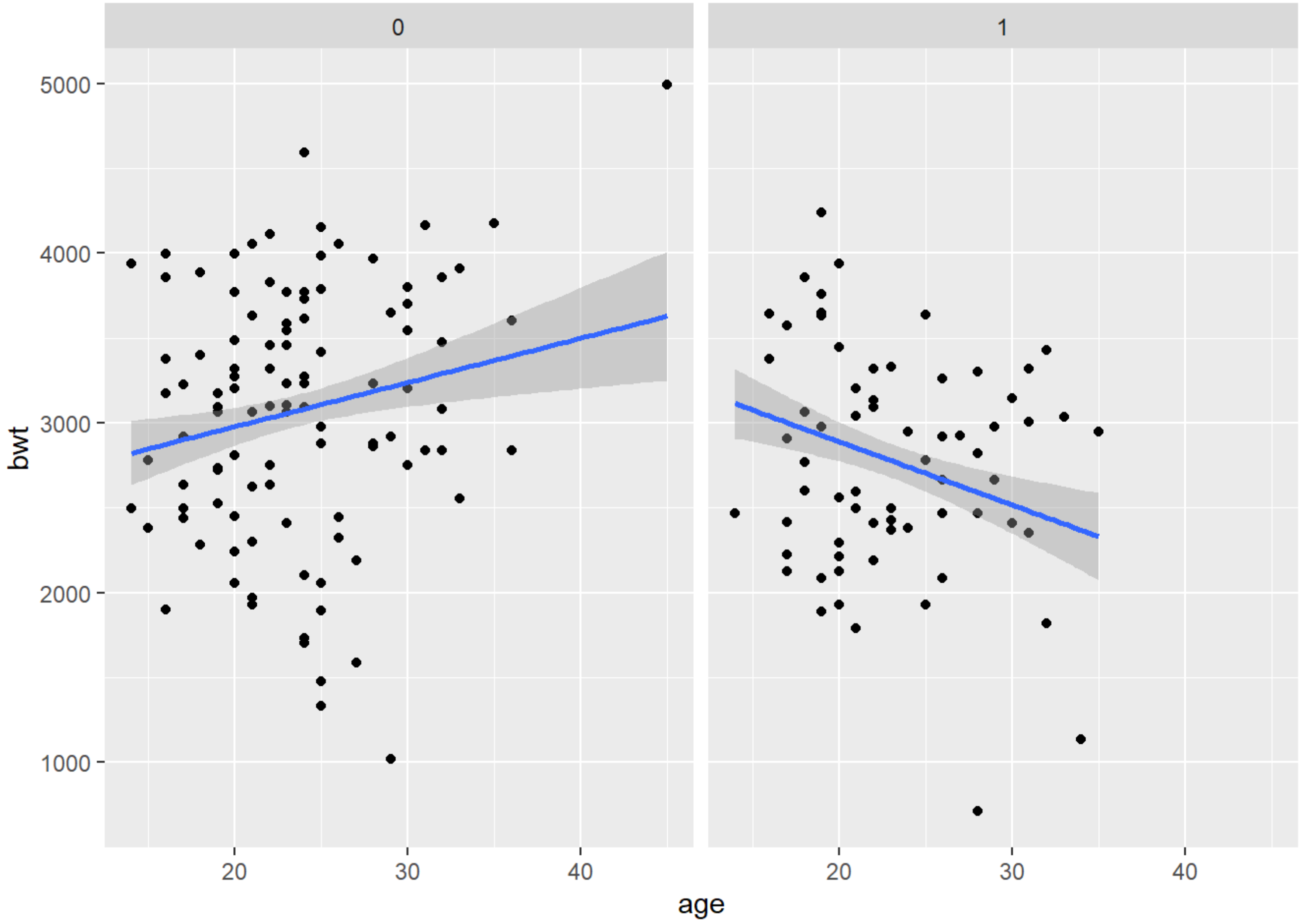
```
ggplot(TR, aes(x=factor(smoke), y=bwt)) +
 geom_boxplot()
```



smoke vs bwt | smoke

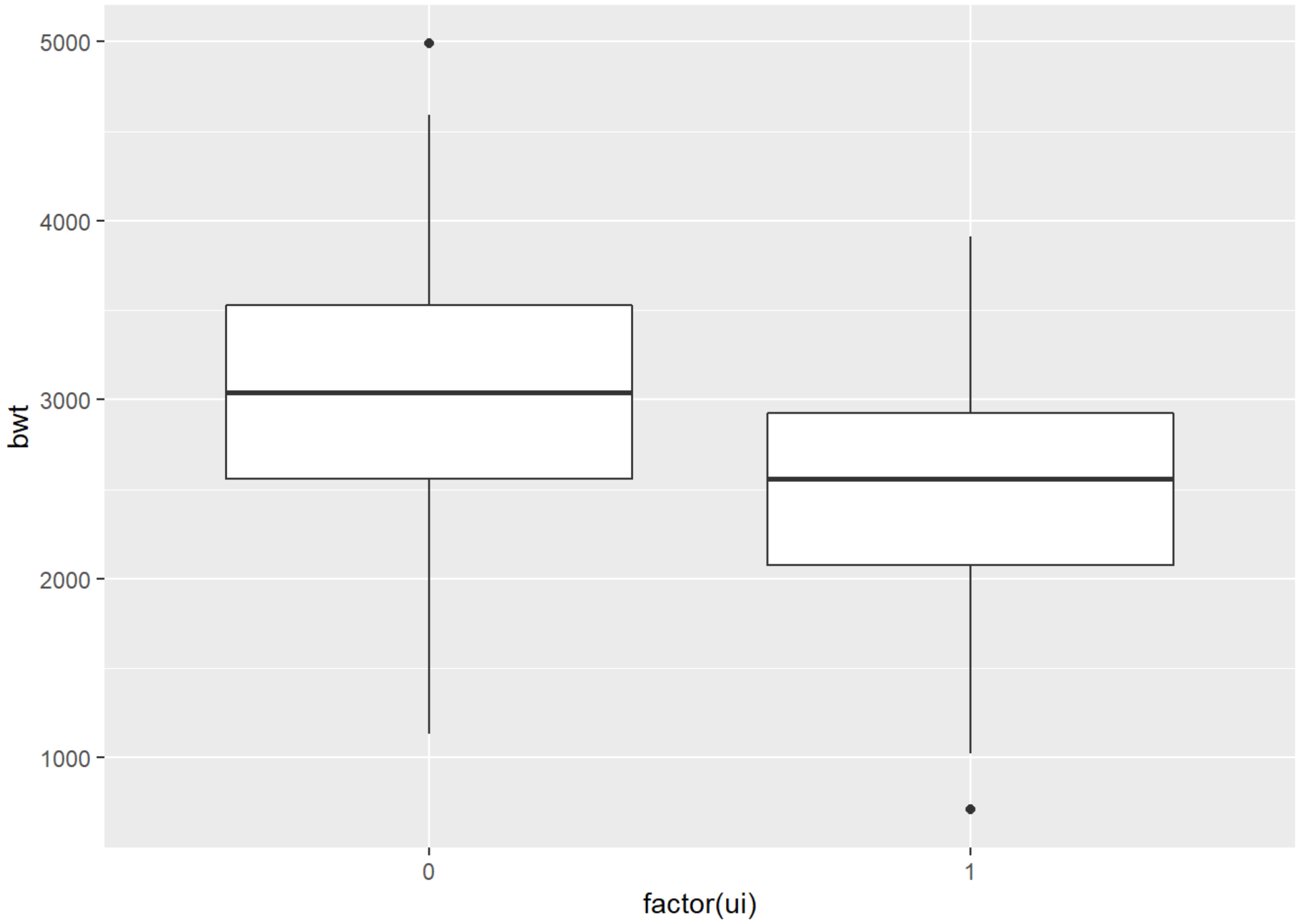
```
ggplot(TR, aes(x=age, y=bwt)) +
geom_point() +
 geom_smooth(method='lm') +
 facet_wrap(~smoke)
```

```
## `geom_smooth()` using formula 'y ~ x'
```



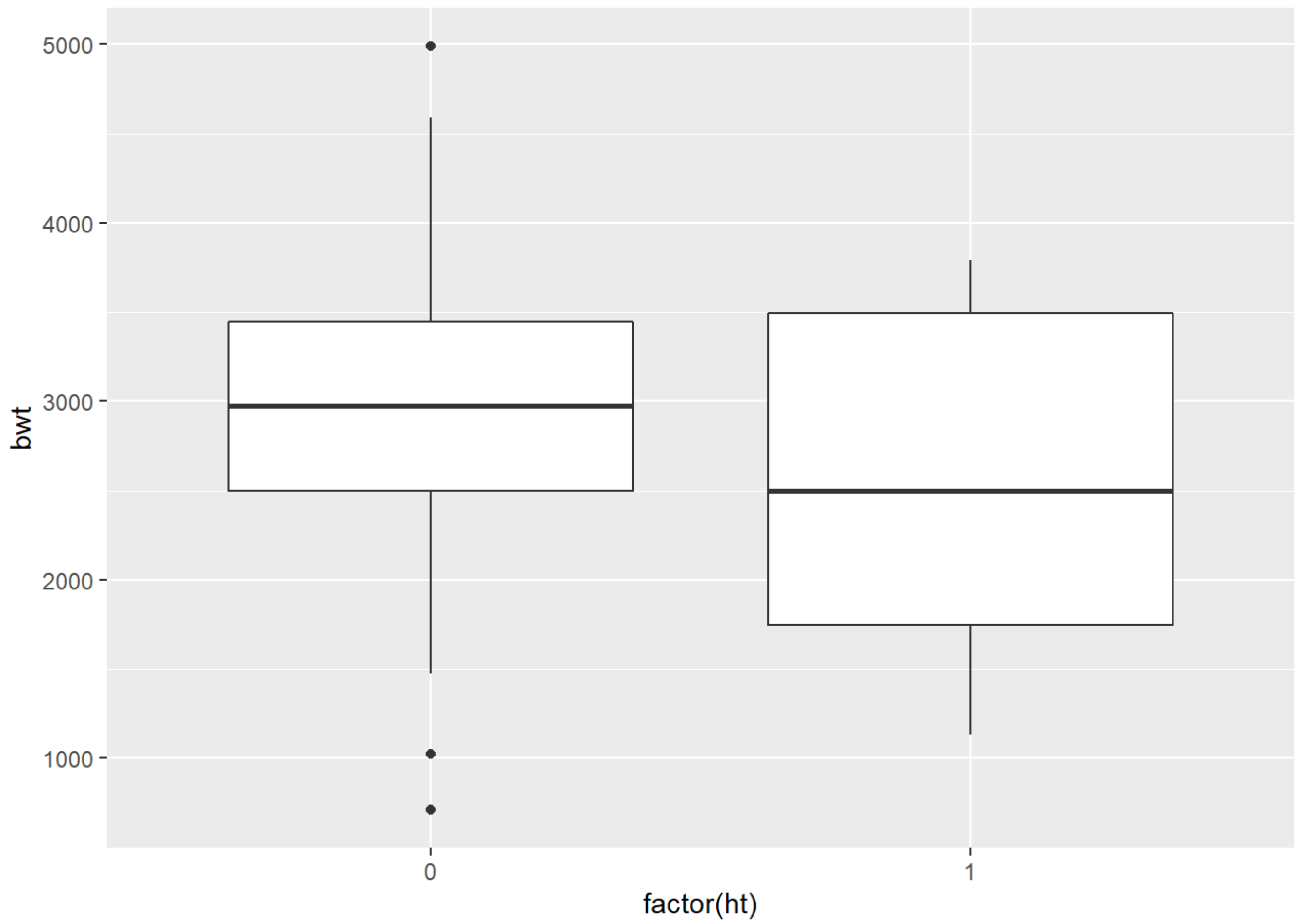
ui vs bwt

```
ggplot(TR, aes(x=factor(ui), y=bwt)) +
 geom_boxplot()
```



ht vs bwt

```
ggplot(TR, aes(x=factor(ht), y=bwt)) +
 geom_boxplot()
```



모형적합

- 종속변수(y): bwt
- 독립변수: age, ftv, ptl, race, smoke, ht, ui, lwtkg

```
Rlm <- lm(bwt ~ age+ftv+ptl+race+smoke+ht+ui+lwtkg, data=TR)
summary(Rlm)
```

```
##
## Call:
## lm(formula = bwt ~ age + ftv + ptl + race + smoke + ht + ui +
##     lwtkg, data = TR)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -1764.09  -408.30    34.75   403.55  1906.63
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 3253.286    210.830  15.431  < 2e-16 ***
## age          -13.826      6.566  -2.106 0.035891 *
## ftv          -17.028     32.346  -0.526 0.598914
## ptl           69.538     67.750   1.026 0.305381
## race2       -482.914    102.111  -4.729 3.22e-06 ***
## race3       -405.009     78.490  -5.160 4.05e-07 ***
## smoke       -403.857     71.581  -5.642 3.36e-08 ***
## ht          -561.130    143.300  -3.916 0.000107 ***
## ui          -529.388     97.343  -5.438 9.82e-08 ***
## lwtkg          8.410      2.520   3.337 0.000934 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 621.7 on 368 degrees of freedom
## Multiple R-squared:  0.2407, Adjusted R-squared:  0.2222
## F-statistic: 12.96 on 9 and 368 DF,  p-value: < 2.2e-16
```

```
TROUT <-
 TR %>%
 mutate(
 yh=predict(Rlm),
 e=residuals(Rlm))
head(TROUT)
```

```
##     id low  bwt age ftv race ptl smoke ht ui lwtkg       yh         e
## 1  284   0 3643  16   0    1   0     1  0  0  61.2 3142.923 500.07680
## 3  623   0 3175  16   0    3   0     0  0  0  49.9 3046.734 128.26628
## 5  400   0 2835  31   3    1   0     0  0  1  45.4 2626.029 208.97145
## 7  103   0 3770  24   0    3   1     0  0  0  49.9 3005.661 764.33935
## 9  602   0 2977  25   0    2   0     0  0  0  56.7 2901.581  75.41884
## 11  79   0 3444  20   0    2   0     1  0  0  54.9 2551.718 892.28246
```

```
mean(TROUT$e^2) # MSE
```

```
## [1] 376236.7
```

