> setwd('C:/Rdirectory/data\_mining/data\_mining\_covid');

> covid\_train = read.csv("covid\_train.csv", header=T);

> str(covid\_train);

'data.frame': 32356 obs. of 16 variables:

$ sex : int 1 1 1 1 1 1 1 1 1 1 ...

$ patient\_type : int 2 2 2 2 2 2 2 2 2 2 ...

$ pneumonia : int 2 2 1 1 1 2 1 1 1 2 ...

$ age : num 25 52 51 67 59 52 54 78 80 40 ...

$ diabetes : int 2 2 2 1 1 1 2 2 1 2 ...

$ copd : int 2 2 2 2 2 2 2 2 2 2 ...

$ asthma : int 2 2 2 2 2 2 2 2 2 2 ...

$ inmsupr : int 2 2 2 2 2 2 2 2 2 2 ...

$ hypertension : int 2 2 2 1 2 1 2 1 1 2 ...

$ other\_disease : int 2 2 2 2 2 2 2 2 2 2 ...

$ cardiovascular : int 2 2 2 2 2 1 2 2 2 2 ...

$ obesity : int 2 1 2 1 2 2 2 1 1 2 ...

$ renal\_chronic : int 2 2 2 2 2 2 2 2 2 2 ...

$ tobacco : int 2 1 2 2 2 2 2 2 2 2 ...

$ contact\_other\_covid: int 1 1 1 1 1 2 2 1 1 2 ...

$ is\_dead : int 2 2 2 2 2 2 2 1 2 2 ...

> head(covid\_train);

sex patient\_type pneumonia age diabetes copd asthma inmsupr hypertension other\_disease cardiovascular obesity renal\_chronic

1 1 2 2 25 2 2 2 2 2 2 2 2 2

2 1 2 2 52 2 2 2 2 2 2 2 1 2

3 1 2 1 51 2 2 2 2 2 2 2 2 2

4 1 2 1 67 1 2 2 2 1 2 2 1 2

5 1 2 1 59 1 2 2 2 2 2 2 2 2

6 1 2 2 52 1 2 2 2 1 2 1 2 2

tobacco contact\_other\_covid is\_dead

1 2 1 2

2 1 1 2

3 2 1 2

4 2 1 2

5 2 1 2

6 2 2 2

> covid\_test = read.csv("covid\_test.csv", header=T);

> str(covid\_test);

'data.frame': 5920 obs. of 16 variables:

$ sex : int 1 1 1 1 1 1 1 1 1 1 ...

$ patient\_type : int 2 2 2 2 2 2 2 2 2 2 ...

$ pneumonia : int 2 2 1 1 1 1 2 2 2 1 ...

$ age : num 52 36 0 85 75 19 46 60 62 39 ...

$ diabetes : int 2 2 2 2 2 2 2 2 1 2 ...

$ copd : int 2 2 2 1 2 2 2 2 2 2 ...

$ asthma : int 2 2 2 2 2 2 2 2 2 2 ...

$ inmsupr : int 2 2 2 2 2 2 2 2 2 2 ...

$ hypertension : int 2 2 2 1 1 2 2 2 2 1 ...

$ other\_disease : int 1 2 2 2 2 2 2 2 2 2 ...

$ cardiovascular : int 2 2 2 1 2 2 2 2 2 2 ...

$ obesity : int 2 2 2 1 2 2 2 2 2 2 ...

$ renal\_chronic : int 2 2 2 2 2 2 2 2 2 2 ...

$ tobacco : int 2 2 2 2 2 2 2 2 2 2 ...

$ contact\_other\_covid: int 2 2 2 2 2 2 1 2 2 2 ...

$ is\_dead : int 2 2 1 2 2 2 2 2 2 2 ...

> head(covid\_test);

sex patient\_type pneumonia age diabetes copd asthma inmsupr hypertension other\_disease cardiovascular obesity renal\_chronic

1 1 2 2 52 2 2 2 2 2 1 2 2 2

2 1 2 2 36 2 2 2 2 2 2 2 2 2

3 1 2 1 0 2 2 2 2 2 2 2 2 2

4 1 2 1 85 2 1 2 2 1 2 1 1 2

5 1 2 1 75 2 2 2 2 1 2 2 2 2

6 1 2 1 19 2 2 2 2 2 2 2 2 2

tobacco contact\_other\_covid is\_dead

1 2 2 2

2 2 2 2

3 2 2 1

4 2 2 2

5 2 2 2

6 2 2 2

> covid\_dead\_train = read.csv("covid\_dead\_train.csv", header=T);

> str(covid\_dead\_train);

'data.frame': 3600 obs. of 16 variables:

$ sex : int 1 1 1 1 1 1 1 1 1 1 ...

$ patient\_type : int 2 2 2 2 2 2 2 2 2 2 ...

$ pneumonia : int 1 1 2 1 1 1 1 1 1 2 ...

$ age : int 78 65 58 78 92 70 71 65 56 72 ...

