Untitled

Wenhao Pan

12/6/2021

```
library(dplyr)
```

```
##
## Attaching package: 'dplyr'
```

```
## The following objects are masked from 'package:stats':
##
##     filter, lag
```

```
## The following objects are masked from 'package:base':
##
##     intersect, setdiff, setequal, union
```

```
library(ggplot2)
train <- read.csv("data/Train.csv")
test <- read.csv("data/test.csv")
head(train)
```

```
##   ATTEND BFACIL  BMI CIG_0 DBWT DMAR FAGECOMB FEDUC FRACE6 LD_INDL MAGER
## 1      1      1 22.3 FALSE 3572    2       23     4      1       N    22
## 2      1      1 28.3 FALSE 3355    1       31     6      5       N    31
## 3      1      1 22.3 FALSE 3550    1       39     1      1       Y    38
## 4      1      1 30.2 FALSE 3190    1       32     7      1       Y    36
## 5      1      1 21.9 FALSE 2725    1       32     4      1       N    32
## 6      1      1 18.9 FALSE 3577    1       26     3      2       N    29
##   MBSTATE_REC MEDUC MRAVE6 M_Ht_In NO_INFEC NO_MMORB NO_RISKS PAY_REC PRECARE
## 1           1     5      1      64        1        1        1       2       1
## 2           2     6      4      64        1        1        0       2       1
## 3           2     2      1      60        1        1        0       1       1
## 4           1     7      1      67        1        1        1       2       1
## 5           1     7      1      67        1        1        1       2       1
## 6           1     6      2      67        1        1        0       2       1
##   PREVIS PRIORDEAD PRIORLIVE PRIORTERM RDMETH_REC RESTATUS RF_CESAR SEX
## 1     11     FALSE     FALSE     FALSE          1        1        N   M
## 2     13     FALSE      TRUE      TRUE          1        2        N   M
## 3     18     FALSE      TRUE      TRUE          1        2        N   F
## 4      9     FALSE     FALSE      TRUE          1        2        N   M
## 5     11     FALSE     FALSE     FALSE          1        2        N   M
## 6     12     FALSE      TRUE      TRUE          4        1        Y   F
##   PREG_LEN WTGAIN_PER FIRST_BIRTH
```