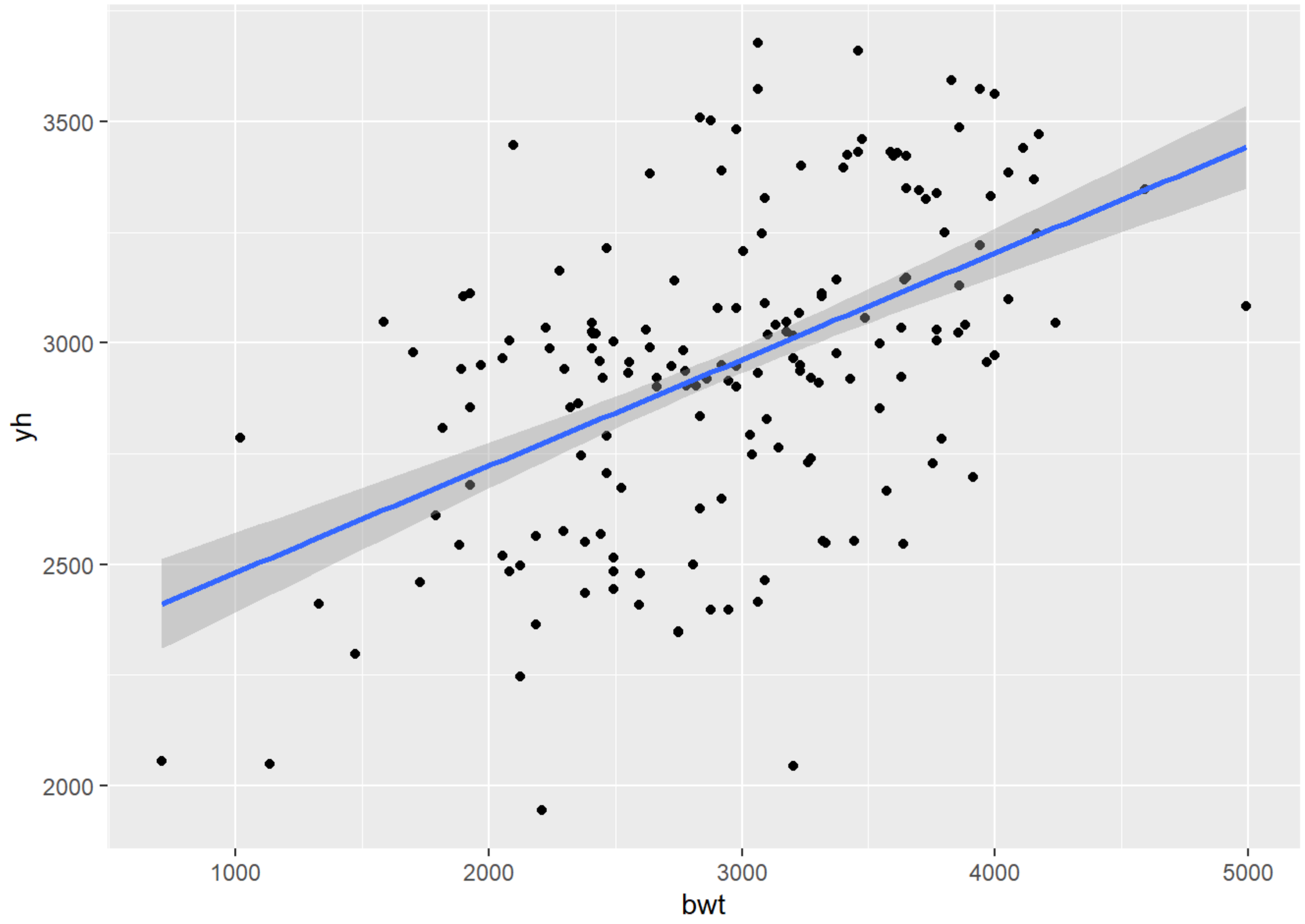
```
mean(abs(TROUT$e)) # MAE
```

```
## [1] 492.1593
```

모형검토(TR)

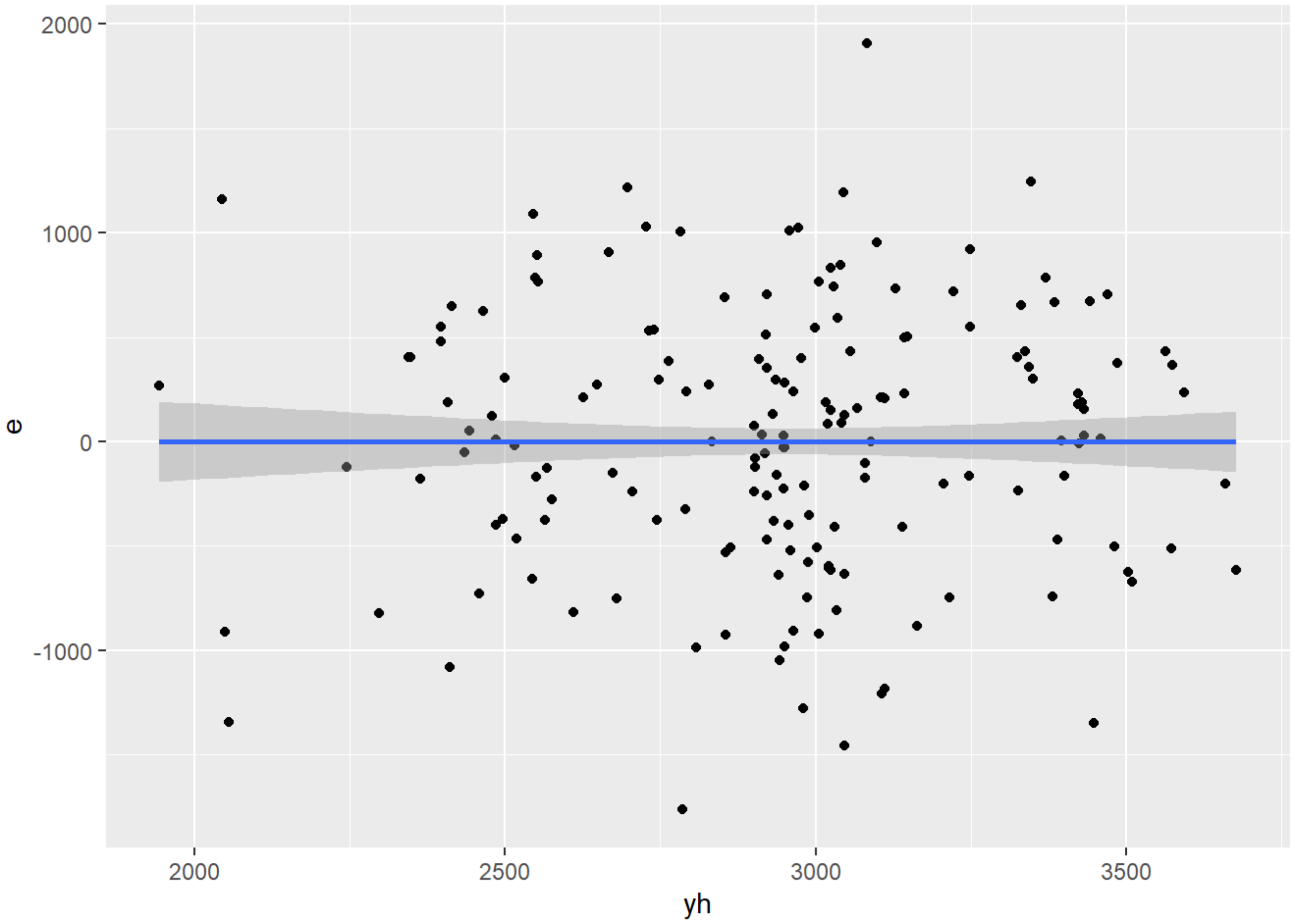
```
ggplot(TROUT, aes(x=bwt, y=yh)) +
geom_point() +
geom_smooth(method='lm')
```

```
## `geom_smooth()` using formula 'y ~ x'
```



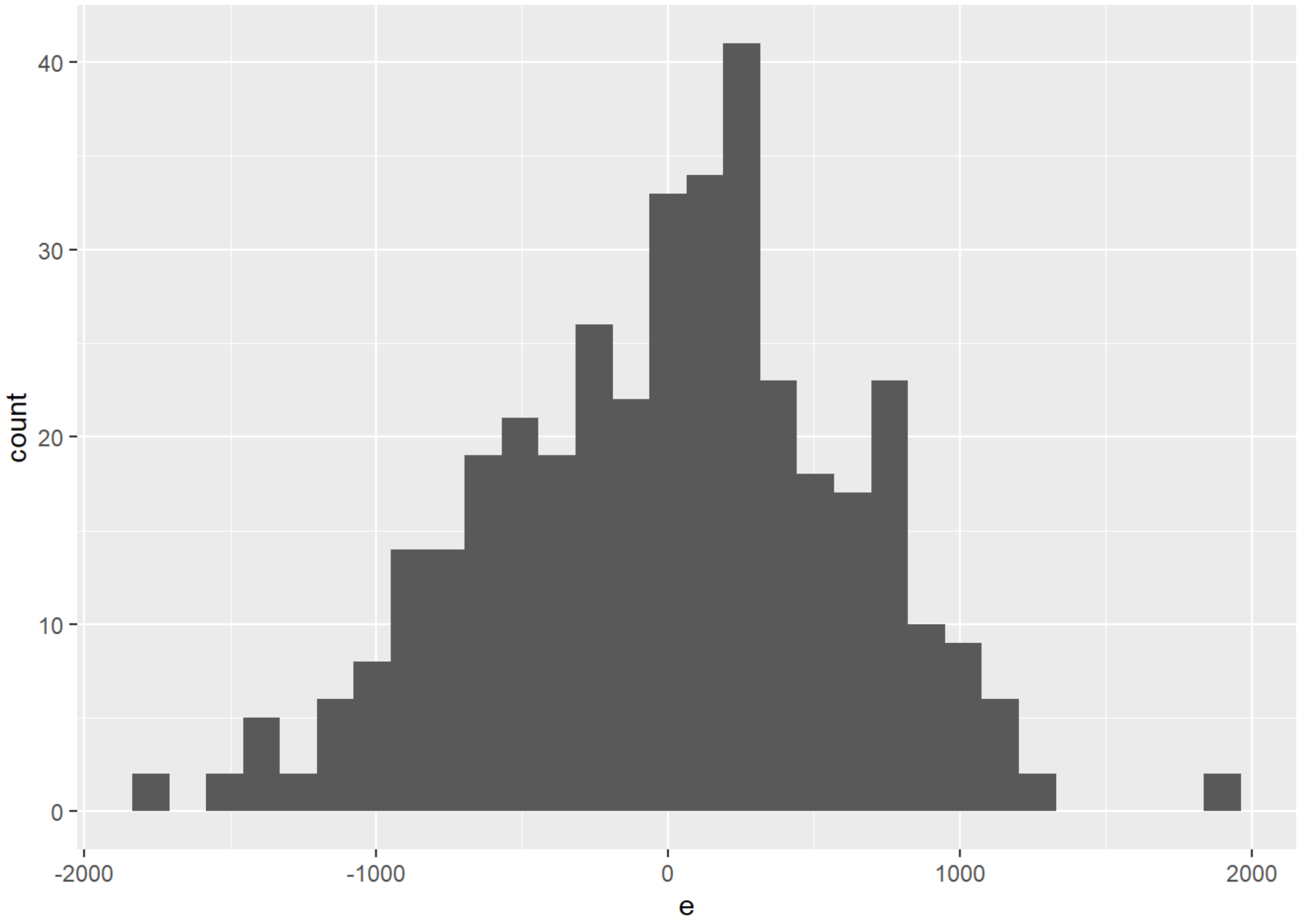
```
ggplot(TROUT, aes(x=yh, y=e)) +
geom_point() +
 geom_smooth(method='lm')
```

```
## `geom_smooth()` using formula 'y ~ x'
```



```
ggplot(TROUT, aes(x=e)) +
geom_histogram()
```

```
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```



```
TROUT %>% summarize(mn=mean(e), sd=sd(e), min=min(e), max=max(e))
```

```
##             mn       sd       min     max
## 1 3.429587e-14 614.1943 -1764.091 1906.63
```

```
summary(TROUT$e)
```

```
##     Min.  1st Qu.   Median     Mean  3rd Qu.     Max. 
## -1764.09  -408.30    34.75     0.00   403.55  1906.63
```

모형평가(TS)

```
TSOUT <-
TS %>%
mutate(yh=predict(Rlm, TS), e=bwt-yh)
head(TSOUT)
```

```
##     id low  bwt age ftv race ptl smoke ht ui lwtkg       yh          e
## 2  101   0 3728  24   1    1   0     0  0  0  49.9 3324.103  403.89653
## 4  645   0 3430  32   4    1   1     1  0  0  60.8 2919.764  510.23603
## 6   98   0 3651  19   0    1   0     1  0  0  66.7 3147.701  503.29884
## 8  726   1 2187  27   0    2   0     0  0  1  59.0 2363.884 -176.88449
## 10 326   1 1588  23   1    3   0     0  0  1  44.0 2353.912 -765.91184
## 12 270   0 3460  22   1    1   0     0  0  0  59.4 3431.655   28.34472
```

