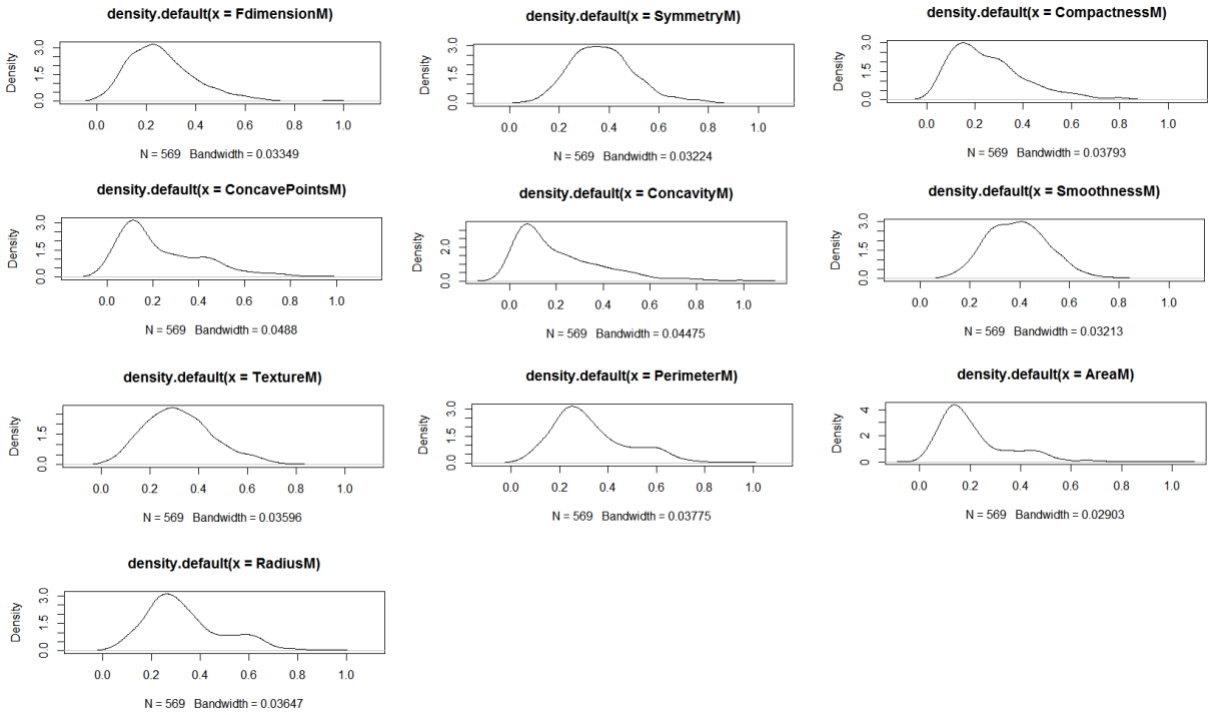
IMSE 514 – MULTIVARIATE STATISTICS
HOMEWORK 3

SURESH OOTY

Data Analysis:

The data was analyzed in detailed in an exploratory approach in comparing the deviation, range of values and distribution.

Normality Assumption:



Density plots of all Mean values

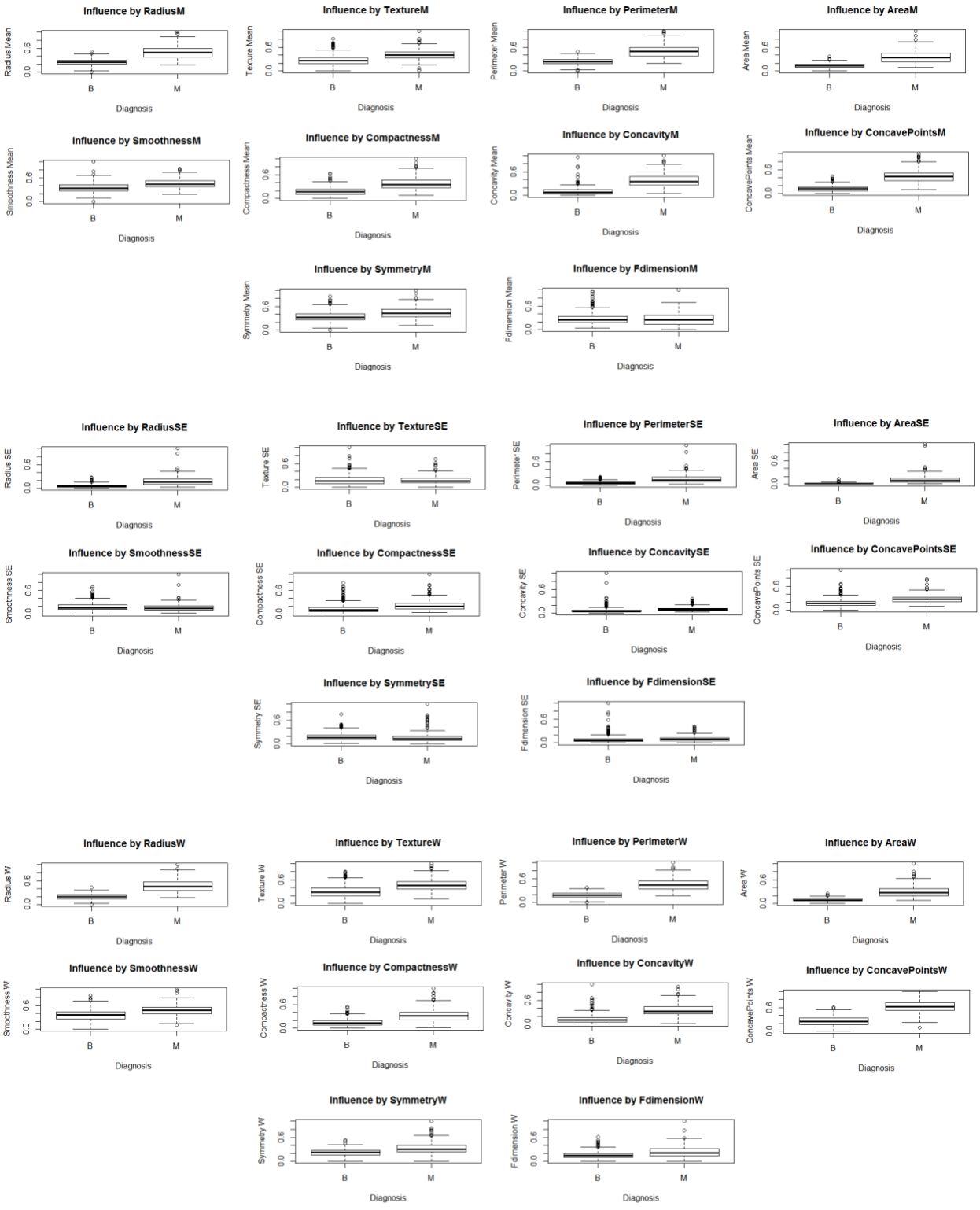
As shown above, all the variables followed a normal distribution.

Data Normalization:

As the data has large variation of ranges among the variables, min-max transformation was applied to normalize the data, to guarantee that none of the variable will have higher influence.

$$x' = \frac{x - \min_x}{max_x - min_x}$$

Variable Selection:



Boxplots of every variable against Diagnosis ('B' or 'M')

Criteria applied for variable Selection: (in a prioritized order)