```
## 1        9  0.3923077        TRUE
## 2        9  0.1393939       FALSE
## 3        8  0.2543860       FALSE
## 4       10  0.1036269        TRUE
## 5        8  0.1571429        TRUE
## 6       10  0.1735537       FALSE
```

```
# Factorize categorical variables
fac_train <- train %>% mutate_if(is.character, as.factor)
fac_train <- fac_train %>% mutate_if(is.logical, as.factor)
fac_train <- fac_train %>% mutate(ATTEND = factor(ATTEND), BFACIL = factor(BFACIL),
                                  DMAR = factor(DMAR), FEDUC = factor(FEDUC),
                                  FRACE6 = factor(FRACE6), MBSTATE_REC = factor(MBSTATE_REC),
                                  MEDUC = factor(MEDUC), MRAVE6 = factor(MRAVE6),
                                  NO_INFEC = factor(NO_INFEC),NO_MMORB = factor(NO_MMORB),
                                  NO_RISKS = factor(NO_RISKS), PAY_REC = factor(PAY_REC),
                                  PRECARE = factor(PRECARE), RDMETH_REC = factor(RDMETH_REC),
                                  RESTATUS = factor(RESTATUS))

fac_test <- test %>% mutate_if(is.character, as.factor)
fac_test <- fac_test %>% mutate_if(is.logical, as.factor)
fac_test <- fac_test %>% mutate(ATTEND = factor(ATTEND), BFACIL = factor(BFACIL),
                                  DMAR = factor(DMAR), FEDUC = factor(FEDUC),
                                  FRACE6 = factor(FRACE6), MBSTATE_REC = factor(MBSTATE_REC),
                                  MEDUC = factor(MEDUC), MRAVE6 = factor(MRAVE6),
                                  NO_INFEC = factor(NO_INFEC),NO_MMORB = factor(NO_MMORB),
                                  NO_RISKS = factor(NO_RISKS), PAY_REC = factor(PAY_REC),
                                  PRECARE = factor(PRECARE), RDMETH_REC = factor(RDMETH_REC),
                                  RESTATUS = factor(RESTATUS))
```

```
best_model.train <- lm(formula = DBWT ~ ATTEND + BFACIL + BMI + CIG_0 + DMAR + FAGECOMB +