$ diabetes : int 2 2 2 2 2 2 1 1 2 2 ...

$ copd : int 2 2 2 2 2 2 1 2 2 2 ...

$ asthma : int 2 2 2 2 2 2 2 2 2 2 ...

$ inmsupr : int 2 2 2 2 2 2 2 2 2 2 ...

$ hypertension : int 1 2 2 2 1 2 1 1 2 1 ...

$ other\_disease : int 2 2 1 2 2 2 2 2 2 2 ...

$ cardiovascular : int 2 2 2 2 2 2 2 2 2 2 ...

$ obesity : int 1 1 2 1 2 1 1 1 2 2 ...

$ renal\_chronic : int 2 2 2 2 2 2 2 2 2 1 ...

$ tobacco : int 2 2 2 2 2 2 1 2 2 2 ...

$ contact\_other\_covid: int 1 2 1 1 2 2 1 2 2 2 ...

$ day\_cnt : int 10 7 1 6 3 16 21 14 30 20 ...

> head(covid\_dead\_train);

sex patient\_type pneumonia age diabetes copd asthma inmsupr hypertension other\_disease cardiovascular obesity renal\_chronic

1 1 2 1 78 2 2 2 2 1 2 2 1 2

2 1 2 1 65 2 2 2 2 2 2 2 1 2

3 1 2 2 58 2 2 2 2 2 1 2 2 2

4 1 2 1 78 2 2 2 2 2 2 2 1 2

5 1 2 1 92 2 2 2 2 1 2 2 2 2

6 1 2 1 70 2 2 2 2 2 2 2 1 2

tobacco contact\_other\_covid day\_cnt

1 2 1 10

2 2 2 7

3 2 1 1

4 2 1 6

5 2 2 3

6 2 2 16

> covid\_dead\_test = read.csv("covid\_dead\_test.csv", header=T);

> str(covid\_dead\_test);

'data.frame': 420 obs. of 16 variables:

$ sex : int 1 1 1 1 1 1 1 1 1 1 ...

$ patient\_type : int 2 2 2 2 2 2 2 2 2 2 ...

$ pneumonia : int 1 1 1 1 1 1 1 1 1 1 ...

$ age : int 66 68 49 61 47 56 68 72 78 64 ...

$ diabetes : int 1 2 2 1 1 2 1 1 1 1 ...

$ copd : int 2 2 2 2 2 2 2 1 2 2 ...

$ asthma : int 2 2 2 2 1 2 2 2 2 2 ...

$ inmsupr : int 2 2 2 2 2 2 2 2 2 2 ...

$ hypertension : int 1 2 2 1 2 2 1 1 2 2 ...

$ other\_disease : int 2 2 2 2 2 2 2 2 2 2 ...

$ cardiovascular : int 2 2 2 2 2 2 2 2 2 2 ...

$ obesity : int 1 2 1 2 2 2 1 1 1 2 ...

$ renal\_chronic : int 2 2 2 2 2 2 2 2 2 2 ...

$ tobacco : int 2 2 2 2 2 2 2 2 2 2 ...

$ contact\_other\_covid: int 1 2 2 2 2 2 2 2 2 2 ...

$ day\_cnt : int 3 5 16 4 18 10 9 6 6 8 ...

> head(covid\_dead\_test);

sex patient\_type pneumonia age diabetes copd asthma inmsupr hypertension other\_disease cardiovascular obesity renal\_chronic

1 1 2 1 66 1 2 2 2 1 2 2 1 2

2 1 2 1 68 2 2 2 2 2 2 2 2 2

3 1 2 1 49 2 2 2 2 2 2 2 1 2

4 1 2 1 61 1 2 2 2 1 2 2 2 2

5 1 2 1 47 1 2 1 2 2 2 2 2 2

6 1 2 1 56 2 2 2 2 2 2 2 2 2

tobacco contact\_other\_covid day\_cnt

1 2 1 3

2 2 2 5

3 2 2 16

4 2 2 4

5 2 2 18

6 2 2 10

> regressionFit = lm(is\_dead ~ ., data=covid\_train);

> summary(regressionFit);

Call:

lm(formula = is\_dead ~ ., data = covid\_train)

Residuals:

Min 1Q Median 3Q Max

-1.1267 -0.3810 -0.1290 0.4417 0.9735

Coefficients: (2 not defined because of singularities)

Estimate Std. Error t value Pr(>|t|)