```
mean(TSOUT$e^2) # MSE
```

```
## [1] 437372.7
```

```
mean(abs(TSOUT$e)) # MAE
```

```
## [1] 539.0403
```

```
TSOUT %>% summarize(mn=mean(e), sd=sd(e), min=min(e), max=max(e))
```

```
##          mn       sd       min     max
## 1 -1.103086 662.2172 -1764.091 1906.63
```

```
summary(TSOUT$e)
```

```
##      Min.   1st Qu.    Median      Mean   3rd Qu.      Max.
## -1764.091  -470.794    13.717    -1.103   495.133  1906.630
```

스코어

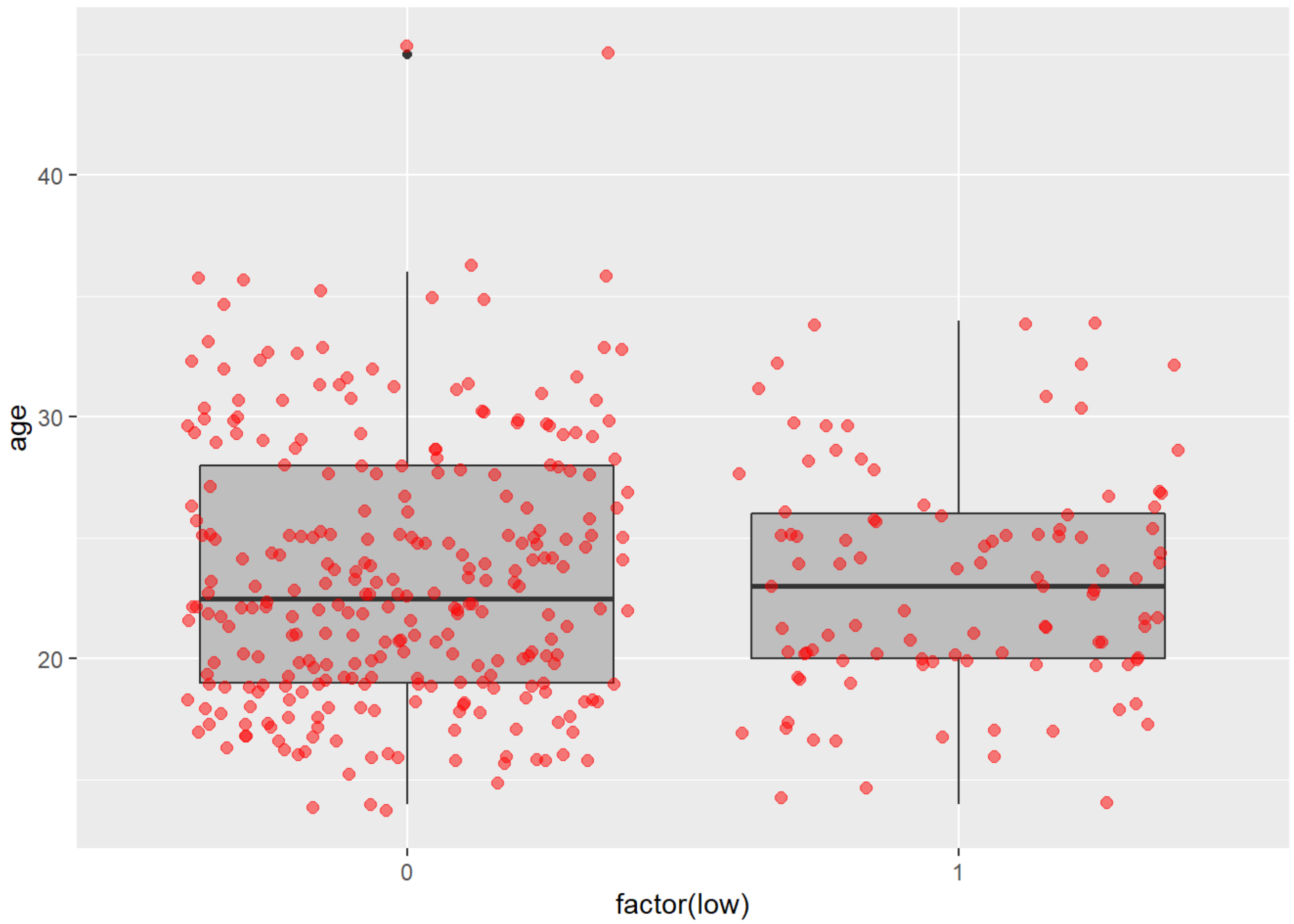
```
SC <- read.csv(text='
age,ftv,race,ptl,smoke,ht,ui,lwtkg
30,0,3,0,1,0,0,60
40,0,3,0,1,0,0,60
30,0,3,0,0,0,0,60
40,0,3,0,0,0,0,60
30,0,3,0,1,1,0,60
40,0,3,0,1,1,0,60
30,0,3,0,0,1,0,60
40,0,3,0,0,1,0,60
30,0,3,0,1,1,1,60
40,0,3,0,1,1,1,60
30,0,3,0,0,1,1,60
40,0,3,0,0,1,1,60
')
SC <- SC %>% mutate(race=factor(race, levels=1:3))
SCOUT <-
 SC %>%
 mutate(yh = predict(Rlm, SC))
SCOUT
```

```
##     age  ftv  race  ptl  smoke  ht  ui  lwtkg        yh
## 1    30    0     3    0      1   0   0     60  2534.252
## 2    40    0     3    0      1   0   0     60  2395.987
## 3    30    0     3    0      0   0   0     60  2938.109
## 4    40    0     3    0      0   0   0     60  2799.844
## 5    30    0     3    0      1   1   0     60  1973.122
## 6    40    0     3    0      1   1   0     60  1834.858
## 7    30    0     3    0      0   1   0     60  2376.979
## 8    40    0     3    0      0   1   0     60  2238.715
## 9    30    0     3    0      1   1   1     60  1443.734
## 10   40    0     3    0      1   1   1     60  1305.470
## 11   30    0     3    0      0   1   1     60  1847.591
## 12   40    0     3    0      0   1   1     60  1709.327
```

2. 로지스틱회귀모형

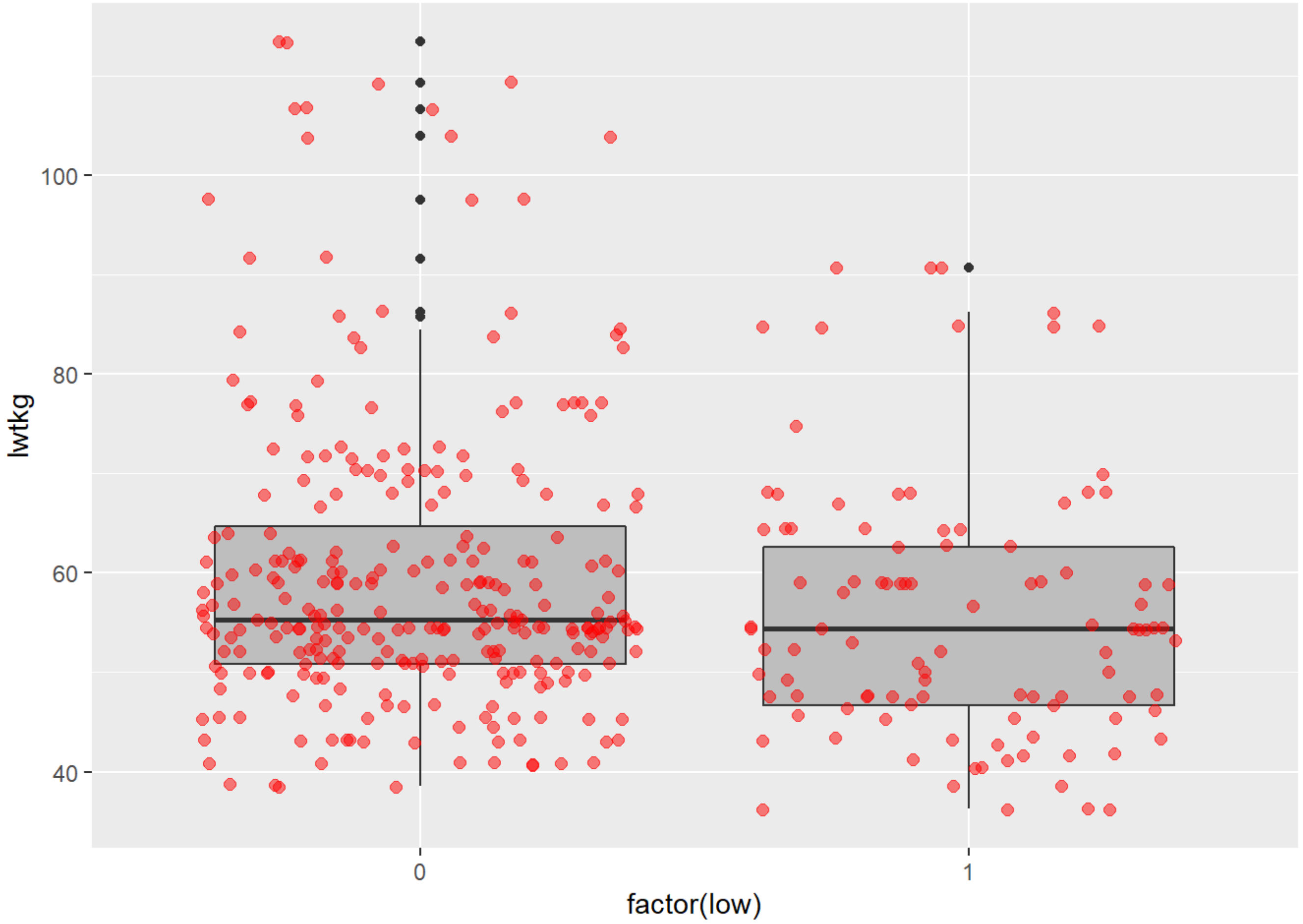
age vs low

```
ggplot(TR, aes(x=factor(low), y=age)) +
geom_boxplot(fill='gray') +
 geom_jitter(color='red', alpha=0.5, size=2)
```



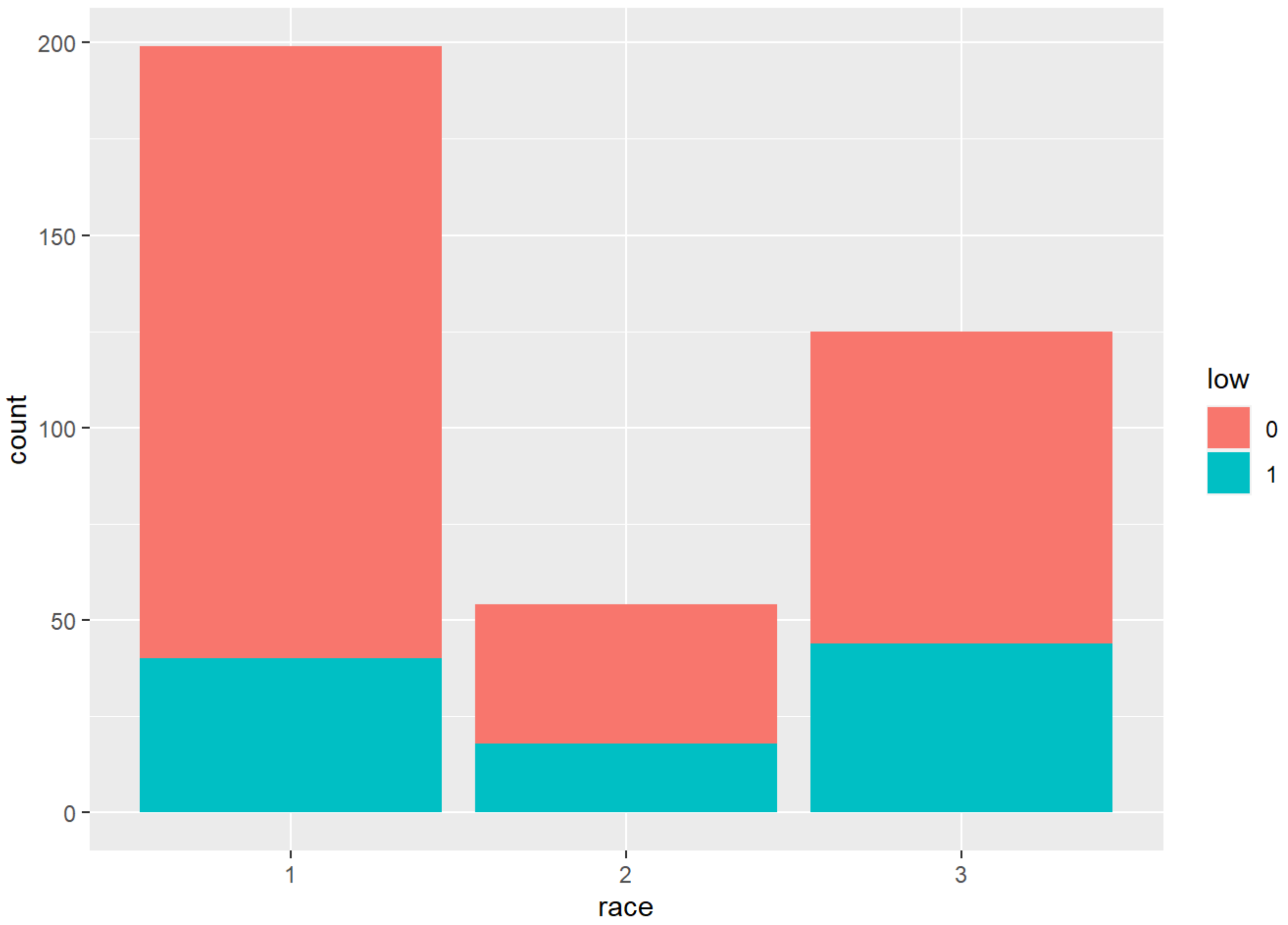
lwtkg vs low