    FEDUC + FRACE6 + LD_INDL + MBSTATE_REC + MEDUC + MRAVE6 +
    M_Ht_In + NO_INFEC + NO_MMORB + NO_RISKS + PAY_REC + PRECARE +
    PREVIS + PRIORLIVE + RDMETH_REC + RESTATUS + SEX + PREG_LEN +
    WTGAIN_PER, data = fac_train)

best_model.test <- lm(formula = DBWT ~ ATTEND + BFACIL + BMI + CIG_0 + DMAR + FAGECOMB +
    FEDUC + FRACE6 + LD_INDL + MBSTATE_REC + MEDUC + MRAVE6 +
    M_Ht_In + NO_INFEC + NO_MMORB + NO_RISKS + PAY_REC + PRECARE +
    PREVIS + PRIORLIVE + RDMETH_REC + RESTATUS + SEX + PREG_LEN +
    WTGAIN_PER, data = fac_test)
```

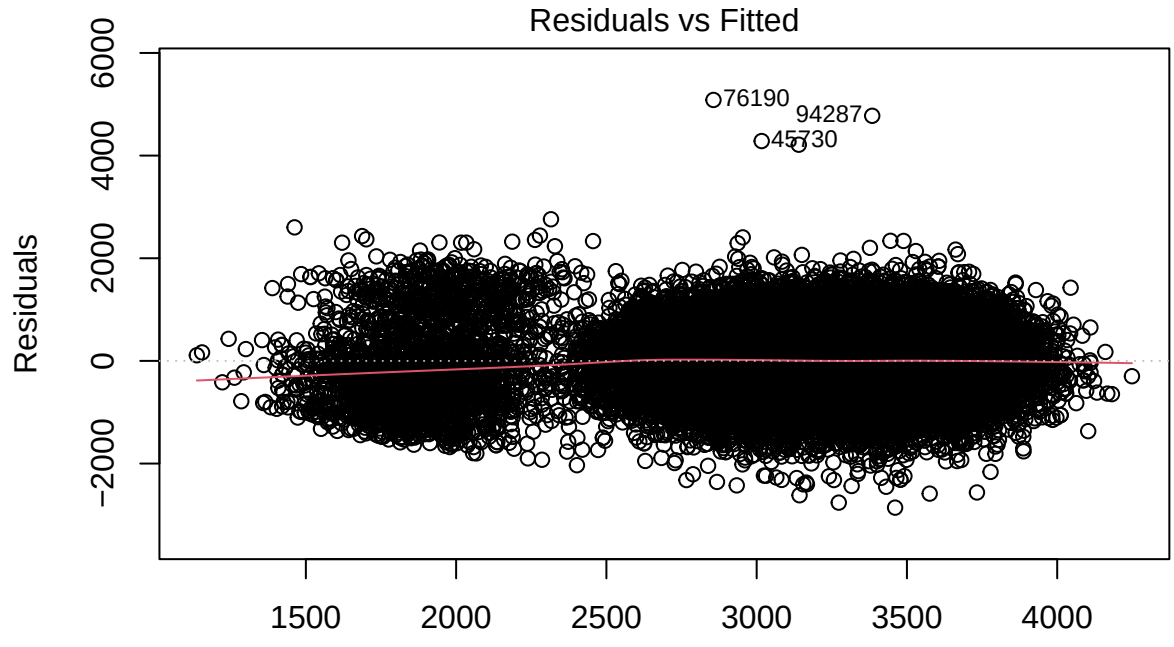
```
# Model interpretation
summary(best_model.test)
```

```
##
## Call:
## lm(formula = DBWT ~ ATTEND + BFACIL + BMI + CIG_0 + DMAR + FAGECOMB +
##     FEDUC + FRACE6 + LD_INDL + MBSTATE_REC + MEDUC + MRAVE6 +
##     M_Ht_In + NO_INFEC + NO_MMORB + NO_RISKS + PAY_REC + PRECARE +
##     PREVIS + PRIORLIVE + RDMETH_REC + RESTATUS + SEX + PREG_LEN +
##     WTGAIN_PER, data = fac_test)
##
```

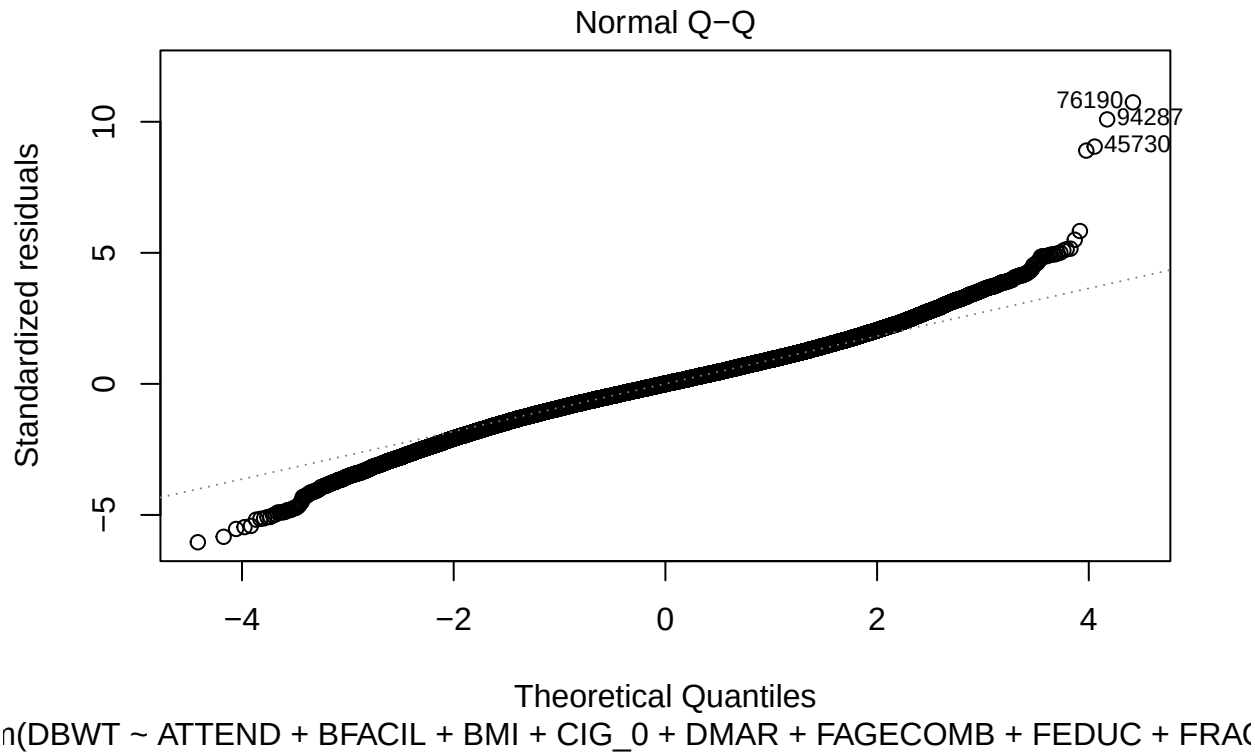
```
## Residuals:
##     Min      1Q  Median      3Q     Max
## -3234.4  -286.7    -0.4   293.4  4748.5
##
## Coefficients:
##                 Estimate Std. Error  t value Pr(>|t|)
## (Intercept)     735.8315    45.8617   16.045  < 2e-16 ***
## ATTEND2           1.1668     5.4311    0.215 0.829889
## ATTEND3          50.2761     5.4714    9.189  < 2e-16 ***
## ATTEND4          64.9249    20.5491    3.160 0.001581 **
## ATTEND5          41.4274    19.1359    2.165 0.030398 *
## BFACIL2          61.0096    19.4925    3.130 0.001749 **
## BFACIL3          98.2900    20.1982    4.866 1.14e-06 ***
## BFACIL4        -128.3833    50.6455   -2.535 0.011248 *
## BFACIL5           1.4782   137.1424    0.011 0.991400
## BFACIL6          33.1712   101.2269    0.328 0.743146
## BFACIL7          28.6799    57.2432    0.501 0.616360
## BMI              17.6336     0.2893   60.959  < 2e-16 ***
## CIG_0TRUE      -112.0420     5.9159  -18.939  < 2e-16 ***
## DMAR2           -14.5697     4.0298   -3.616 0.000300 ***