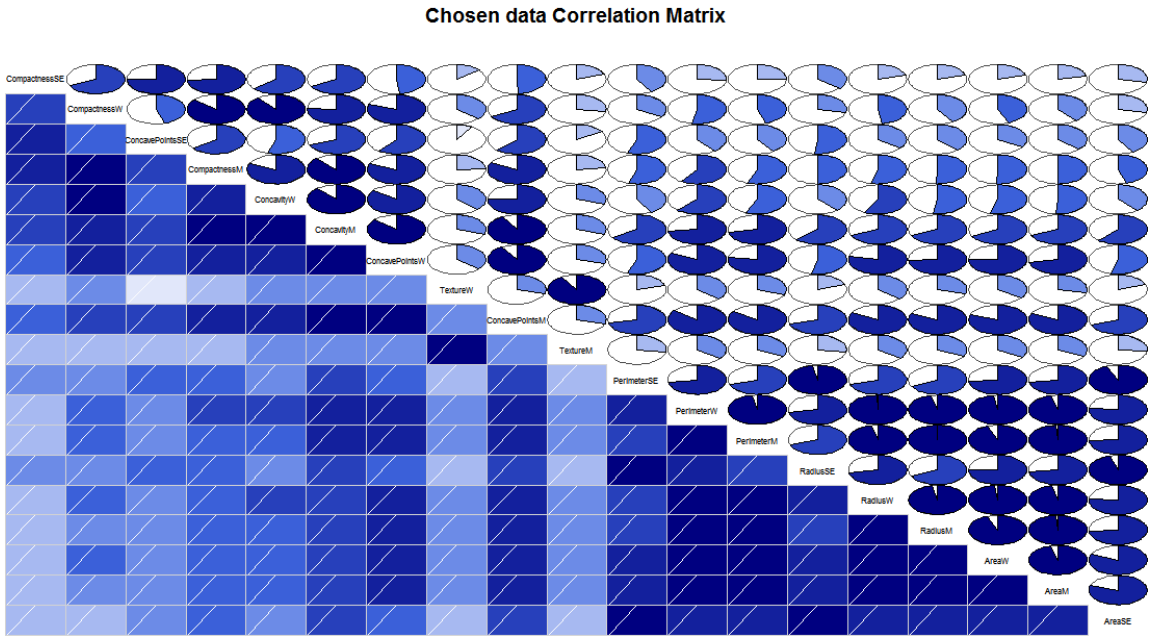
1. Variable's Median (of the 5 point summary) against "M" is not close to that against "B"
2. Variable's Minimum (of the 5 point summary) against "M" is close to or above the median of (5 point summary) against "B".
3. Less or No outliers identified in box plots.

Based on the above criteria, the following variables have been chosen.

- RadiusM
- TextureM
- PerimeterM
- AreaM
- CompactnessM
- ConcavityM
- ConcavepointsM
- RadiusSE
- PerimeterSE
- AreaSE
- CompactnessSE
- ConcavepointsSE
- RadiusW
- TextureW
- PerimeterW
- AreaW
- CompactnessW
- ConcavityW
- ConcavepointsW

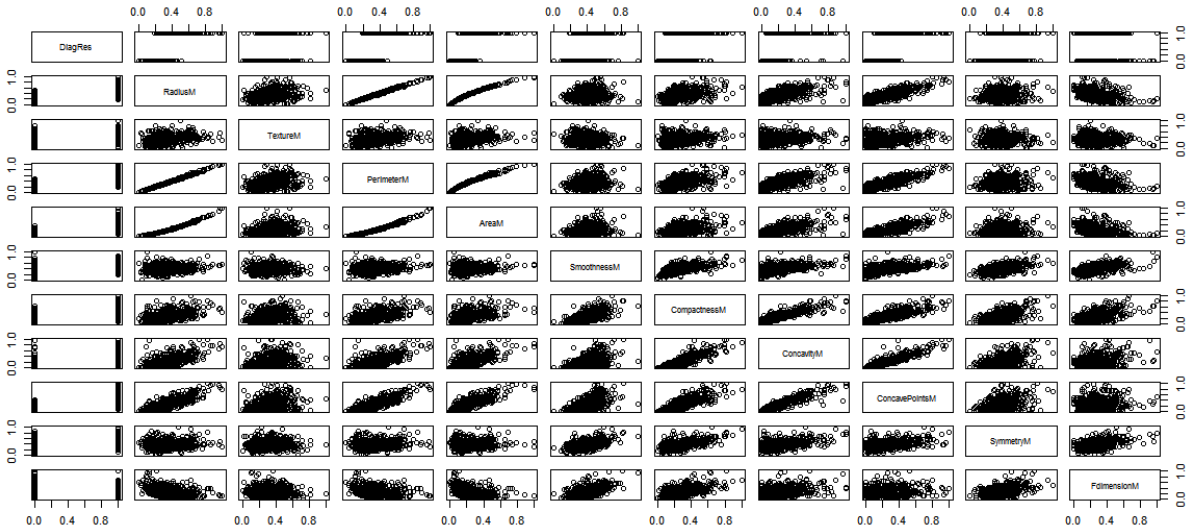
Correlation Matrix:

When the correlation was tried on all the parameters based on the chosen set above. The following was observed.