```
ggplot(TR, aes(x=factor(low), y=lwtkg)) +
geom_boxplot(fill='gray') +
 geom_jitter(color='red', alpha=0.5, size=2)
```



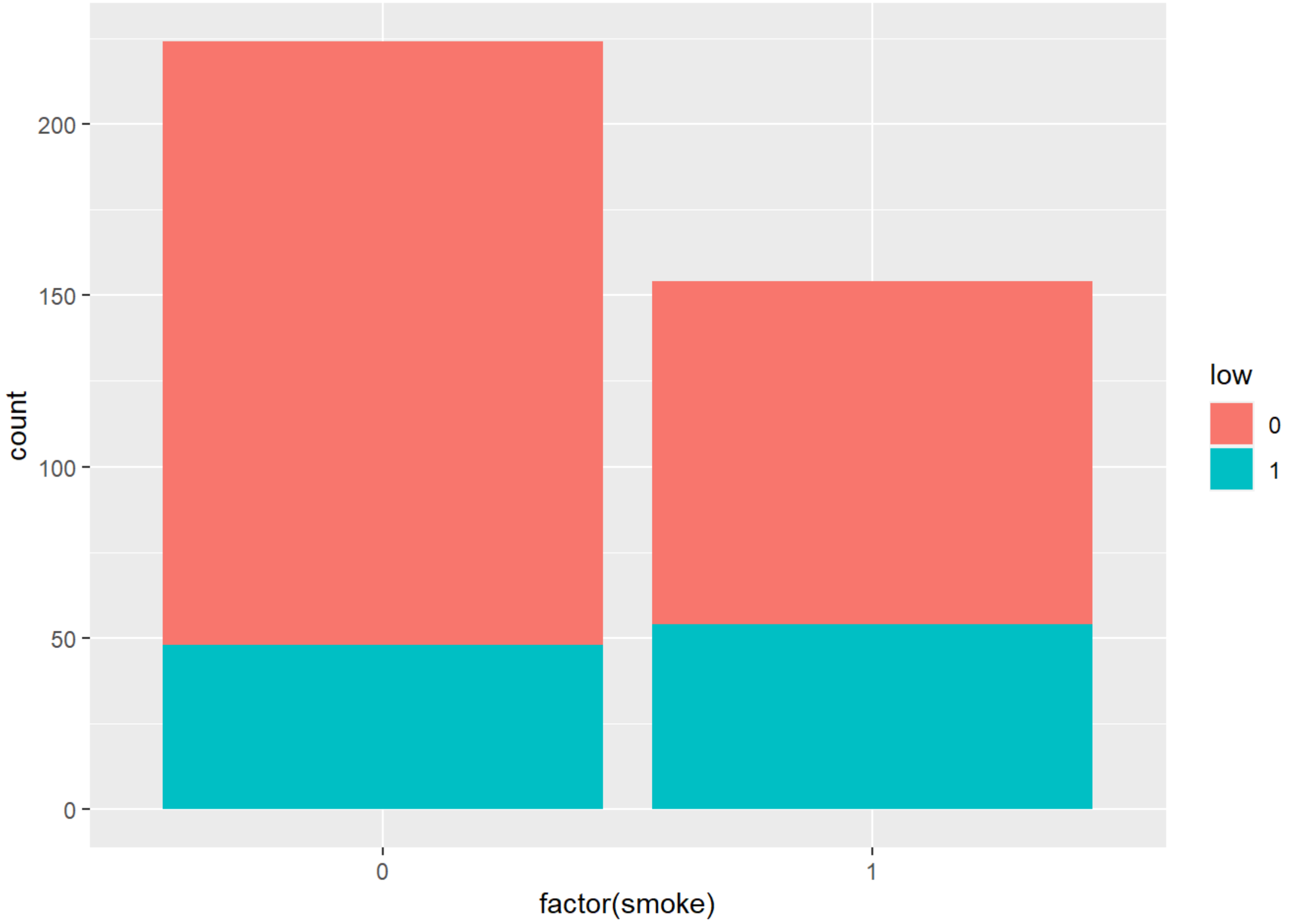
race vs low

```
ggplot(TR, aes(x=race, fill=low)) +
geom_bar()
```



smoke vs low

```
ggplot(TR, aes(x=factor(smoke), fill=low)) +
geom_bar()
```



모형적합

- 종속변수(y): low
- 독립변수: age, ftv, ptl, race, smoke, ht, ui, lwtkg

```
Mglm <- glm(low ~ age+ftv+ptl+race+smoke+ht+ui+lwtkg, data=TR, family=binomial)
summary(Mglm)
```

```
##
## Call:
## glm(formula = low ~ age + ftv + ptl + race + smoke + ht + ui +
##     lwtkg, family = binomial, data = TR)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -1.5607  -0.7525  -0.5753   0.7439   2.3179
##
## Coefficients:
##             Estimate Std. Error z value Pr(>|z|)
## (Intercept) -1.61373    0.84823  -1.902 0.057110 .
## age          0.02641    0.02670   0.989 0.322740
## ftv          0.07754    0.12425   0.624 0.532590
## ptl          0.28110    0.23679   1.187 0.235183
## race2        1.31087    0.39963   3.280 0.001037 **
## race3        1.21348    0.31857   3.809 0.000139 ***
## smoke        1.07547    0.29131   3.692 0.000223 ***
## ht           1.81267    0.51481   3.521 0.000430 ***
## ui           0.88840    0.34136   2.603 0.009253 **
## lwtkg       -0.02613    0.01035  -2.525 0.011574 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 440.82  on 377  degrees of freedom
## Residual deviance: 384.51  on 368  degrees of freedom
## AIC: 404.51
##
## Number of Fisher Scoring iterations: 4
```

모형검토(TR)

```
TROUT <-
 TR %>% dplyr::select(low) %>%
 mutate(
 ph = predict(Mglm, type='response'),
 yh = factor(ifelse(ph>=0.5, 1, 0)))
head(TROUT)
```

```
##    low        ph yh
## 1    0 0.1525332  0
## 3    0 0.2172778  0
## 5    0 0.2972587  0
## 7    0 0.3123296  0
## 9    0 0.2452283  0
## 11   0 0.4666107  0
```

```
confusionMatrix(TROUT$yh, TROUT$low, positive='1')
```

```
## Confusion Matrix and Statistics
## 
##           Reference
## Prediction   0   1
##          0 262  73
##          1  14  29
## 
##                Accuracy : 0.7698          
##                  95% CI : (0.7241, 0.8113)
##     No Information Rate : 0.7302          
##     P-Value [Acc > NIR] : 0.04475         
## 
##                   Kappa : 0.2857          
## 
##  Mcnemar's Test P-Value : 5.027e-10       
## 
##             Sensitivity : 0.28431         
##             Specificity : 0.94928         
##          Pos Pred Value : 0.67442         
##          Neg Pred Value : 0.78209         
##              Prevalence : 0.26984         
##          Detection Rate : 0.07672         
##    Detection Prevalence : 0.11376         
##       Balanced Accuracy : 0.61679         
## 
##        'Positive' Class : 1               
## 
```

모형평가(TS)

```
TSOUT <-
TS %>%
mutate(yh=predict(Rlm, TS), e=bwt-yh)
head(TSOUT)
```

```
##     id low  bwt age ftv race ptl smoke ht ui lwtkg       yh          e
## 2  101   0 3728  24   1    1   0     0  0  0  49.9 3324.103  403.89653
## 4  645   0 3430  32   4    1   1     1  0  0  60.8 2919.764  510.23603
## 6   98   0 3651  19   0    1   0     1  0  0  66.7 3147.701  503.29884
## 8  726   1 2187  27   0    2   0     0  0  1  59.0 2363.884 -176.88449
## 10 326   1 1588  23   1    3   0     0  0  1  44.0 2353.912 -765.91184
## 12 270   0 3460  22   1    1   0     0  0  0  59.4 3431.655   28.34472
```

```
mean(TSOUT$e^2) # MSE
```

```
## [1] 437372.7
```

```
mean(abs(TSOUT$e)) # MAE
```

```
## [1] 539.0403
```

```
TSOUT %>% summarize(mn=mean(e), sd=sd(e), min=min(e), max=max(e))
```

```
##          mn       sd       min     max
## 1 -1.103086 662.2172 -1764.091 1906.63
```

```
summary(TSOUT$e)
```

```
##      Min.   1st Qu.    Median      Mean   3rd Qu.      Max.
## -1764.091  -470.794    13.717    -1.103   495.133  1906.630
```

```
TSOUT <-
 TS %>% dplyr::select(low) %>%
 mutate(
 ph = predict(Mglm, TS, type='response'),
 yh = factor(ifelse(ph>=0.5, 1, 0)))
head(TSOUT)
```

```
##    low         ph yh
## 2    0 0.09918719  0
## 4    0 0.33388549  0
## 6    0 0.14438209  0
## 8    1 0.43952015  0
## 10   1 0.50583584  1
## 12   0 0.07534600  0
```

```
confusionMatrix(TSOUT$yh, TSOUT$low, positive='1')
```

```
## Confusion Matrix and Statistics
##
##           Reference
## Prediction   0   1
##          0 230  95
##          1  14  39
##
##                Accuracy : 0.7116
##                  95% CI : (0.6631, 0.7568)
##     No Information Rate : 0.6455
##     P-Value [Acc > NIR] : 0.003789
##
##                   Kappa : 0.2705
##
##  Mcnemar's Test P-Value : 1.822e-14
##
##             Sensitivity : 0.2910
##             Specificity : 0.9426
##          Pos Pred Value : 0.7358
##          Neg Pred Value : 0.7077
##              Prevalence : 0.3545
##          Detection Rate : 0.1032
##    Detection Prevalence : 0.1402
##       Balanced Accuracy : 0.6168
##
##        'Positive' Class : 1
##
```

스코어

```
SC <- read.csv(text='
age,ftv,race,ptl,smoke,ht,ui,lwtkg
30,0,3,0,1,0,0,60
40,0,3,0,1,0,0,60
30,0,3,0,0,0,0,60
40,0,3,0,0,0,0,60
30,0,3,0,1,1,0,60
40,0,3,0,1,1,0,60
30,0,3,0,0,1,0,60
40,0,3,0,0,1,0,60
30,0,3,0,1,1,1,60
40,0,3,0,1,1,1,60
30,0,3,0,0,1,1,60
40,0,3,0,0,1,1,60
')
SC <- SC %>% mutate(race=factor(race, levels=1:3))
SCOUT <-
 SC %>%
 mutate(
 ph = predict(Mglm, SC, type='response'),
 yh = factor(ifelse(ph>=0.5, 1, 0)))
SCOUT
```

```
##    age ftv race ptl smoke ht ui lwtkg        ph yh
## 1   30   0    3   0     1  0  0    60 0.4749339  0
## 2   40   0    3   0     1  0  0    60 0.5408376  1
## 3   30   0    3   0     0  0  0    60 0.2358051  0
## 4   40   0    3   0     0  0  0    60 0.2866411  0
## 5   30   0    3   0     1  1  0    60 0.8471371  1
## 6   40   0    3   0     1  1  0    60 0.8782953  1
## 7   30   0    3   0     0  1  0    60 0.6540417  1
## 8   40   0    3   0     0  1  0    60 0.7111378  1
## 9   30   0    3   0     1  1  1    60 0.9309080  1
## 10  40   0    3   0     1  1  1    60 0.9460779  1
## 11  30   0    3   0     0  1  1    60 0.8213108  1
## 12  40   0    3   0     0  1  1    60 0.8568435  1
```

3. 나무모형

```
library(yardstick)
```

```
## Warning: 패키지 'yardstick'는 R 버전 4.1.2에서 작성되었습니다
```

```
## For binary classification, the first factor level is assumed to be the event.
## Use the argument `event_level = "second"` to alter this as needed.
```

```
##
## 다음의 패키지를 부착합니다: 'yardstick'
```

```
## The following objects are masked from 'package:caret':
##
##     precision, recall, sensitivity, specificity
```

```
## The following object is masked from 'package:readr':
##
##     spec
```

```
library(ROCR)
```

```
## Warning: 패키지 'ROCR'는 R 버전 4.1.2에서 작성되었습니다
```

```
library(pROC)
```

```
## Warning: 패키지 'pROC'는 R 버전 4.1.2에서 작성되었습니다
```

```
## Type 'citation("pROC")' for a citation.
```

```
##
## 다음의 패키지를 부착합니다: 'pROC'
## The following objects are masked from 'package:stats':
##
##     cov, smooth, var
```

```
library(rpart)
library(rpart.plot)
```

```
## Warning: 패키지 'rpart.plot'는 R 버전 4.1.2에서 작성되었습니다
```

```
TR %>%
group_by(low) %>%
summarize_if(is.numeric,
             list(mn='mean', sd='sd', min='min', max='max'))
```