(Intercept) 1.238965 0.057632 21.498 < 2e-16 \*\*\*

sex NA NA NA NA

patient\_type NA NA NA NA

pneumonia 0.308114 0.006036 51.049 < 2e-16 \*\*\*

age -0.005592 0.000143 -39.095 < 2e-16 \*\*\*

diabetes 0.059541 0.005754 10.348 < 2e-16 \*\*\*

copd -0.016383 0.011281 -1.452 0.14644

asthma -0.014946 0.014270 -1.047 0.29493

inmsupr 0.017241 0.013079 1.318 0.18743

hypertension 0.030859 0.005948 5.189 2.13e-07 \*\*\*

other\_disease 0.032470 0.012166 2.669 0.00761 \*\*

cardiovascular 0.001250 0.010668 0.117 0.90669

obesity 0.033247 0.005850 5.683 1.33e-08 \*\*\*

renal\_chronic 0.062355 0.010855 5.744 9.32e-09 \*\*\*

tobacco -0.035682 0.012384 -2.881 0.00396 \*\*

contact\_other\_covid -0.066049 0.005663 -11.662 < 2e-16 \*\*\*

---

Signif. codes: 0 ‘\*\*\*’ 0.001 ‘\*\*’ 0.01 ‘\*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.451 on 32342 degrees of freedom

Multiple R-squared: 0.1867, Adjusted R-squared: 0.1864

F-statistic: 571.1 on 13 and 32342 DF, p-value: < 2.2e-16

> prediction = predict(regressionFit, covid\_test);

Warning message:

In predict.lm(regressionFit, covid\_test) :

prediction from a rank-deficient fit may be misleading

> summary(prediction);

Min. 1st Qu. Median Mean 3rd Qu. Max.

1.073 1.360 1.494 1.538 1.705 2.129

> prediction = round(prediction);

> comparison=cbind(covid\_test,prediction);

> comparison=as.data.frame(comparison);

> print(paste("test 건수 : ",nrow(covid\_test)));

[1] "test 건수 : 5920"

> predictCorrect = comparison[comparison$is\_dead == comparison$prediction,];

> print(paste("사망여부 예측성공 건수 : ", nrow(predictCorrect)));

[1] "사망여부 예측성공 건수 : 3617"

> print(paste("사망여부 예측 정확도 : " ,nrow(predictCorrect)/nrow(covid\_test))); # 61.1%

[1] "사망여부 예측 정확도 : 0.61097972972973"

> regressionFit\_dead = lm(day\_cnt ~ ., data=covid\_dead\_train);

> summary(regressionFit\_dead);

Call:

lm(formula = day\_cnt ~ ., data = covid\_dead\_train)

Residuals:

Min 1Q Median 3Q Max

-11.840 -5.118 -1.803 3.269 43.002

Coefficients: (2 not defined because of singularities)

Estimate Std. Error t value Pr(>|t|)

(Intercept) 6.789655 2.815601 2.411 0.01594 \*

sex NA NA NA NA

patient\_type NA NA NA NA

pneumonia -1.146756 0.420694 -2.726 0.00644 \*\*

age -0.005432 0.008215 -0.661 0.50855

diabetes 0.371892 0.264079 1.408 0.15914

copd 0.748439 0.510778 1.465 0.14293

asthma 0.089652 0.765361 0.117 0.90676

inmsupr 1.076194 0.672527 1.600 0.10964

hypertension 0.197530 0.272466 0.725 0.46852

other\_disease -1.026140 0.606404 -1.692 0.09070 .

cardiovascular 0.734718 0.490061 1.499 0.13390

obesity -0.024487 0.273988 -0.089 0.92879

renal\_chronic 1.587909 0.500762 3.171 0.00153 \*\*

tobacco -0.244416 0.633829 -0.386 0.69980

contact\_other\_covid -0.477998 0.292047 -1.637 0.10178

---

Signif. codes: 0 ‘\*\*\*’ 0.001 ‘\*\*’ 0.01 ‘\*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 7.422 on 3586 degrees of freedom

Multiple R-squared: 0.01077, Adjusted R-squared: 0.007186

F-statistic: 3.004 on 13 and 3586 DF, p-value: 0.0002075

> prediction\_dead = predict(regressionFit\_dead, covid\_dead\_test);

Warning message:

In predict.lm(regressionFit\_dead, covid\_dead\_test) :

prediction from a rank-deficient fit may be misleading

> summary(prediction\_dead);

Min. 1st Qu. Median Mean 3rd Qu. Max.

7.363 10.231 10.919 10.734 11.378 12.730

> prediction\_dead = round(prediction\_dead);

> comparison\_dead=cbind(covid\_dead\_test,prediction\_dead);

> comparison\_dead=as.data.frame(comparison\_dead);

> print(paste("test 건수 : ", nrow(covid\_dead\_test)));

[1] "test 건수 : 420"

> # 투병일수 예측성공 기준 설정

> deadPredictCorrectCreteria = 10;

> deadPredictCorrect = comparison\_dead[abs(comparison\_dead$day\_cnt-comparison\_dead$prediction\_dead)<=deadPredictCorrectCreteria, 0];

> print(paste("투병일수 예측성공 건수(",deadPredictCorrectCreteria,"일) : " , nrow(deadPredictCorrect)));

[1] "투병일수 예측성공 건수( 10 일) : 387"

> print(paste("투병일수 예측 정확도(",deadPredictCorrectCreteria,"일) : ", nrow(deadPredictCorrect) / nrow(covid\_dead\_test)));

[1] "투병일수 예측 정확도( 10 일) : 0.921428571428571"

>