We spilt the data sets into two parts:

Surgical Unit Training Data: We will build and develop our model.

Surgical Unit Holdout Data: We will test our model.

There are 8 variables in the model, we will

> round(cor(SurgicalUnit\_training),2)

X1 X2 X3 X4 X5 X6 X7 X8 Y lnY

X1 1.00 0.09 -0.15 0.50 -0.02 0.04 -0.10 0.22 0.35 0.25

X2 0.09 1.00 -0.02 0.37 -0.05 0.12 0.13 -0.08 0.42 0.47

X3 -0.15 -0.02 1.00 0.42 -0.01 0.14 -0.09 0.12 0.58 0.65

X4 0.50 0.37 0.42 1.00 -0.21 0.30 -0.02 0.13 0.67 0.65

X5 -0.02 -0.05 -0.01 -0.21 1.00 0.01 0.15 -0.11 -0.12 -0.14

X6 0.04 0.12 0.14 0.30 0.01 1.00 0.04 -0.06 0.17 0.23

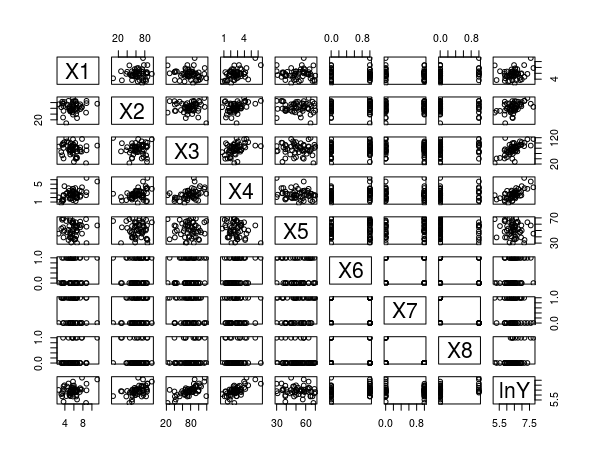
X7 -0.10 0.13 -0.09 -0.02 0.15 0.04 1.00 -0.51 -0.18 -0.13

X8 0.22 -0.08 0.12 0.13 -0.11 -0.06 -0.51 1.00 0.42 0.37

Y 0.35 0.42 0.58 0.67 -0.12 0.17 -0.18 0.42 1.00 0.93

lnY 0.25 0.47 0.65 0.65 -0.14 0.23 -0.13 0.37 0.93 1.00

No sign of multicollinearity, X2,X3 and X4 are highly correlated with lnY.



There are couple of points look like out outliers, this needs to be examined

during the model building process. We will start with StepWise.

install.packages("olsrr")

library(olsrr)

model <- **lm**(lnY ~ X1+X2+X3+X4+X5+X6+X7+X8, data = SurgicalUnit\_training)

f1 <- **ols\_step\_all\_possible**(model)

**plot**(f1)

install.packages("leaps")

library(leaps)

ex<- SurgicalUnit\_training

ex.r2<-leaps( x=cbind(X1,X2,X3,X4,X5,X6,X7,X8),y=lnY, method='r2', nbest=6)

p<-seq( min(ex.r2$size),max(ex.r2$size) )

ind<-as.data.frame(ex.r2[c(3:4)])

ind<-ind[with(ind, order(size,r2)), ]

plot(ind[,c(1:2)] ,ylab=expression(R^2), xlab='p' ,col="red",pch=16)

Rp2 = by( data=ex.r2[4],INDICES=factor(ex.r2$size), FUN=max)

lines( Rp2 ~ p,col="blue" )

b<-regsubsets(lnY ~ X1+X2+X3+X4+X5+X6+X7+X8, data = SurgicalUnit\_training)

rs<-summary(b)

rs$which

|  |
| --- |
| > rs$which  (Intercept) X1 X2 X3 X4 X5 X6 X7 X8  1 TRUE FALSE FALSE TRUE FALSE FALSE FALSE FALSE FALSE  2 TRUE FALSE TRUE TRUE FALSE FALSE FALSE FALSE FALSE  3 TRUE FALSE TRUE TRUE FALSE FALSE FALSE FALSE TRUE  4 TRUE TRUE TRUE TRUE FALSE FALSE FALSE FALSE TRUE  5 TRUE TRUE TRUE TRUE FALSE FALSE TRUE FALSE TRUE  6 TRUE TRUE TRUE TRUE FALSE TRUE TRUE FALSE TRUE  7 TRUE TRUE TRUE TRUE FALSE TRUE TRUE TRUE TRUE  8 TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE |
|  |
| |  | | --- | | The best one predictor is X3. Lets calculate AIC  > AICp<- 54\*log(rs$rss/54)+(2:9)\*2  > plot(AICp~I(1:8),ylab="AIC",xlab="Number of Predictors") | |