```
## # A tibble: 2 x 37
##   low   id_mn bwt_mn age_mn ftv_mn ptl_mn smoke_mn  ht_mn ui_mn lwtkg_mn id_sd
##   <fct> <dbl>  <dbl>  <dbl>  <dbl>  <dbl>    <dbl>  <dbl> <dbl>    <dbl> <dbl>
## 1 0      343.  3277.   23.4  0.779  0.138    0.362 0.0362 0.101     59.9  213.
## 2 1      439.  2071.   23.2  0.686  0.294    0.529 0.118  0.235     55.8  211.
## # ... with 26 more variables: bwt_sd <dbl>, age_sd <dbl>, ftv_sd <dbl>,
## #   ptl_sd <dbl>, smoke_sd <dbl>, ht_sd <dbl>, ui_sd <dbl>, lwtkg_sd <dbl>,
## #   id_min <int>, bwt_min <int>, age_min <int>, ftv_min <int>, ptl_min <int>,
## #   smoke_min <int>, ht_min <int>, ui_min <int>, lwtkg_min <dbl>, id_max <int>,
## #   bwt_max <int>, age_max <int>, ftv_max <int>, ptl_max <int>,
## #   smoke_max <int>, ht_max <int>, ui_max <int>, lwtkg_max <dbl>
```

모형적합

- 종속변수(y): low
- 독립변수: age, ftv, ptl, race, smoke, ht, ui, lwtkg

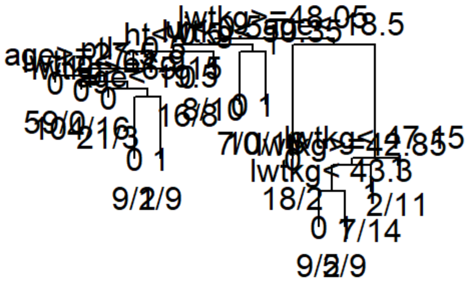
```
Mr <- rpart(low~ age+ftv+ptl+race+smoke+ht+ui+lwtkg, data=TR)
summary(Mr)
```

```
## Call:
## rpart(formula = low ~ age + ftv + ptl + race + smoke + ht + ui +
##     lwtkg, data = TR)
##   n= 378
##
##           CP nsplit rel error    xerror       xstd
## 1 0.06372549      0 1.0000000 1.0000000 0.08460744
## 2 0.02941176      2 0.8725490 0.9117647 0.08209521
## 3 0.02287582      4 0.8137255 0.9313725 0.08268161
## 4 0.01960784      7 0.7450980 0.9117647 0.08209521
## 5 0.01568627      8 0.7254902 0.9215686 0.08239048
## 6 0.01000000     13 0.6470588 0.9705882 0.08380571
##
## Variable importance
## lwtkg   age    ht    ui smoke   ptl  race   ftv
##    42    27     9     8     4     4     3     3
##
## Node number 1: 378 observations,    complexity param=0.06372549
##   predicted class=0  expected loss=0.2698413  P(node) =1
##     class counts:   276   102
##    probabilities: 0.730 0.270
##   left son=2 (299 obs) right son=3 (79 obs)
##   Primary splits:
##       lwtkg < 48.05  to the right, improve=8.907336, (0 missing)
##       ptl   < 0.5    to the left,  improve=6.387964, (0 missing)
##       ui    < 0.5    to the left,  improve=4.431380, (0 missing)
##       race  splits as  LRR,        improve=3.982504, (0 missing)
##       ht    < 0.5    to the left,  improve=3.548908, (0 missing)
##   Surrogate splits:
##       ptl < 2.5    to the left,  agree=0.799, adj=0.038, (0 split)
##
## Node number 2: 299 observations,    complexity param=0.02941176
##   predicted class=0  expected loss=0.2140468  P(node) =0.7910053
##     class counts:   235    64
##    probabilities: 0.786 0.214
##   left son=4 (266 obs) right son=5 (33 obs)
##   Primary splits:
##       ui    < 0.5    to the left,  improve=5.440466, (0 missing)
##       ht    < 0.5    to the left,  improve=4.467566, (0 missing)
##       ptl   < 0.5    to the left,  improve=3.648543, (0 missing)
##       smoke < 0.5    to the left,  improve=2.699448, (0 missing)
##       lwtkg < 91.15  to the right, improve=1.447077, (0 missing)
##   Surrogate splits:
##       lwtkg < 49.2   to the right, agree=0.91, adj=0.182, (0 split)
##
## Node number 3: 79 observations,    complexity param=0.06372549
##   predicted class=0  expected loss=0.4810127  P(node) =0.2089947
##     class counts:    41    38
##    probabilities: 0.519 0.481
##   left son=6 (20 obs) right son=7 (59 obs)
##   Primary splits:
##       age   < 18.5   to the left,  improve=7.7752410, (0 missing)
##       lwtkg < 47.15  to the left,  improve=4.1493320, (0 missing)
##       race  splits as  LRR,        improve=2.3988820, (0 missing)
##       ptl   < 0.5    to the left,  improve=1.1944610, (0 missing)
##       ftv   < 1.5    to the left,  improve=0.9753763, (0 missing)
##
## Node number 4: 266 observations,    complexity param=0.01960784
##   predicted class=0  expected loss=0.1804511  P(node) =0.7037037
##     class counts:   218    48
##    probabilities: 0.820 0.180
##   left son=8 (248 obs) right son=9 (18 obs)
##   Primary splits:
##       ht    < 0.5    to the left,  improve=5.432964, (0 missing)
##       smoke < 0.5    to the left,  improve=2.454143, (0 missing)
##       lwtkg < 62.35  to the left,  improve=1.823645, (0 missing)
##       ptl   < 0.5    to the left,  improve=1.779098, (0 missing)
##       age   < 27.5   to the right, improve=1.297999, (0 missing)
##   Surrogate splits:
##       lwtkg < 105.25 to the left,  agree=0.944, adj=0.167, (0 split)
##
## Node number 5: 33 observations,    complexity param=0.02941176
##   predicted class=0  expected loss=0.4848485  P(node) =0.08730159
##     class counts:    17    16
##    probabilities: 0.515 0.485
##   left son=10 (7 obs) right son=11 (26 obs)
##   Primary splits:
##       lwtkg < 50.35  to the left,  improve=4.1771560, (0 missing)
##       age   < 26     to the left,  improve=0.9353979, (0 missing)
##       ftv   < 0.5    to the right, improve=0.7495544, (0 missing)
##       race  splits as  LRR,        improve=0.4310023, (0 missing)
##       ptl   < 0.5    to the left,  improve=0.3805007, (0 missing)
##
## Node number 6: 20 observations
##   predicted class=0  expected loss=0.1  P(node) =0.05291005
##     class counts:    18     2
##    probabilities: 0.900 0.100
##
## Node number 7: 59 observations,    complexity param=0.02287582
##   predicted class=1  expected loss=0.3898305  P(node) =0.1560847
##     class counts:    23    36
##    probabilities: 0.390 0.610
##   left son=14 (46 obs) right son=15 (13 obs)
##   Primary splits:
##       lwtkg < 47.15  to the left,  improve=1.8570940, (0 missing)
##       age   < 24.5   to the left,  improve=1.3521790, (0 missing)
##       race  splits as  LRR,        improve=0.9820823, (0 missing)
##       ftv   < 0.5    to the right, improve=0.2190787, (0 missing)
##       ui    < 0.5    to the right, improve=0.0550982, (0 missing)
##   Surrogate splits:
##       age   < 31.5  to the left,  agree=0.831, adj=0.231, (0 split)
##       race splits as  LRL,        agree=0.831, adj=0.231, (0 split)
##
## Node number 8: 248 observations,    complexity param=0.01568627
##   predicted class=0  expected loss=0.1532258  P(node) =0.6560847
##     class counts:   210    38
##    probabilities: 0.847 0.153
##   left son=16 (224 obs) right son=17 (24 obs)
##   Primary splits:
##       ptl   < 0.5    to the left,  improve=1.7238860, (0 missing)
##       age   < 27.5   to the right, improve=1.6698710, (0 missing)
##       smoke < 0.5    to the left,  improve=1.6571640, (0 missing)
##       lwtkg < 68.7   to the right, improve=1.2585640, (0 missing)
##       ftv   < 0.5    to the right, improve=0.7108106, (0 missing)
## 
## Node number 9: 18 observations
##   predicted class=1  expected loss=0.4444444  P(node) =0.04761905
##     class counts:     8    10
##    probabilities: 0.444 0.556 
## 
## Node number 10: 7 observations
##   predicted class=0  expected loss=0  P(node) =0.01851852
##     class counts:     7     0
##    probabilities: 1.000 0.000 
## 
## Node number 11: 26 observations
##   predicted class=1  expected loss=0.3846154  P(node) =0.06878307
##     class counts:    10    16
##    probabilities: 0.385 0.615 
## 
## Node number 14: 46 observations,    complexity param=0.02287582
##   predicted class=1  expected loss=0.4565217  P(node) =0.1216931
##     class counts:    21    25
##    probabilities: 0.457 0.543 
##   left son=28 (25 obs) right son=29 (21 obs)
##   Primary splits:
##       lwtkg < 42.85  to the right, improve=1.17275400, (0 missing)
##       age   < 24.5   to the left,  improve=1.17275400, (0 missing)
##       race  splits as  L-R,        improve=0.31536570, (0 missing)
##       ftv   < 1.5    to the right, improve=0.03661327, (0 missing)
##       ptl   < 0.5    to the left,  improve=0.03144410, (0 missing)
##   Surrogate splits:
##       smoke < 0.5    to the left,  agree=0.739, adj=0.429, (0 split)
##       age   < 25.5   to the right, agree=0.696, adj=0.333, (0 split)
##       ftv   < 1.5    to the right, agree=0.630, adj=0.190, (0 split)
##       ui    < 0.5    to the left,  agree=0.587, adj=0.095, (0 split)
##       ptl   < 1.5    to the left,  agree=0.565, adj=0.048, (0 split)
## 
## Node number 15: 13 observations
##   predicted class=1  expected loss=0.1538462  P(node) =0.03439153
##     class counts:     2    11
##    probabilities: 0.154 0.846 
## 
## Node number 16: 224 observations,    complexity param=0.01568627
##   predicted class=0  expected loss=0.1339286  P(node) =0.5925926
##     class counts:   194    30
##    probabilities: 0.866 0.134 
##   left son=32 (59 obs) right son=33 (165 obs)
##   Primary splits:
##       age   < 27.5   to the right, improve=2.8733770, (0 missing)
##       lwtkg < 62.35  to the left,  improve=0.7248772, (0 missing)
##       race  splits as  LRR,        improve=0.6686513, (0 missing)
##       smoke < 0.5    to the left,  improve=0.4222373, (0 missing)
##       ftv   < 0.5    to the right, improve=0.1428571, (0 missing)
##   Surrogate splits:
##       lwtkg < 88.9   to the right, agree=0.754, adj=0.068, (0 split)
##       ftv   < 3.5    to the right, agree=0.746, adj=0.034, (0 split)
```

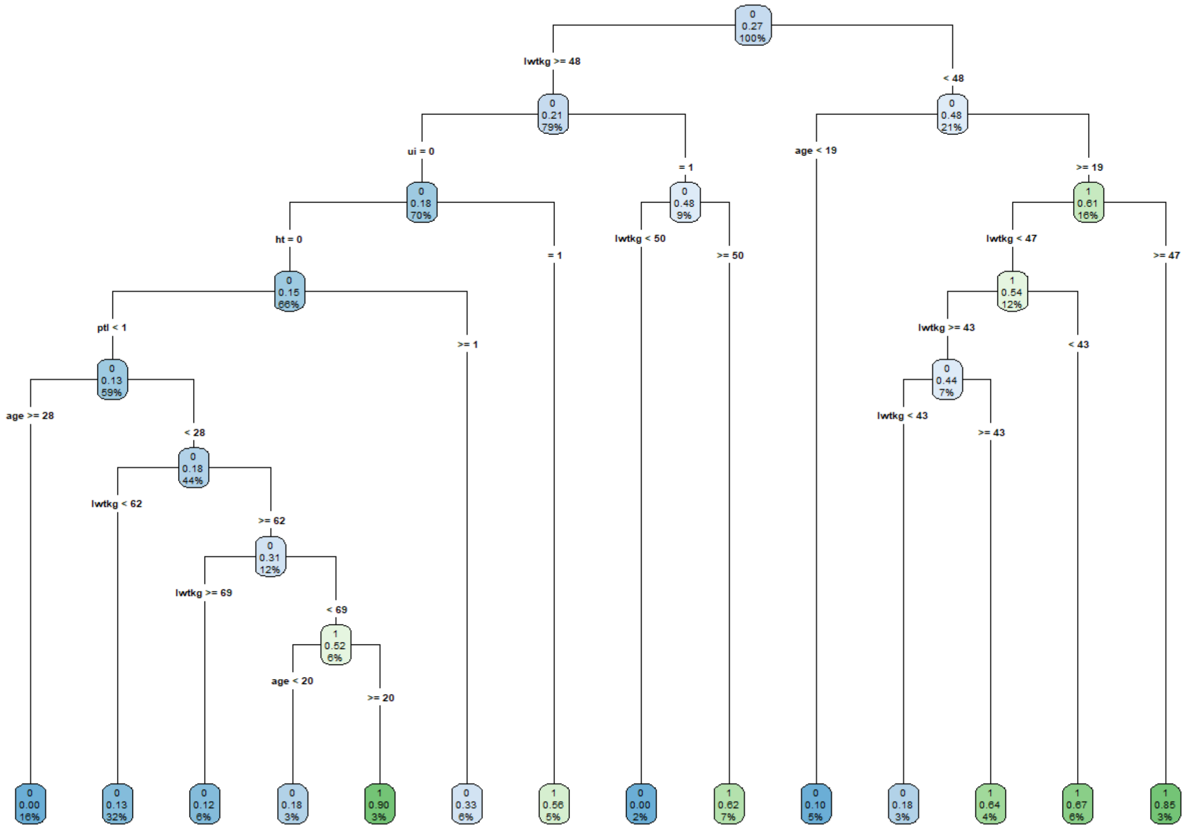
```
##
## Node number 17: 24 observations
##   predicted class=0  expected loss=0.3333333  P(node) =0.06349206
##     class counts:    16     8
##    probabilities: 0.667 0.333
##
## Node number 28: 25 observations,    complexity param=0.02287582
##   predicted class=0  expected loss=0.44  P(node) =0.06613757
##     class counts:    14    11
##    probabilities: 0.560 0.440
##   left son=56 (11 obs) right son=57 (14 obs)
##   Primary splits:
##       lwtkg < 43.3   to the left,  improve=2.6187010, (0 missing)
##       age   < 25.5   to the left,  improve=1.5148050, (0 missing)
##       ftv   < 0.5    to the left,  improve=0.8533333, (0 missing)
##       ptl   < 0.5    to the left,  improve=0.3358730, (0 missing)
##       race  splits as  L-R,        improve=0.2290909, (0 missing)
##   Surrogate splits:
##       smoke < 0.5    to the right, agree=0.72, adj=0.364, (0 split)
##       age   < 19.5   to the left,  agree=0.68, adj=0.273, (0 split)
##       ptl   < 2      to the right, agree=0.68, adj=0.273, (0 split)
##       ftv   < 2.5    to the left,  agree=0.64, adj=0.182, (0 split)
##       ht    < 0.5    to the right, agree=0.64, adj=0.182, (0 split)
##
## Node number 29: 21 observations
##   predicted class=1  expected loss=0.3333333  P(node) =0.05555556
##     class counts:     7    14
##    probabilities: 0.333 0.667
##
## Node number 32: 59 observations
##   predicted class=0  expected loss=0  P(node) =0.1560847
##     class counts:    59     0
##    probabilities: 1.000 0.000
##
## Node number 33: 165 observations,    complexity param=0.01568627
##   predicted class=0  expected loss=0.1818182  P(node) =0.4365079
##     class counts:   135    30
##    probabilities: 0.818 0.182
##   left son=66 (120 obs) right son=67 (45 obs)
##   Primary splits:
##       lwtkg < 61.9   to the left,  improve=2.0686870, (0 missing)
##       race  splits as  LLR,        improve=0.7100941, (0 missing)
##       age   < 16.5   to the left,  improve=0.5044997, (0 missing)
##       smoke < 0.5    to the left,  improve=0.4293757, (0 missing)
##       ftv   < 1.5    to the left,  improve=0.3030303, (0 missing)
##
## Node number 56: 11 observations
##   predicted class=0  expected loss=0.1818182  P(node) =0.02910053
##     class counts:     9     2
##    probabilities: 0.818 0.182
##
## Node number 57: 14 observations
##   predicted class=1  expected loss=0.3571429  P(node) =0.03703704
##     class counts:     5     9
##    probabilities: 0.357 0.643
##
```