## FAGECOMB         -0.7815     0.2599   -3.006 0.002645 **
## FEDUC2          -33.5748    12.3492   -2.719 0.006553 **
## FEDUC3          -12.8980    11.8490   -1.089 0.276366
## FEDUC4           -0.0240    12.2021   -0.002 0.998430
## FEDUC5           14.9161    12.9615    1.151 0.249816
## FEDUC6           23.8039    12.4563    1.911 0.056008 .
## FEDUC7           20.7946    13.2769    1.566 0.117298
## FEDUC8           23.2517    14.7803    1.573 0.115687
## FRACE62         -59.0333     7.4136   -7.963 1.70e-15 ***
## FRACE63          25.9479    19.9766    1.299 0.193977
## FRACE64        -129.4172    11.1927  -11.563  < 2e-16 ***
## FRACE65          34.1317    35.8022    0.953 0.340420
## FRACE66         -34.6185    10.1885   -3.398 0.000680 ***
## LD_INDLY         33.7033     3.4875    9.664  < 2e-16 ***
## MBSTATE_REC2     50.9542     4.6024   11.071  < 2e-16 ***
## MEDUC2          -49.8276    13.3999   -3.718 0.000201 ***
## MEDUC3          -46.9394    12.9213   -3.633 0.000281 ***
## MEDUC4          -48.4026    13.1743   -3.674 0.000239 ***
## MEDUC5          -43.8374    13.7365   -3.191 0.001417 **
## MEDUC6          -36.6991    13.4435   -2.730 0.006337 **
## MEDUC7          -41.5035    14.0175   -2.961 0.003069 **
## MEDUC8          -44.1664    15.8629   -2.784 0.005366 **
## MRAVE62        -117.0954     7.8266  -14.961  < 2e-16 ***
## MRAVE63          56.3258    19.3623    2.909 0.003626 **
## MRAVE64         -15.1501    11.0018   -1.377 0.168498
## MRAVE65         -57.7706    37.2961   -1.549 0.121392
## MRAVE66         -41.0986    10.2077   -4.026 5.67e-05 ***
## M_Ht_In          30.2528     0.5626   53.771  < 2e-16 ***
## NO_INFEC1        10.1860    10.6451    0.957 0.338635
## NO_MMORB1       -85.7331    12.4763   -6.872 6.38e-12 ***
## NO_RISKS1       136.7509     4.3092   31.735  < 2e-16 ***
## PAY_REC2         21.6449     4.0891    5.293 1.20e-07 ***
## PAY_REC3         30.3789     8.9678    3.388 0.000705 ***
## PAY_REC4         23.2334     8.4093    2.763 0.005731 **
```