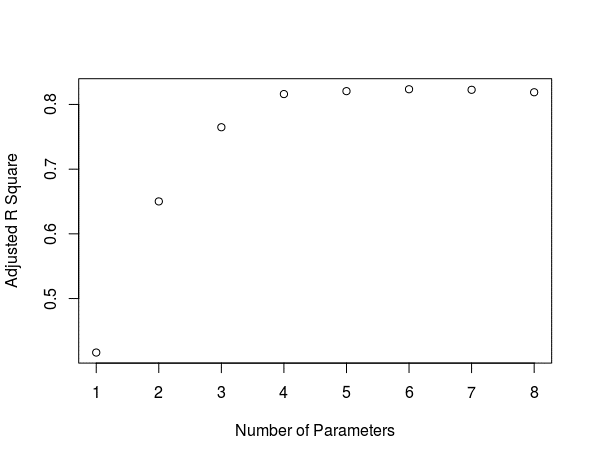
We see in that the AIC is minimized by a choice of four predictors, X1,X2,X3, and X8 as determined by the logical matrix above.

Another measure is widely used is adjusted R square.

plot(1:8,rs$adjr2, ylab="Adjusted R Square",xlab="Number of Parameters")

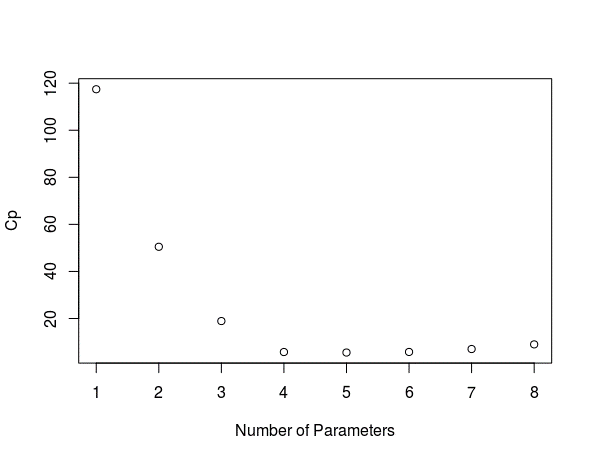
> which.max(rs$adjr2)

[1] 6



The same result, after 4 parameters Rsquare doesn’t increase significantly.

plot(1:8,rs$cp, ylab="Cp",xlab="Number of Parameters")



The same conclusion, the elbow point is for 4 parameters.

h<-lm.influence(model)$hat

> sort(h,decreasing = TRUE)

38 28 43 32 52 23 50

0.40774374 0.32347020 0.31296529 0.29966399 0.28671024 0.26138040 0.26097122

13 42 17 48 18 22 54

0.24810677 0.23927406 0.23800235 0.23281732 0.21675431 0.21170653 0.20463386

5 40 7 25 45 37 20

0.19731884 0.19042916 0.18423215 0.18230663 0.17860268 0.17391749 0.17201034

3 6 34 33 15 35 29

0.16831310 0.16049667 0.16015490 0.15896748 0.15883071 0.14901762 0.14838092

4 16 11 49 8 14 31

0.14419759 0.14220443 0.14067116 0.14003789 0.13907463 0.12692191 0.12462407

19 21 2 27 46 10 47

0.12427787 0.12325226 0.11878224 0.11600466 0.11588024 0.11156504 0.11113955

51 12 26 36 41 9 53

0.11006966 0.10850204 0.10502848 0.10003819 0.09676497 0.09469537 0.09291697

44 30 1 39 24

0.09108298 0.08761814 0.08157404 0.06588190 0.06001475

model1<- **lm**(lnY ~ X1+X2+X3+X8, data = SurgicalUnit\_training)

|  |
| --- |
| > summary(model1)  Call:  lm(formula = lnY ~ X1 + X2 + X3 + X8, data = SurgicalUnit\_training)  Residuals:  Min 1Q Median 3Q Max  -0.45307 -0.16149 -0.02779 0.12073 0.59524  Coefficients:  Estimate Std. Error t value Pr(>|t|)  (Intercept) 3.852419 0.192695 19.992 < 2e-16 \*\*\*  X1 0.073323 0.018973 3.865 0.000327 \*\*\*  X2 0.014185 0.001731 8.196 9.58e-11 \*\*\*  X3 0.015453 0.001396 11.072 6.15e-15 \*\*\*  X8 0.352968 0.077191 4.573 3.29e-05 \*\*\*  ---  Signif. codes: 0 ‘\*\*\*’ 0.001 ‘\*\*’ 0.01 ‘\*’ 0.05 ‘.’ 0.1 ‘ ’ 1  Residual standard error: 0.2109 on 49 degrees of freedom  Multiple R-squared: 0.8299, Adjusted R-squared: 0.816  F-statistic: 59.76 on 4 and 49 DF, p-value: < 2.2e-16 |
| Nothing seems out of ordinary. Lets check the influential points |
| |  | | --- | | > h<-lm.influence(model1)$hat  > sort(h,decreasing = TRUE)  38 28 42 52 32 23 48  0.30591386 0.29137119 0.22622720 0.22207637 0.22024279 0.18851889 0.18401937  5 13 17 43 7 22 18  0.16926387 0.15552396 0.14991900 0.14070242 0.13697626 0.12822176 0.12710129  34 15 54 20 37 50 12  0.11817802 0.11735414 0.10891514 0.10726460 0.10592819 0.09536812 0.08528388  45 31 4 46 51 10 47  0.08314749 0.08086296 0.08031643 0.07561627 0.07316798 0.07270145 0.07217125  6 3 25 36 11 8 14  0.06946491 0.05876010 0.05774766 0.05604245 0.05459422 0.05380426 0.05367197  16 29 19 30 41 9 40  0.05220468 0.04909290 0.04520681 0.04450786 0.03905601 0.03853052 0.03815696  27 1 53 44 21 2 24  0.03648132 0.03519035 0.03516694 0.03471569 0.03284933 0.03145926 0.02838043  39 33 35 49 26  0.02825540 0.02765796 0.02703079 0.02569701 0.02392005 | |