```
## Node number 66: 120 observations
##   predicted class=0  expected loss=0.1333333  P(node) =0.3174603
##     class counts:   104    16
##    probabilities: 0.867 0.133 
## 
## Node number 67: 45 observations,    complexity param=0.01568627
##   predicted class=0  expected loss=0.3111111  P(node) =0.1190476
##     class counts:    31    14
##    probabilities: 0.689 0.311 
##   left son=134 (24 obs) right son=135 (21 obs)
##   Primary splits:
##       lwtkg < 69.15  to the right, improve=3.56269800, (0 missing)
##       age   < 22.5   to the left,  improve=1.77164800, (0 missing)
##       race  splits as  LLR,        improve=1.12347500, (0 missing)
##       smoke < 0.5    to the left,  improve=0.21601440, (0 missing)
##       ftv   < 1.5    to the left,  improve=0.08033274, (0 missing)
##   Surrogate splits:
##       age   < 20.5   to the right, agree=0.689, adj=0.333, (0 split)
##       race  splits as  RLR,        agree=0.667, adj=0.286, (0 split)
##       ftv   < 1.5    to the left,  agree=0.600, adj=0.143, (0 split)
##       smoke < 0.5    to the left,  agree=0.600, adj=0.143, (0 split)
## 
## Node number 134: 24 observations
##   predicted class=0  expected loss=0.125  P(node) =0.06349206
##     class counts:    21     3
##    probabilities: 0.875 0.125 
## 
## Node number 135: 21 observations,    complexity param=0.01568627
##   predicted class=1  expected loss=0.4761905  P(node) =0.05555556
##     class counts:    10    11
##    probabilities: 0.476 0.524 
##   left son=270 (11 obs) right son=271 (10 obs)
##   Primary splits:
##       age   < 19.5   to the left,  improve=5.4034630, (0 missing)
##       lwtkg < 66.9   to the left,  improve=2.9125540, (0 missing)
##       smoke < 0.5    to the right, improve=1.1852810, (0 missing)
##       ftv   < 0.5    to the right, improve=0.5852814, (0 missing)
##   Surrogate splits:
##       lwtkg < 67.55  to the left,  agree=0.714, adj=0.4, (0 split)
##       smoke < 0.5    to the right, agree=0.619, adj=0.2, (0 split)
##       ftv   < 1.5    to the left,  agree=0.571, adj=0.1, (0 split)
##       race  splits as  R-L,        agree=0.571, adj=0.1, (0 split)
## 
## Node number 270: 11 observations
##   predicted class=0  expected loss=0.1818182  P(node) =0.02910053
##     class counts:     9     2
##    probabilities: 0.818 0.182 
## 
## Node number 271: 10 observations
##   predicted class=1  expected loss=0.1  P(node) =0.02645503
##     class counts:     1     9
##    probabilities: 0.100 0.900
```

```
plot(Mr, margin=1)
text(Mr, use.n=TRUE)
```



```
rpart.plot(Mr, type=4)
```



모형검토(TR)

```
TROUT <-
 TR %>% dplyr::select(low) %>%
 mutate(
 ph = predict(Mr, type='prob')[,2],
 yh = factor(ifelse(ph>=0.5, 1, 0)))
head(TROUT)
```

```
##    low        ph yh
## 1    0 0.1333333  0
## 3    0 0.1333333  0
## 5    0 0.6428571  1
## 7    0 0.3333333  0
## 9    0 0.1333333  0
## 11   0 0.1333333  0
```

```
confusionMatrix(TROUT$yh, TROUT$low, positive='1')
```

```
## Confusion Matrix and Statistics
##
##           Reference
## Prediction   0   1
##          0 243  33
##          1  33  69
##
##                Accuracy : 0.8254
##                  95% CI : (0.7833, 0.8623)
##     No Information Rate : 0.7302
##     P-Value [Acc > NIR] : 8.954e-06
##
##                   Kappa : 0.5569
##
##  Mcnemar's Test P-Value : 1
##
##             Sensitivity : 0.6765
##             Specificity : 0.8804
##          Pos Pred Value : 0.6765
##          Neg Pred Value : 0.8804
##              Prevalence : 0.2698
##          Detection Rate : 0.1825
##    Detection Prevalence : 0.2698
##       Balanced Accuracy : 0.7785
##
##        'Positive' Class : 1
##
```

```
TR <-
TR %>%
  mutate(
    ph = predict (Mr, type='prob')[,2],
    yh = factor(ifelse(ph>=0.5, 1, 0)))
head(TR)
```

```
##      id low  bwt age ftv race ptl smoke ht ui lwtkg        ph yh
## 1   284   0 3643  16   0    1   0     1  0  0  61.2 0.1333333  0
## 3   623   0 3175  16   0    3   0     0  0  0  49.9 0.1333333  0
## 5   400   0 2835  31   3    1   0     0  0  1  45.4 0.6428571  1
## 7   103   0 3770  24   0    3   1     0  0  0  49.9 0.3333333  0
## 9   602   0 2977  25   0    2   0     0  0  0  56.7 0.1333333  0
## 11   79   0 3444  20   0    2   0     1  0  0  54.9 0.1333333  0
```

모형평가(TS)

```
TSOUT <-
TS %>%
mutate(yh=predict(Mr, TS), e=bwt-yh)
head(TSOUT)
```

```
##      id low  bwt age ftv race ptl smoke ht ui lwtkg      yh.0      yh.1      e.0
## 2   101   0 3728  24   1    1   0     0  0  0  49.9 0.8666667 0.1333333 3727.133
## 4   645   0 3430  32   4    1   1     1  0  0  60.8 0.6666667 0.3333333 3429.333
## 6    98   0 3651  19   0    1   0     1  0  0  66.7 0.8181818 0.1818182 3650.182
## 8   726   1 2187  27   0    2   0     0  0  1  59.0 0.3846154 0.6153846 2186.615
## 10  326   1 1588  23   1    3   0     0  0  1  44.0 0.3571429 0.6428571 1587.643
## 12  270   0 3460  22   1    1   0     0  0  0  59.4 0.8666667 0.1333333 3459.133
##          e.1
## 2   3727.867
## 4   3429.667
## 6   3650.818
## 8   2186.385
## 10  1587.357
## 12  3459.867
```

```
mean(TSOUT$e^2) # MSE
```

```
## [1] 9189335
```

```
mean(abs(TSOUT$e)) # MAE
```

```
## [1] 2937.188
```

```
TSOUT %>% summarize(mn=mean(e), sd=sd(e), min=min(e), max=max(e))
```

```
##         mn      sd      min  max
## 1 2937.188 750.338 708.3846 4990
```

```
summary(TSOUT$e)
```

```
##        0                 1
##  Min.   : 708.6   Min.   : 708.4
##  1st Qu.:2380.1   1st Qu.:2380.4
##  Median :2976.0   Median :2976.9
##  Mean   :2937.0   Mean   :2937.4
##  3rd Qu.:3571.1   3rd Qu.:3571.2
##  Max.   :4989.0   Max.   :4990.0
```

스코어

```
SC <- read.csv(text='
age,ftv,race,ptl,smoke,ht,ui,lwtkg
30,0,3,0,1,0,0,60
40,0,3,0,1,0,0,60
30,0,3,0,0,0,0,60
40,0,3,0,0,0,0,60
30,0,3,0,1,1,0,60
40,0,3,0,1,1,0,60
30,0,3,0,0,1,0,60
40,0,3,0,0,1,0,60
30,0,3,0,1,1,1,60
40,0,3,0,1,1,1,60
30,0,3,0,0,1,1,60
40,0,3,0,0,1,1,60
')
SC <- SC %>% mutate(race=factor(race, levels=1:3))

SCOUT <-
  SC %>%
  mutate(
    ph = predict(Mr, SC, type='prob')[,2],
    yh = factor(ifelse(ph>=0.5, 1, 0)))
SCOUT
```

```
##    age ftv race ptl smoke ht ui lwtkg        ph yh
## 1   30   0    3   0     1  0  0    60 0.0000000  0
## 2   40   0    3   0     1  0  0    60 0.0000000  0
## 3   30   0    3   0     0  0  0    60 0.0000000  0
## 4   40   0    3   0     0  0  0    60 0.0000000  0
## 5   30   0    3   0     1  1  0    60 0.5555556  1
## 6   40   0    3   0     1  1  0    60 0.5555556  1
## 7   30   0    3   0     0  1  0    60 0.5555556  1
## 8   40   0    3   0     0  1  0    60 0.5555556  1
## 9   30   0    3   0     1  1  1    60 0.6153846  1
## 10  40   0    3   0     1  1  1    60 0.6153846  1
## 11  30   0    3   0     0  1  1    60 0.6153846  1
## 12  40   0    3   0     0  1  1    60 0.6153846  1
```

4. 랜덤포레스트

```
library(randomForest)
```

```
## Warning: 패키지 'randomForest'는 R 버전 4.1.2에서 작성되었습니다
```

```
## randomForest 4.6-14
```

```
## Type rfNews() to see new features/changes/bug fixes.
```

```
##
## 다음의 패키지를 부착합니다: 'randomForest'
```

```
## The following object is masked from 'package:dplyr':
##
##     combine
```

```
## The following object is masked from 'package:ggplot2':
##
##     margin
```

모형적합

- 종속변수(y): low
- 독립변수: age, ftv, ptl, race, smoke, ht, ui, lwtkg

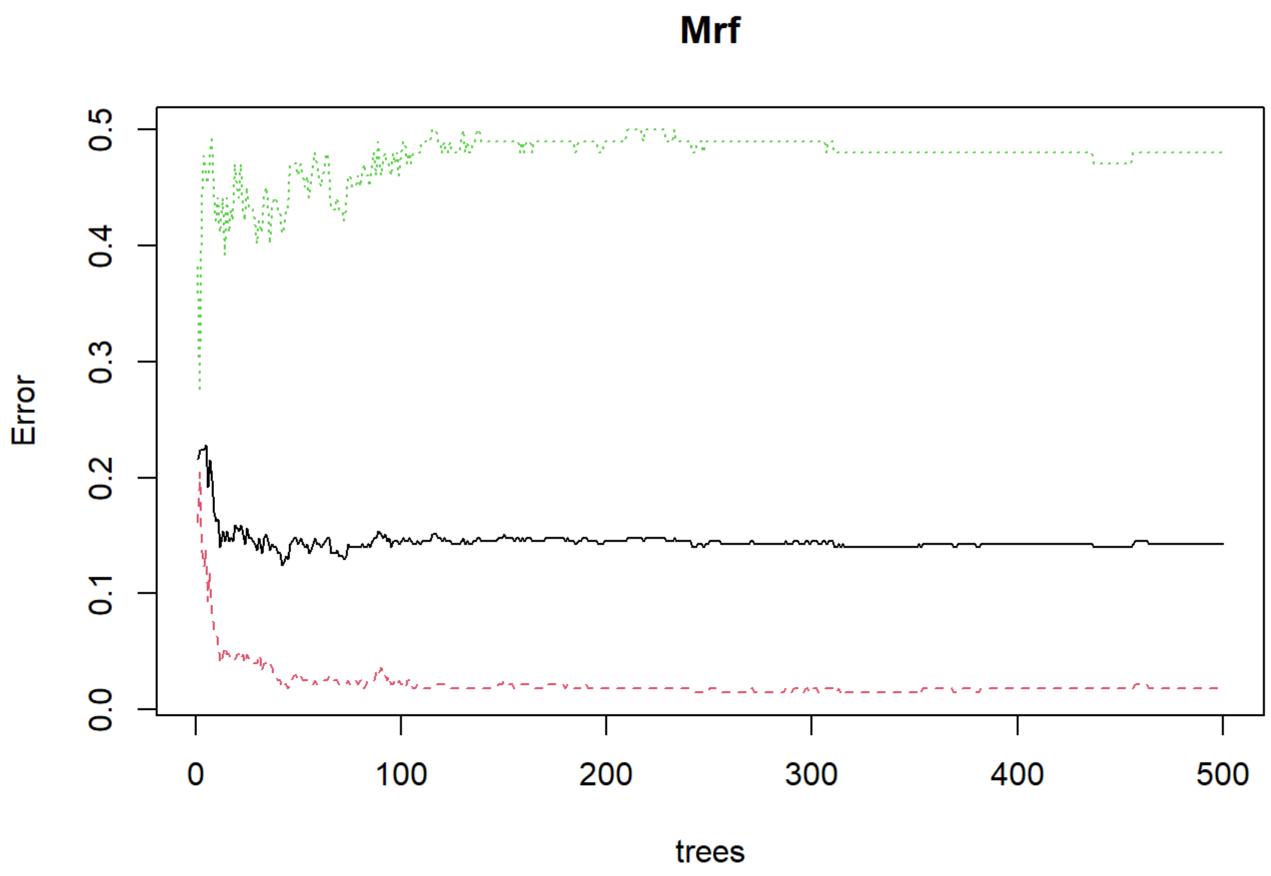
```
Mrf <- randomForest(low~ age+ftv+ptl+race+smoke+ht+ui+lwtkg, data=TR)
summary(Mr)
```