```
## PRECARE1        -74.2166    16.8008   -4.417 1.00e-05 ***
## PRECARE2        -29.3632    16.8015   -1.748 0.080527 .
## PRECARE3         32.4566    17.8112    1.822 0.068419 .
## PREVIS           10.1167     0.4451   22.731  < 2e-16 ***
## PRIORLIVETRUE   104.4111     3.5127   29.724  < 2e-16 ***
## RDMETH_REC2     114.5958    11.2779   10.161  < 2e-16 ***
## RDMETH_REC3     -40.4676     4.1799   -9.681  < 2e-16 ***
## RDMETH_REC4     126.3587     6.1697   20.480  < 2e-16 ***
## RESTATUS2       -15.6220     3.3105   -4.719 2.37e-06 ***
## RESTATUS3       -22.6633     9.1419   -2.479 0.013175 *
## RESTATUS4       -62.5508    30.5216   -2.049 0.040426 *
## SEXM            117.0748     2.9964   39.071  < 2e-16 ***
## PREG_LEN8      -531.5709     5.8048  -91.574  < 2e-16 ***
## PREG_LEN9      -112.3346     4.3056  -26.091  < 2e-16 ***
## PREG_LENEarly -1466.5952    10.8932 -134.634  < 2e-16 ***
## PREG_LENLate   -110.9548    15.6498   -7.090 1.35e-12 ***
## WTGAIN_PER      894.2647    16.3115   54.824  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 473 on 99935 degrees of freedom
## Multiple R-squared:  0.3238, Adjusted R-squared:  0.3234
## F-statistic: 747.7 on 64 and 99935 DF,  p-value: < 2.2e-16
```

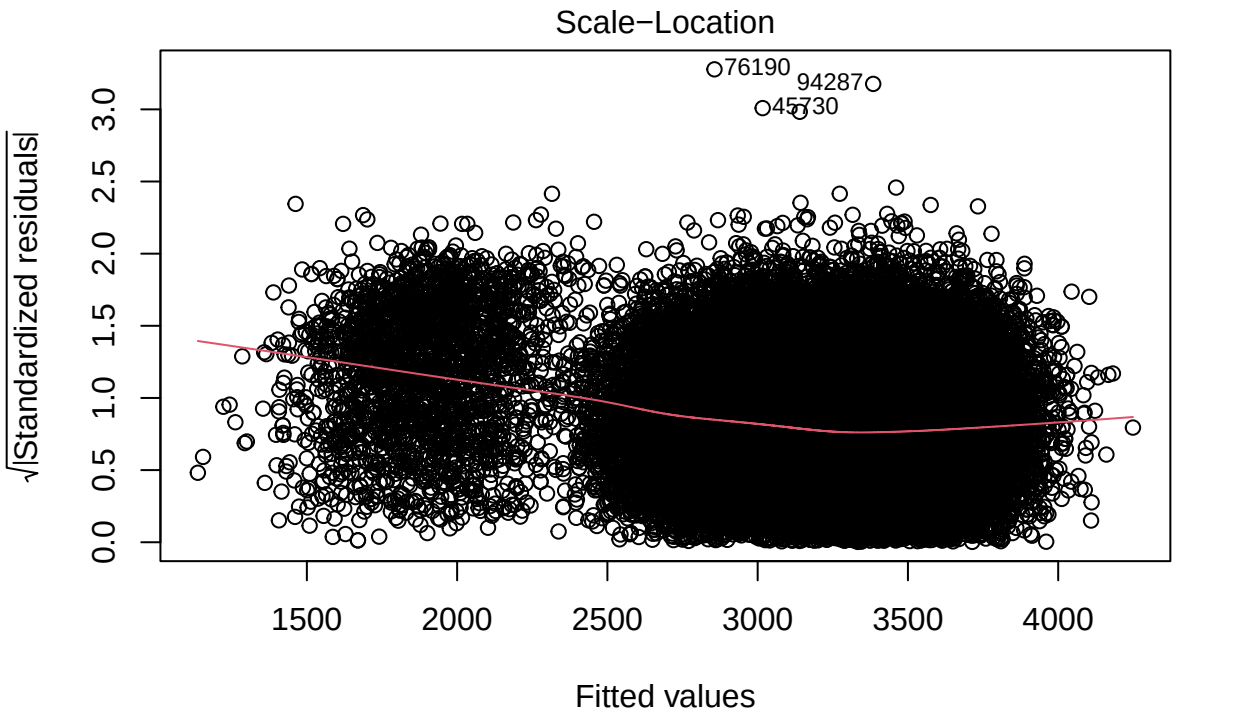
```
# Model diagnostic
plot(best_model.train)
```



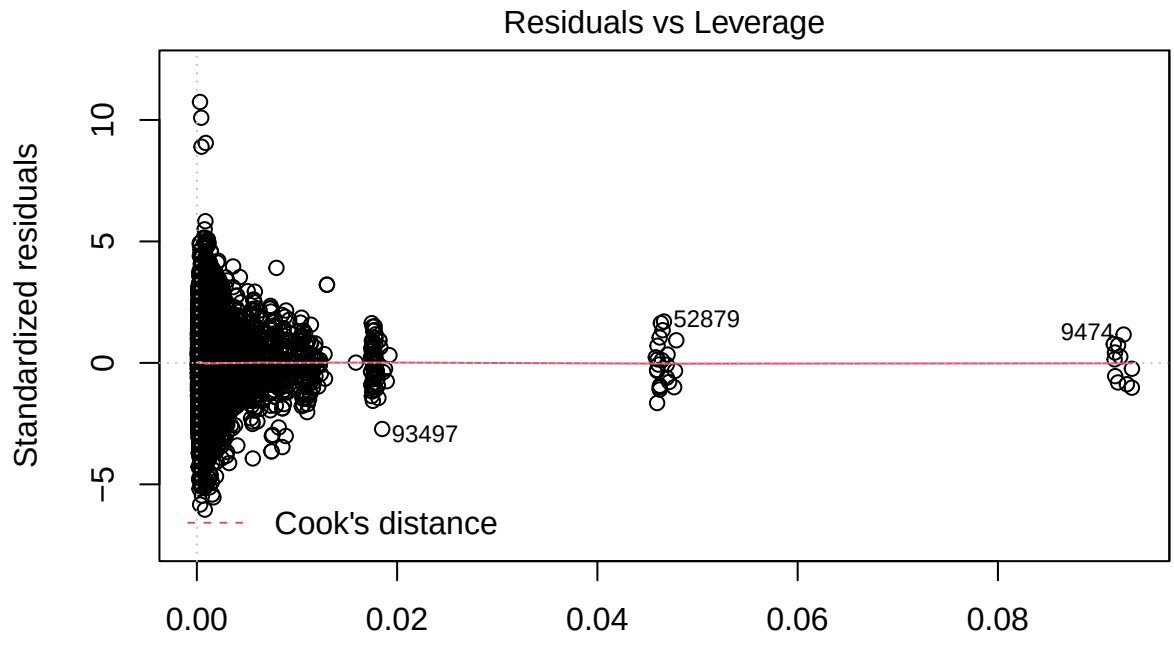
n(DBWT ~ ATTEND + BFACIL + BMI + CIG_0 + DMAR + FAGECOMB + FEDUC + FRAC