When the correlation was studied for all the mean values, a natural correlation was observed between the variables RadiusM, PerimeterM and AreaM compared to a slightly not so direct correlation between ConcavityM & ConcavepointsM.

From correlation matrix, a set of reasonable correlations were noted on the response variable ‘Diagnosis’ by the same set of above variables, that were noted for mutual correlation.



```
> cordata<-cbind.data.frame(newset[,c(2,3,4,5,6,7,8,9,10,11)])
> rcorr(as.matrix(cordata),type = "pearson")
```

	RadiusM	TextureM	PerimeterM	AreaM	SmoothnessM	CompactnessM	ConcavityM
RadiusM	1.00	0.32	1.00	0.99	0.17	0.51	0.68
TextureM	0.32	1.00	0.33	0.32	-0.02	0.24	0.30
PerimeterM	1.00	0.33	1.00	0.99	0.21	0.56	0.72
AreaM	0.99	0.32	0.99	1.00	0.18	0.50	0.69
SmoothnessM	0.17	-0.02	0.21	0.18	1.00	0.66	0.52
CompactnessM	0.51	0.24	0.56	0.50	0.66	1.00	0.88
ConcavityM	0.68	0.30	0.72	0.69	0.52	0.88	1.00
ConcavePointsM	0.82	0.29	0.85	0.82	0.55	0.83	0.92
SymmetryM	0.15	0.07	0.18	0.15	0.56	0.60	0.50
FdimensionM	-0.31	-0.08	-0.26	-0.28	0.58	0.57	0.34

	ConcavePointsM	SymmetryM	FdimensionM
RadiusM	0.82	0.15	-0.31
TextureM	0.29	0.07	-0.08
PerimeterM	0.85	0.18	-0.26
AreaM	0.82	0.15	-0.28
SmoothnessM	0.55	0.56	0.58
CompactnessM	0.83	0.60	0.57
ConcavityM	0.92	0.50	0.34
ConcavePointsM	1.00	0.46	0.17
SymmetryM	0.46	1.00	0.48
FdimensionM	0.17	0.48	1.00

Model Construction:

However, the model was constructed based on the chosen data from boxplots show before.

```
> cfull<-glm(Diagnosis~RadiusM+TextureM+PerimeterM+AreaM+CompactnessM+ConcavityM+ConcavePointsM+RadiusSE+PerimeterSE+AreaSE+Compa
ctnessSE+ConcavePointsSE+RadiusW+TextureW+PerimeterW+AreaW+CompactnessW+ConcavityW+ConcavePointsW,family = binomial("logit"),data
= chosedata)
Warning message:
glm.fit: fitted probabilities numerically 0 or 1 occurred
> summary(cfull)

Call:
glm(formula = Diagnosis ~ RadiusM + TextureM + PerimeterM + AreaM +
    CompactnessM + ConcavityM + ConcavePointsM + RadiusSE + PerimeterSE +
    AreaSE + CompactnessSE + ConcavePointsSE + RadiusW + TextureW +
    PerimeterW + AreaW + CompactnessW + ConcavityW + ConcavePointsW,
    family = binomial("logit"), data = chosedata)

Deviance Residuals:
     Min        1Q    Median        3Q       Max
-1.83860  -0.00432  -0.00026   0.00000   2.82138

Coefficients:
               Estimate Std. Error z value Pr(>|z|)
(Intercept)     -21.780      7.554  -2.883  0.00393 **
RadiusM        -155.004    223.399  -0.694  0.48778
TextureM          4.883      9.651   0.506  0.61284
PerimeterM       11.318    228.603   0.050  0.96051
AreaM            87.955    128.224   0.686  0.49275
```

```
CompactnessM     -28.668     18.376  -1.560  0.11874
ConcavityM       -13.850     20.881  -0.663  0.50715
ConcavePointsM    47.536     16.421   2.895  0.00379 **
RadiusSE         -41.102     95.636  -0.430  0.66736
PerimeterSE      -46.241     66.692  -0.693  0.48810
AreaSE           248.968    162.163   1.535  0.12471
CompactnessSE    -31.033     15.506  -2.001  0.04536 *
ConcavePointsSE   26.125     17.373   1.504  0.13264
RadiusW           62.872     92.643   0.679  0.49736
TextureW          13.585      8.543   1.590  0.11177
PerimeterW        40.652     78.074   0.521  0.60259
AreaW             -1.325    127.801  -0.010  0.99173
CompactnessW      36.329     24.236   1.499  0.13389
ConcavityW        17.969     16.138   1.113  0.26553
ConcavePointsW    -4.675     13.017  -0.359  0.71949
```