```
## Call:
## rpart(formula = low ~ age + ftv + ptl + race + smoke + ht + ui + 
##     lwtkg, data = TR)
##   n= 378 
## 
##           CP nsplit rel error    xerror       xstd
## 1 0.06372549      0 1.0000000 1.0000000 0.08460744
## 2 0.02941176      2 0.8725490 0.9117647 0.08209521
## 3 0.02287582      4 0.8137255 0.9313725 0.08268161
## 4 0.01960784      7 0.7450980 0.9117647 0.08209521
## 5 0.01568627      8 0.7254902 0.9215686 0.08239048
## 6 0.01000000     13 0.6470588 0.9705882 0.08380571
## 
## Variable importance
## lwtkg   age    ht    ui smoke   ptl  race   ftv 
##    42    27     9     8     4     4     3     3 
## 
## Node number 1: 378 observations,    complexity param=0.06372549
##   predicted class=0  expected loss=0.2698413  P(node) =1
##     class counts:   276   102
##    probabilities: 0.730 0.270 
##   left son=2 (299 obs) right son=3 (79 obs)
##   Primary splits:
##       lwtkg < 48.05 to the right, improve=8.907336, (0 missing)
##       ptl   < 0.5   to the left,  improve=6.387964, (0 missing)
##       ui    < 0.5   to the left,  improve=4.431380, (0 missing)
##       race  splits as  LRR,       improve=3.982504, (0 missing)
##       ht    < 0.5   to the left,  improve=3.548908, (0 missing)
##   Surrogate splits:
##       ptl < 2.5   to the left,  agree=0.799, adj=0.038, (0 split)
## 
## Node number 2: 299 observations,    complexity param=0.02941176
##   predicted class=0  expected loss=0.2140468  P(node) =0.7910053
##     class counts:   235    64
##    probabilities: 0.786 0.214 
##   left son=4 (266 obs) right son=5 (33 obs)
##   Primary splits:
##       ui    < 0.5   to the left,  improve=5.440466, (0 missing)
##       ht    < 0.5   to the left,  improve=4.467566, (0 missing)
##       ptl   < 0.5   to the left,  improve=3.648543, (0 missing)
##       smoke < 0.5   to the left,  improve=2.699448, (0 missing)
##       lwtkg < 91.15 to the right, improve=1.447077, (0 missing)
##   Surrogate splits:
##       lwtkg < 49.2  to the right, agree=0.91, adj=0.182, (0 split)
## 
## Node number 3: 79 observations,    complexity param=0.06372549
##   predicted class=0  expected loss=0.4810127  P(node) =0.2089947
##     class counts:    41    38
##    probabilities: 0.519 0.481 
##   left son=6 (20 obs) right son=7 (59 obs)
##   Primary splits:
##       age   < 18.5  to the left,  improve=7.7752410, (0 missing)
##       lwtkg < 47.15 to the left,  improve=4.1493320, (0 missing)
##       race  splits as  LRR,       improve=2.3988820, (0 missing)
##       ptl   < 0.5   to the left,  improve=1.1944610, (0 missing)
##       ftv   < 1.5    to the left,  improve=0.9753763, (0 missing)
## 
## Node number 4: 266 observations,    complexity param=0.01960784
##   predicted class=0  expected loss=0.1804511  P(node) =0.7037037
##     class counts:   218    48
##    probabilities: 0.820 0.180 
##   left son=8 (248 obs) right son=9 (18 obs)
##   Primary splits:
##       ht    < 0.5    to the left,  improve=5.432964, (0 missing)
##       smoke < 0.5    to the left,  improve=2.454143, (0 missing)
##       lwtkg < 62.35  to the left,  improve=1.823645, (0 missing)
##       ptl   < 0.5    to the left,  improve=1.779098, (0 missing)
##       age   < 27.5   to the right, improve=1.297999, (0 missing)
##   Surrogate splits:
##       lwtkg < 105.25 to the left,  agree=0.944, adj=0.167, (0 split)
## 
## Node number 5: 33 observations,    complexity param=0.02941176
##   predicted class=0  expected loss=0.4848485  P(node) =0.08730159
##     class counts:    17    16
##    probabilities: 0.515 0.485 
##   left son=10 (7 obs) right son=11 (26 obs)
##   Primary splits:
##       lwtkg < 50.35  to the left,  improve=4.1771560, (0 missing)
##       age   < 26     to the left,  improve=0.9353979, (0 missing)
##       ftv   < 0.5    to the right, improve=0.7495544, (0 missing)
##       race  splits as  LRR,        improve=0.4310023, (0 missing)
##       ptl   < 0.5    to the left,  improve=0.3805007, (0 missing)
## 
## Node number 6: 20 observations
##   predicted class=0  expected loss=0.1  P(node) =0.05291005
##     class counts:    18     2
##    probabilities: 0.900 0.100 
## 
## Node number 7: 59 observations,    complexity param=0.02287582
##   predicted class=1  expected loss=0.3898305  P(node) =0.1560847
##     class counts:    23    36
##    probabilities: 0.390 0.610 
##   left son=14 (46 obs) right son=15 (13 obs)
##   Primary splits:
##       lwtkg < 47.15  to the left,  improve=1.8570940, (0 missing)
##       age   < 24.5   to the left,  improve=1.3521790, (0 missing)
##       race  splits as  LRR,        improve=0.9820823, (0 missing)
##       ftv   < 0.5    to the right, improve=0.2190787, (0 missing)
##       ui    < 0.5    to the right, improve=0.0550982, (0 missing)
##   Surrogate splits:
##       age  < 31.5   to the left,  agree=0.831, adj=0.231, (0 split)
##       race splits as  LRL,        agree=0.831, adj=0.231, (0 split)
## 
## Node number 8: 248 observations,    complexity param=0.01568627
##   predicted class=0  expected loss=0.1532258  P(node) =0.6560847
##     class counts:   210    38
##    probabilities: 0.847 0.153 
##   left son=16 (224 obs) right son=17 (24 obs)
##   Primary splits:
##       ptl   < 0.5    to the left,  improve=1.7238860, (0 missing)
##       age   < 27.5   to the right, improve=1.6698710, (0 missing)
##       smoke < 0.5    to the left,  improve=1.6571640, (0 missing)
##       lwtkg < 68.7   to the right, improve=1.2585640, (0 missing)
##       ftv   < 0.5    to the right, improve=0.7108106, (0 missing)
##
## Node number 9: 18 observations
##   predicted class=1  expected loss=0.4444444  P(node) =0.04761905
##     class counts:     8    10
##    probabilities: 0.444 0.556
##
## Node number 10: 7 observations
##   predicted class=0  expected loss=0  P(node) =0.01851852
##     class counts:     7     0
##    probabilities: 1.000 0.000
##
## Node number 11: 26 observations
##   predicted class=1  expected loss=0.3846154  P(node) =0.06878307
##     class counts:    10    16
##    probabilities: 0.385 0.615
##
## Node number 14: 46 observations,    complexity param=0.02287582
##   predicted class=1  expected loss=0.4565217  P(node) =0.1216931
##     class counts:    21    25
##    probabilities: 0.457 0.543
##   left son=28 (25 obs) right son=29 (21 obs)
##   Primary splits:
##       lwtkg < 42.85  to the right, improve=1.17275400, (0 missing)
##       age   < 24.5   to the left,  improve=1.17275400, (0 missing)
##       race  splits as  L-R,        improve=0.31536570, (0 missing)
##       ftv   < 1.5    to the right, improve=0.03661327, (0 missing)
##       ptl   < 0.5    to the left,  improve=0.03144410, (0 missing)
##   Surrogate splits:
##       smoke < 0.5    to the left,  agree=0.739, adj=0.429, (0 split)
##       age   < 25.5   to the right, agree=0.696, adj=0.333, (0 split)
##       ftv   < 1.5    to the right, agree=0.630, adj=0.190, (0 split)
##       ui    < 0.5    to the left,  agree=0.587, adj=0.095, (0 split)
##       ptl   < 1.5    to the left,  agree=0.565, adj=0.048, (0 split)
##
## Node number 15: 13 observations
##   predicted class=1  expected loss=0.1538462  P(node) =0.03439153
##     class counts:     2    11
##    probabilities: 0.154 0.846
##
## Node number 16: 224 observations,    complexity param=0.01568627
##   predicted class=0  expected loss=0.1339286  P(node) =0.5925926
##     class counts:   194    30
##    probabilities: 0.866 0.134
##   left son=32 (59 obs) right son=33 (165 obs)
##   Primary splits:
##       age   < 27.5   to the right, improve=2.8733770, (0 missing)
##       lwtkg < 62.35  to the left,  improve=0.7248772, (0 missing)
##       race  splits as  LRR,        improve=0.6686513, (0 missing)
##       smoke < 0.5    to the left,  improve=0.4222373, (0 missing)
##       ftv   < 0.5    to the right, improve=0.1428571, (0 missing)
##   Surrogate splits:
##       lwtkg < 88.9   to the right, agree=0.754, adj=0.068, (0 split)
##       ftv   < 3.5    to the right, agree=0.746, adj=0.034, (0 split)
```

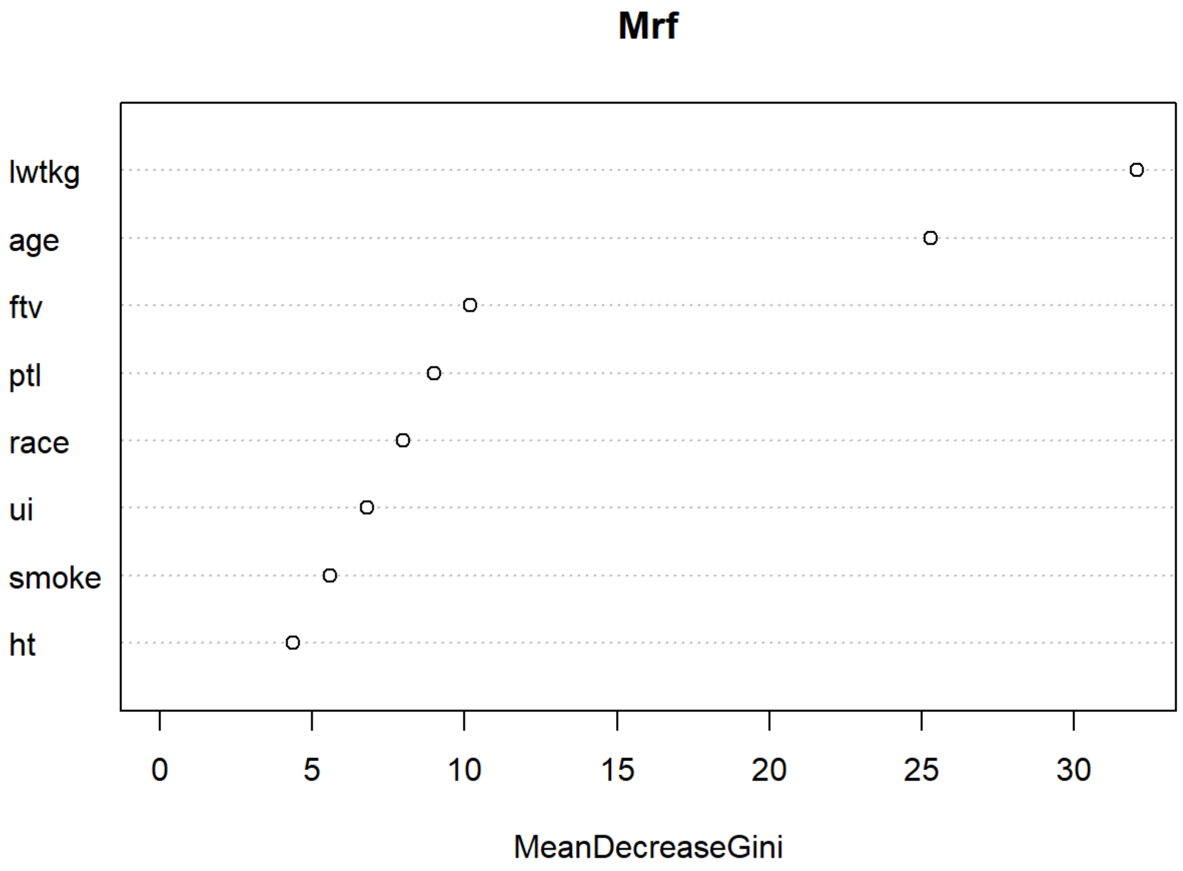
```
##
## Node number 17: 24 observations
##   predicted class=0  expected loss=0.3333333  P(node) =0.06349206
##     class counts:    16     8
##    probabilities: 0.667 0.333
##
## Node number 28: 25 observations,    complexity param=0.02287582
##   predicted class=0  expected loss=0.44  P(node) =0.06613757
##     class counts:    14    11
##    probabilities: 0.560 0.440
##   left son=56 (11 obs) right son=57 (14 obs)
##   Primary splits:
##       lwtkg < 43.3   to the left,  improve=2.6187010, (0 missing)
##       age   < 25.5   to the left,  improve=1.5148050, (0 missing)
##       ftv   < 0.5    to the left,  improve=0.8533333, (0 missing)
##       ptl   < 0.5    to the left,  improve=0.3358730, (0 missing)
##       race  splits as  L-R,        improve=0.2290909, (0 missing)
##   Surrogate splits:
##       smoke < 0.5    to the right, agree=0.72, adj=0.364, (0 split)
##       age   < 19.5   to the left,  agree=0.68, adj=0.273, (0 split)
##       ptl   < 2      to the right, agree=0.68, adj=0.273, (0 split)
##       ftv   < 2.5    to the left,  agree=0.64, adj=0.182, (0 split)
##       ht    < 0.5    to the right, agree=0.64, adj=0.182, (0 split)
##
## Node number 29: 21 observations
##   predicted class=1  expected loss=0.3333333  P(node) =0.05555556
##     class counts:     7    14
##    probabilities: 0.333 0.667
##
## Node number 32: 59 observations
##   predicted class=0  expected loss=0  P(node) =0.1560847
##     class counts:    59     0
##    probabilities: 1.000 0.000
##
## Node number 33: 165 observations,    complexity param=0.01568627
##   predicted class=0  expected loss=0.1818182  P(node) =0.4365079
##     class counts:   135    30
##    probabilities: 0.818 0.182
##   left son=66 (120 obs) right son=67 (45 obs)
##   Primary splits:
##       lwtkg < 61.9   to the left,  improve=2.0686870, (0 missing)
##       race  splits as  LLR,        improve=0.7100941, (0 missing)
##       age   < 16.5   to the left,  improve=0.5044997, (0 missing)
##       smoke < 0.5    to the left,  improve=0.4293757, (0 missing)
##       ftv   < 1.5    to the left,  improve=0.3030303, (0 missing)
##
## Node number 56: 11 observations
##   predicted class=0  expected loss=0.1818182  P(node) =0.02910053
##     class counts:     9     2
##    probabilities: 0.818 0.182
##
## Node number 57: 14 observations
##   predicted class=1  expected loss=0.3571429  P(node) =0.03703704
##     class counts:     5     9
##    probabilities: 0.357 0.643
##
```