n(DBWT ~ ATTEND + BFACIL + BMI + CIG_0 + DMAR + FAGECOMB + FEDUC + FRAC



n(DBWT ~ ATTEND + BFACIL + BMI + CIG_0 + DMAR + FAGECOMB + FEDUC + FRAC



n(DBWT ~ ATTEND + BFACIL + BMI + CIG_0 + DMAR + FAGECOMB + FEDUC + FRA(

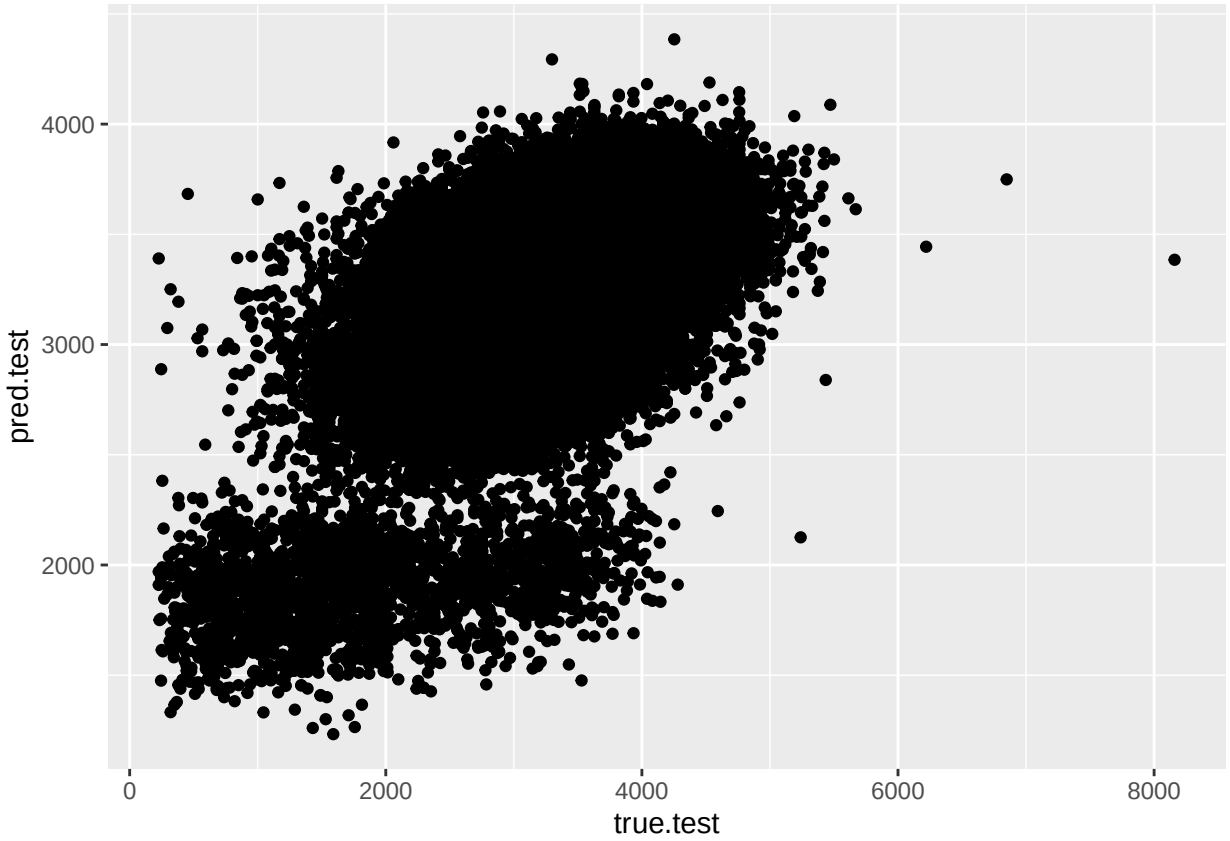
```
# Model Prediction
MSE.train <- mean(best_model.train$residuals ^ 2)
pred.test <- predict(best_model.train, fac_test)
MSE.test <- mean((pred.test - fac_test$DBWT) ^ 2)
MSE.train
```

```
## [1] 223944.6
```

```
MSE.test
```

```
## [1] 223894.9
```

```
true.test <- fac_test$DBWT
test.pred.df <- data.frame(cbind(true.test, pred.test))
ggplot(test.pred.df, aes(x = true.test, y = pred.test)) +
  geom_point()
```


4000
3000
2000
pred.test
0
2000
4000
6000
8000
true.test