Between AreaW & PerimeterM , PerimeterM was dropped.

```
Call:
glm(formula = Diagnosis ~ RadiusM + TextureM + AreaM + CompactnessM +
    ConcavityM + ConcavePointsM + RadiusSE + PerimeterSE + AreaSE +
    CompactnessSE + ConcavePointsSE + RadiusW + TextureW + PerimeterW +
    AreaW + CompactnessW + ConcavityW + ConcavePointsW, family = binomial("logit"),
    data = chosedata)

Deviance Residuals:
     Min        1Q    Median        3Q       Max
-1.83537  -0.00439  -0.00026   0.00000   2.83197

Coefficients:
                Estimate Std. Error z value Pr(>|z|)
(Intercept)      -21.731      7.480  -2.905  0.00367 **
RadiusM         -145.331    107.909  -1.347  0.17805
TextureM           5.115      8.447   0.605  0.54485
AreaM             88.876    126.955   0.700  0.48389
CompactnessM     -28.295     16.748  -1.689  0.09113 .
ConcavityM       -13.362     18.383  -0.727  0.46731
ConcavePointsM    47.594     16.371   2.907  0.00365 **
RadiusSE         -43.608     80.999  -0.538  0.59032
PerimeterSE      -45.630     65.428  -0.697  0.48555
AreaSE           252.847    142.062   1.780  0.07510 .
CompactnessSE    -31.382     13.836  -2.268  0.02332 *
ConcavePointsSE   26.259     17.129   1.533  0.12526
RadiusW           62.385     92.018   0.678  0.49779
TextureW          13.436      7.979   1.684  0.09221 .
PerimeterW        41.447     76.338   0.543  0.58717
AreaW             -1.606    127.564  -0.013  0.98996
CompactnessW      36.713     22.964   1.599  0.10988
ConcavityW        17.660     14.842   1.190  0.23410
ConcavePointsW    -4.690     12.992  -0.361  0.71812
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 751.440  on 568  degrees of freedom
```

```
Residual deviance:  49.154  on 550  degrees of freedom
AIC: 87.154

Number of Fisher Scoring iterations: 11
```

AreaW is highly insignificant. Now a careful judgement between AreaM & AreaW is made based on correlation. AreaM is retained.

```
Deviance Residuals:
     Min        1Q    Median        3Q       Max
-1.83412  -0.00440  -0.00026   0.00000   2.83095

Coefficients:
                Estimate Std. Error z value Pr(>|z|)
(Intercept)      -21.747      7.379  -2.947  0.00321 **
RadiusM         -144.466     83.121  -1.738  0.08221 .
TextureM           5.112      8.444   0.605  0.54492
AreaM             87.820     95.228   0.922  0.35642
CompactnessM     -28.288     16.735  -1.690  0.09095 .
ConcavityM       -13.366     18.371  -0.728  0.46691
ConcavePointsM    47.593     16.351   2.911  0.00361 **
RadiusSE         -43.340     78.043  -0.555  0.57867
PerimeterSE      -45.404     62.881  -0.722  0.47025
AreaSE           251.753    112.190   2.244  0.02483 *
CompactnessSE    -31.369     13.793  -2.274  0.02295 *
ConcavePointsSE   26.266     17.090   1.537  0.12430
RadiusW           61.625     69.395   0.888  0.37452
TextureW          13.432      7.973   1.685  0.09206 .
PerimeterW        41.176     73.210   0.562  0.57382
CompactnessW      36.725     22.935   1.601  0.10932
ConcavityW        17.644     14.787   1.193  0.23279
ConcavePointsW    -4.699     12.959  -0.363  0.71689
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 751.440  on 568  degrees of freedom
Residual deviance:  49.154  on 551  degrees of freedom
AIC: 85.154

Number of Fisher Scoring iterations: 11
```