> 2\*5/54

[1] 0.1851852

2\*p/n gives the limit 0.18, however this is not a large dataset and graph above does not indicate influential point. To be on the safest side, I will exclude observation 38 and refit the model.

model2<- **lm**(lnY ~ X1+X2+X3+X8, data = SurgicalUnit\_training[-c(38),])

summary(model2)

plot(model2)

> summary(model2)

Call:

lm(formula = lnY ~ X1 + X2 + X3 + X8, data = SurgicalUnit\_training[-c(38),

])

Residuals:

Min 1Q Median 3Q Max

-0.45562 -0.14867 -0.03083 0.14549 0.64099

Coefficients:

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

(Intercept) 3.895349 0.194167 20.062 < 2e-16 \*\*\*

X1 0.072871 0.018843 3.867 0.000331 \*\*\*

X2 0.012983 0.001951 6.655 2.48e-08 \*\*\*

X3 0.016030 0.001455 11.017 9.68e-15 \*\*\*

X8 0.337933 0.077515 4.360 6.85e-05 \*\*\*

---

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

Residual standard error: 0.2094 on 48 degrees of freedom

Multiple R-squared: 0.8318, Adjusted R-squared: 0.8178

F-statistic: 59.35 on 4 and 48 DF, p-value: < 2.2e-16

h<-lm.influence(model2)$hat

sort(h,decreasing = TRUE)

> sort(h,decreasing = TRUE)

28 52 42 32 23 48 17

0.29504947 0.26139503 0.25958509 0.22028611 0.20307764 0.18903726 0.17810322

5 13 43 22 7 18 15

0.17142044 0.15667209 0.14318610 0.13858520 0.13753055 0.12728017 0.11910492

34 54 20 37 46 50 4

0.11894598 0.11126305 0.10794300 0.10601113 0.09761472 0.09542726 0.08810699

6 12 45 51 31 47 36

0.08712148 0.08544367 0.08323551 0.08228347 0.08092258 0.07495199 0.07388603

10 25 3 11 14 8 16

0.07293899 0.06750293 0.06362588 0.05684742 0.05467303 0.05434010 0.05357164

29 40 30 27 41 19 9

0.04995934 0.04933047 0.04837719 0.04684450 0.04634439 0.04521237 0.04094777

53 21 44 1 2 39 24

0.03963319 0.03934853 0.03867787 0.03718244 0.03228511 0.02998584 0.02958241

33 35 49 26

0.02957034 0.02872186 0.02597808 0.02504815

> summary(influence.measures(model1))

Potentially influential observations of

lm(formula = lnY ~ X1 + X2 + X3 + X8, data = SurgicalUnit\_training) :

dfb.1\_ dfb.X1 dfb.X2 dfb.X3 dfb.X8 dffit cov.r cook.d hat

13 0.06 -0.01 -0.06 -0.06 0.02 -0.09 1.31\_\* 0.00 0.16

17 0.44 -0.15 0.63 -1.14\_\* -0.01 1.42\_\* 0.45\_\* 0.33 0.15

28 -0.20 0.24 0.05 0.07 0.10 0.31 1.53\_\* 0.02 0.29\_\*

38 -0.22 0.02 0.70 -0.42 0.20 -0.86 1.34\_\* 0.15 0.31\_\*

42 -0.04 0.04 0.04 -0.01 -0.05 -0.09 1.43\_\* 0.00 0.23

52 -0.07 0.07 -0.17 0.24 -0.28 -0.39 1.35\_\* 0.03 0.22

Cook's distance and leverage are used to detect highly influential data points, i.e. data points that can have a large effect on the outcome and accuracy of the regression. For large sample sizes, a rough guideline is to consider Cook's distance values above 1 to indicate highly influential points and leverage values greater than 2 times the number of predictors divided by the sample size to indicate high leverage observations. High leverage observations are ones which have predictor values very far from their averages, which can greatly influence the fitted model.