```
## Node number 66: 120 observations
##   predicted class=0  expected loss=0.1333333  P(node) =0.3174603
##     class counts:   104    16
##    probabilities: 0.867 0.133 
## 
## Node number 67: 45 observations,    complexity param=0.01568627
##   predicted class=0  expected loss=0.3111111  P(node) =0.1190476
##     class counts:    31    14
##    probabilities: 0.689 0.311 
##   left son=134 (24 obs) right son=135 (21 obs)
##   Primary splits:
##       lwtkg < 69.15  to the right, improve=3.56269800, (0 missing)
##       age   < 22.5   to the left,  improve=1.77164800, (0 missing)
##       race  splits as  LLR,        improve=1.12347500, (0 missing)
##       smoke < 0.5    to the left,  improve=0.21601440, (0 missing)
##       ftv   < 1.5    to the left,  improve=0.08033274, (0 missing)
##   Surrogate splits:
##       age   < 20.5   to the right, agree=0.689, adj=0.333, (0 split)
##       race  splits as  RLR,        agree=0.667, adj=0.286, (0 split)
##       ftv   < 1.5    to the left,  agree=0.600, adj=0.143, (0 split)
##       smoke < 0.5    to the left,  agree=0.600, adj=0.143, (0 split)
## 
## Node number 134: 24 observations
##   predicted class=0  expected loss=0.125  P(node) =0.06349206
##     class counts:    21     3
##    probabilities: 0.875 0.125 
## 
## Node number 135: 21 observations,    complexity param=0.01568627
##   predicted class=1  expected loss=0.4761905  P(node) =0.05555556
##     class counts:    10    11
##    probabilities: 0.476 0.524 
##   left son=270 (11 obs) right son=271 (10 obs)
##   Primary splits:
##       age   < 19.5   to the left,  improve=5.4034630, (0 missing)
##       lwtkg < 66.9   to the left,  improve=2.9125540, (0 missing)
##       smoke < 0.5    to the right, improve=1.1852810, (0 missing)
##       ftv   < 0.5    to the right, improve=0.5852814, (0 missing)
##   Surrogate splits:
##       lwtkg < 67.55  to the left,  agree=0.714, adj=0.4, (0 split)
##       smoke < 0.5    to the right, agree=0.619, adj=0.2, (0 split)
##       ftv   < 1.5    to the left,  agree=0.571, adj=0.1, (0 split)
##       race  splits as  R-L,        agree=0.571, adj=0.1, (0 split)
## 
## Node number 270: 11 observations
##   predicted class=0  expected loss=0.1818182  P(node) =0.02910053
##     class counts:     9     2
##    probabilities: 0.818 0.182 
## 
## Node number 271: 10 observations
##   predicted class=1  expected loss=0.1  P(node) =0.02645503
##     class counts:     1     9
##    probabilities: 0.100 0.900
```

```
plot(Mrf)
```



```
varImpPlot(Mrf)
```



```
treesize(Mrf)
```

```
##   [1] 61 35 33 48 54 27 41 24 51 50 54 42 43 43 44 52 33 45 36 48 37 41 49 25 34
##  [26] 43 52 45 49 44 47 60 37 31 47 58 49 39 59 39 60 40 59 27 44 50 37 52 62 47
##  [51] 55 36 45 50 59 44 34 42 50 42 46 37 26 45 38 59 20 32 63 43 60 44 39 31 22
##  [76] 42 59 54 48 33 45 37 45 23 34 45 33 56 39 28 54 38 40 45 43 49 45 52 57 41
## [101] 48 49 52 44 46 50 49 30 51 51 34 64 38 51 23 25 52 54 36 25 40 42 39 28 41
## [126] 37 52 34 44 45 48 51 51 41 29 51 41 48 53 46 54 39 52 51 28 35 35 68 48 59
## [151] 30 47 35 37 54 47 54 37 37 33 47 29 50 39 35 55 37 35 35 39 51 42 39 41 18
## [176] 40 34 54 43 20 39 31 53 26 52 33 47 41 36 36 34 43 44 50 54 42 49 48 32 42
## [201] 41 39 52 48 35 34 46 39 57 43 34 38 45 45 42 39 36 41 28 37 50 50 36 32 42
## [226] 47 50 23 39 63 33 33 52 42 30 47 54 48 44 30 51 44 55 34 39 34 55 38 29 50
## [251] 40 46 42 56 51 41 53 52 45 42 45 46 35 37 56 42 40 24 39 46 28 42 44 36 40
## [276] 44 23 56 55 64 28 64 49 34 31 33 46 49 64 45 47 32 47 46 43 53 41 20 41 38
## [301] 41 38 33 63 52 38 41 27 41 42 55 52 40 52 38 46 34 47 29 44 37 32 42 51 40
## [326] 48 39 53 35 39 53 54 53 24 47 49 48 32 65 37 43 45 46 44 42 45 29 46 41 38
## [351] 44 42 56 51 47 52 50 56 41 45 42 44 46 60 48 52 45 43 50 38 45 30 40 32 49
## [376] 45 47 51 50 34 34 31 40 50 38 45 34 52 49 47 38 45 45 30 33 44 40 45 39 45
## [401] 41 39 43 46 59 38 34 45 50 34 38 50 41 51 46 52 41 37 53 41 56 36 36 40 38
## [426] 40 39 47 47 54 50 55 44 33 57 41 31 33 54 43 47 35 64 50 33 45 54 54 55 50
## [451] 41 35 34 42 57 42 39 42 45 45 59 55 32 27 31 38 51 49 53 39 41 39 50 52 35
## [476] 33 53 57 26 41 55 36 45 25 32 43 46 32 62 35 53 48 47 51 36 28 39 54 65 48
```

모형검토(TR)

```
TROUT <-
 TR %>% dplyr::select(low) %>%
 mutate(
 ph = predict(Mrf, type='prob')[,2],
 yh = factor(ifelse(ph>=0.5, 1, 0)))
head(TROUT)
```

```
##    low          ph yh
## 1    0 0.004878049  0
## 3    0 0.045977011  0
## 5    0 0.121212121  0
## 7    0 0.160427807  0
## 9    0 0.042424242  0
## 11   0 0.082417582  0
```

```
confusionMatrix(TROUT$yh, TROUT$low, positive='1')
```

```
## Confusion Matrix and Statistics
## 
##           Reference
## Prediction   0   1
##          0 271  49
##          1   5  53
## 
##                Accuracy : 0.8571
##                  95% CI : (0.8177, 0.8908)
##     No Information Rate : 0.7302
##     P-Value [Acc > NIR] : 2.366e-09
## 
##                   Kappa : 0.5804
## 
##  Mcnemar's Test P-Value : 4.870e-09
## 
##             Sensitivity : 0.5196
##             Specificity : 0.9819
##          Pos Pred Value : 0.9138
##          Neg Pred Value : 0.8469
##              Prevalence : 0.2698
##          Detection Rate : 0.1402
##    Detection Prevalence : 0.1534
##       Balanced Accuracy : 0.7507
## 
##        'Positive' Class : 1
## 
```

```
TR <-
TR %>%
  mutate(
    ph = predict (Mrf, type='prob')[,2],
    yh = factor(ifelse(ph>=0.5, 1, 0)))
head(TR)
```

```
##     id low  bwt age ftv race ptl smoke ht ui lwtkg          ph yh
## 1  284   0 3643  16   0    1   0     1  0  0  61.2 0.004878049  0
## 3  623   0 3175  16   0    3   0     0  0  0  49.9 0.045977011  0
## 5  400   0 2835  31   3    1   0     0  0  1  45.4 0.121212121  0
## 7  103   0 3770  24   0    3   1     0  0  0  49.9 0.160427807  0
## 9  602   0 2977  25   0    2   0     0  0  0  56.7 0.042424242  0
## 11  79   0 3444  20   0    2   0     1  0  0  54.9 0.082417582  0
```

모형평가(TS)

```
TSOUT <-
TS %>%
mutate(yh=predict(Mrf, TS), e=bwt-yh)
```

```
## Warning in Ops.factor(bwt, yh): 요인(factors)에 대하여 의미있는 '-'가 아닙니다.
```

```
head(TSOUT)
```

```
##     id low  bwt age ftv race ptl smoke ht ui lwtkg yh  e
## 2  101   0 3728  24   1    1   0     0  0  0  49.9  0 NA
## 4  645   0 3430  32   4    1   1     1  0  0  60.8  0 NA
## 6   98   0 3651  19   0    1   0     1  0  0  66.7  0 NA
## 8  726   1 2187  27   0    2   0     0  0  1  59.0  1 NA
## 10 326   1 1588  23   1    3   0     0  0  1  44.0  0 NA
## 12 270   0 3460  22   1    1   0     0  0  0  59.4  0 NA
```

```
mean(TSOUT$e^2) # MSE
```

```
## [1] NA
```

```
mean(abs(TSOUT$e)) # MAE
```

```
## [1] NA
```

```
TSOUT %>% summarize(mn=mean(e), sd=sd(e), min=min(e), max=max(e))
```

```
##   mn sd min max
## 1 NA NA  NA  NA
```

```
summary(TSOUT$e)
```

```
##    Mode    NA's
## logical     378
```

스코어

```
SC <- read.csv(text='
age,ftv,race,ptl,smoke,ht,ui,lwtkg
30,0,3,0,1,0,0,60
40,0,3,0,1,0,0,60
30,0,3,0,0,0,0,60
40,0,3,0,0,0,0,60
30,0,3,0,1,1,0,60
40,0,3,0,1,1,0,60
30,0,3,0,0,1,0,60
40,0,3,0,0,1,0,60
30,0,3,0,1,1,1,60
40,0,3,0,1,1,1,60
30,0,3,0,0,1,1,60
40,0,3,0,0,1,1,60
')
SC <- SC %>% mutate(race=factor(race, levels=1:3))

SCOUT <-
  SC %>%
  mutate(
    ph = predict(Mrf, SC, type='prob')[,2],
    yh = factor(ifelse(ph>=0.5, 1, 0)))
SCOUT
```

```
##    age ftv race ptl smoke ht ui lwtkg    ph yh
## 1   30   0    3   0     1  0  0    60 0.122  0
## 2   40   0    3   0     1  0  0    60 0.200  0
## 3   30   0    3   0     0  0  0    60 0.048  0
## 4   40   0    3   0     0  0  0    60 0.092  0
## 5   30   0    3   0     1  1  0    60 0.590  1
## 6   40   0    3   0     1  1  0    60 0.654  1
## 7   30   0    3   0     0  1  0    60 0.512  1
## 8   40   0    3   0     0  1  0    60 0.532  1
## 9   30   0    3   0     1  1  1    60 0.666  1
## 10  40   0    3   0     1  1  1    60 0.696  1
## 11  30   0    3   0     0  1  1    60 0.616  1
## 12  40   0    3   0     0  1  1    60 0.608  1
```