Dropping ConcavePointsW – as ConcavePointsM is already significant.

```
Deviance Residuals:
     Min        1Q    Median        3Q       Max
-1.77168  -0.00413  -0.00025   0.00000   2.83547

Coefficients:
                Estimate Std. Error z value Pr(>|z|)
(Intercept)      -22.125      7.474  -2.960  0.00307 **
RadiusM         -145.343     83.740  -1.736  0.08263 .
TextureM           6.202      7.938   0.781  0.43459
AreaM             92.219     95.171   0.969  0.33255
CompactnessM     -27.977     16.791  -1.666  0.09567 .
ConcavityM       -12.829     18.354  -0.699  0.48455
ConcavePointsM    45.824     15.969   2.870  0.00411 **
```

```
RadiusSE         -30.854     69.851  -0.442  0.65870
PerimeterSE      -49.651     63.535  -0.781  0.43452
AreaSE           238.892    104.205   2.293  0.02188 *
CompactnessSE    -28.595     11.215  -2.550  0.01078 *
ConcavePointsSE   21.979     11.651   1.886  0.05924 .
RadiusW           54.442     66.758   0.816  0.41478
TextureW          12.758      7.746   1.647  0.09953 .
PerimeterW        46.320     73.381   0.631  0.52790
CompactnessW      32.921     20.004   1.646  0.09982 .
ConcavityW        17.127     14.684   1.166  0.24346
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 751.440  on 568  degrees of freedom
Residual deviance:  49.285  on 552  degrees of freedom
AIC: 83.285

Number of Fisher Scoring iterations: 11
```

RadiusSE is dropped as it is correlated with AreaSE , which turns out to be significant.

```
Deviance Residuals:
     Min        1Q    Median        3Q       Max
-1.86515  -0.00393  -0.00020   0.00000   2.76293

Coefficients:
               Estimate Std. Error z value Pr(>|z|)
(Intercept)     -23.257      7.099  -3.276  0.00105 **
RadiusM         -144.543     83.857  -1.724  0.08476 .
TextureM           5.867      7.880   0.745  0.45653
AreaM             91.493     95.279   0.960  0.33692
CompactnessM     -25.279     15.205  -1.663  0.09640 .
ConcavityM       -18.054     14.096  -1.281  0.20026
ConcavePointsM    47.619     15.287   3.115  0.00184 **
PerimeterSE      -71.141     42.256  -1.684  0.09226 .
AreaSE           214.735     86.695   2.477  0.01325 *
CompactnessSE    -26.675      9.935  -2.685  0.00725 **
ConcavePointsSE   20.720     10.967   1.889  0.05885 .
RadiusW           33.257     46.158   0.720  0.47122
TextureW          13.623      7.517   1.812  0.06995 .
PerimeterW        72.540     44.657   1.624  0.10429
CompactnessW      27.747     15.600   1.779  0.07530 .
ConcavityW        21.075     11.902   1.771  0.07662 .
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 751.440  on 568  degrees of freedom
Residual deviance:  49.483  on 553  degrees of freedom
AIC: 81.483

Number of Fisher Scoring iterations: 11
```

Dropping RadiusW, in comparison with RadiusM & AreaM

```
Deviance Residuals:
     Min        1Q    Median        3Q       Max
-1.90704  -0.00435  -0.00020   0.00000   2.83363

Coefficients:
                Estimate Std. Error z value Pr(>|z|)
(Intercept)      -22.584      7.110  -3.176 0.001492 **
RadiusM         -131.309     82.929  -1.583 0.113334
TextureM           4.643      7.532   0.616 0.537610
AreaM             81.158     95.649   0.848 0.396160
CompactnessM     -25.723     15.144  -1.699 0.089403 .
ConcavityM       -17.677     14.528  -1.217 0.223683
ConcavePointsM    46.369     15.020   3.087 0.002020 **
PerimeterSE      -91.366     33.140  -2.757 0.005833 **
AreaSE           251.964     73.104   3.447 0.000568 ***
CompactnessSE    -28.072      9.835  -2.854 0.004312 **
ConcavePointsSE   21.389     11.038   1.938 0.052663 .
TextureW          14.179      7.355   1.928 0.053894 .
PerimeterW        98.040     30.858   3.177 0.001488 **
CompactnessW      28.448     15.653   1.817 0.069148 .
ConcavityW        20.767     12.259   1.694 0.090254 .
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 751.44  on 568  degrees of freedom
Residual deviance:  49.99  on 554  degrees of freedom
AIC: 79.99

Number of Fisher Scoring iterations: 11
```

TextureM is chosen to be dropped, as the TextureW is almost significant. And from the correlation matrix, TextureW will be completely independent, when TextureM is dropped.

```
Deviance Residuals:
     Min        1Q    Median        3Q       Max
-1.85506  -0.00459  -0.00018   0.00000   2.77434

Coefficients:
                Estimate Std. Error z value Pr(>|z|)
(Intercept)      -23.021      7.000  -3.289 0.001007 **
RadiusM         -128.756     83.112  -1.549 0.121340
AreaM             83.229     95.631   0.870 0.384132
CompactnessM     -24.102     14.794  -1.629 0.103279
ConcavityM       -18.369     14.497  -1.267 0.205134
ConcavePointsM    44.196     14.420   3.065 0.002177 **
PerimeterSE      -84.502     29.901  -2.826 0.004712 **
AreaSE           241.630     68.909   3.507 0.000454 ***
CompactnessSE    -29.295     10.057  -2.913 0.003580 **
ConcavePointsSE   23.646     10.654   2.220 0.026450 *
TextureW          17.922      4.619   3.880 0.000105 ***
PerimeterW        94.321     29.568   3.190 0.001423 **
CompactnessW      29.285     15.775   1.856 0.063402 .
ConcavityW        20.234     12.407   1.631 0.102931
---
```

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

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 751.440  on 568  degrees of freedom
Residual deviance:  50.383  on 555  degrees of freedom
AIC: 78.383

Number of Fisher Scoring iterations: 11
```

As expected. Now the TextureW has become highly significant. Now chosing AreaM to be dropped, in comparison with the other correlated variable PerimeterW.

```
Deviance Residuals:
     Min        1Q    Median        3Q       Max
-2.01291  -0.00467  -0.00015   0.00000   2.90912

Coefficients:
                Estimate Std. Error z value Pr(>|z|)
(Intercept)      -26.730      5.959  -4.485 7.28e-06 ***
RadiusM          -61.638     21.953  -2.808 0.004989 **
CompactnessM     -27.623     14.391  -1.919 0.054929 .
ConcavityM       -12.705     12.874  -0.987 0.323702
ConcavePointsM    41.831     13.689   3.056 0.002244 **
PerimeterSE      -83.794     28.765  -2.913 0.003579 **
AreaSE           238.810     65.244   3.660 0.000252 ***
CompactnessSE    -26.884      8.983  -2.993 0.002763 **
ConcavePointsSE   21.780      9.999   2.178 0.029383 *
TextureW          16.821      4.031   4.173 3.01e-05 ***
PerimeterW        92.979     28.035   3.317 0.000911 ***
CompactnessW      28.062     15.612   1.798 0.072256 .
ConcavityW        17.793     12.498   1.424 0.154559
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 751.440  on 568  degrees of freedom
Residual deviance:  51.168  on 556  degrees of freedom
AIC: 77.168

Number of Fisher Scoring iterations: 11
```

Dropping ConcavityM , in comparison with a related variable CompactnessM. This will make CompactnessM independent.

```
Deviance Residuals:
     Min        1Q    Median        3Q       Max
-2.19813  -0.00685  -0.00026   0.00000   2.94329

Coefficients:
                Estimate Std. Error z value Pr(>|z|)
(Intercept)      -24.985      5.395  -4.631 3.63e-06 ***
RadiusM          -55.960     20.608  -2.715 0.006618 **
CompactnessM     -31.176     14.175  -2.199 0.027849 *
ConcavePointsM    36.909     12.292   3.003 0.002677 **
```

```
PerimeterSE      -75.626     26.916  -2.810 0.004959 **
AreaSE           222.040     61.276   3.624 0.000291 ***
CompactnessSE    -27.482      8.749  -3.141 0.001683 **
ConcavePointsSE   18.207      8.801   2.069 0.038573 *
TextureW          16.097      3.849   4.182 2.89e-05 ***
PerimeterW        85.893     26.404   3.253 0.001142 **
CompactnessW      34.938     14.046   2.487 0.012868 *
ConcavityW         7.270      6.601   1.101 0.270795
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 751.44  on 568  degrees of freedom
Residual deviance:  52.22  on 557  degrees of freedom
AIC: 76.22

Number of Fisher Scoring iterations: 11
```

Dropping ConcavityW , as it is the only insignificant variable.

```
Deviance Residuals:
     Min        1Q    Median        3Q       Max
-2.01524  -0.00693  -0.00029   0.00000   2.82694

Coefficients:
                Estimate Std. Error z value Pr(>|z|)
(Intercept)      -24.743      5.224  -4.737 2.17e-06 ***
RadiusM          -59.210     19.760  -2.996 0.002731 **
CompactnessM     -33.132     13.261  -2.498 0.012475 *
ConcavePointsM    37.217     11.379   3.271 0.001073 **
PerimeterSE      -78.288     26.647  -2.938 0.003304 **
AreaSE           220.689     59.976   3.680 0.000234 ***
CompactnessSE    -27.698      8.702  -3.183 0.001459 **
ConcavePointsSE   21.968      8.324   2.639 0.008308 **
TextureW          16.238      3.759   4.320 1.56e-05 ***
PerimeterW        89.025     25.454   3.497 0.000470 ***
CompactnessW      41.103     12.740   3.226 0.001254 **
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 751.440  on 568  degrees of freedom
Residual deviance:  53.467  on 558  degrees of freedom
AIC: 75.467

Number of Fisher Scoring iterations: 11
```

All variables are shown significant. But the model still continues to have the error. Among all Compactness parameters. Only CompactnessSE was dropped.

```
Deviance Residuals:
     Min        1Q    Median        3Q       Max
-2.22306  -0.03408  -0.00413   0.00003   2.60428
```

```
Coefficients:
                Estimate Std. Error z value Pr(>|z|)
(Intercept)      -15.650      2.814  -5.562 2.67e-08 ***
RadiusM          -46.240     15.327  -3.017 0.002553 **
CompactnessM     -34.322     10.347  -3.317 0.000910 ***
ConcavePointsM    41.975      9.121   4.602 4.19e-06 ***
PerimeterSE      -44.284     19.551  -2.265 0.023509 *
AreaSE           138.270     41.799   3.308 0.000940 ***
ConcavePointsSE   -2.214      5.192  -0.426 0.669815
TextureW          12.039      2.725   4.418 9.94e-06 ***
PerimeterW        67.569     19.615   3.445 0.000572 ***
CompactnessW      19.879      8.148   2.440 0.014698 *
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 751.440  on 568  degrees of freedom
Residual deviance:  69.972  on 559  degrees of freedom
AIC: 89.972

Number of Fisher Scoring iterations: 10
```

Dropping ConcavePointsSE. Not correlated.

```
Deviance Residuals:
     Min        1Q    Median        3Q       Max
-2.17452  -0.03439  -0.00393   0.00003   2.61650

Coefficients:
               Estimate Std. Error z value Pr(>|z|)
(Intercept)     -15.836      2.801  -5.654 1.57e-08 ***
RadiusM         -46.847     15.210  -3.080 0.002070 **
CompactnessM    -35.043     10.302  -3.402 0.000670 ***
ConcavePointsM   41.151      8.894   4.627 3.71e-06 ***
PerimeterSE     -45.558     19.303  -2.360 0.018267 *
AreaSE          136.716     41.157   3.322 0.000894 ***
TextureW         12.005      2.697   4.452 8.52e-06 ***
PerimeterW       69.043     19.324   3.573 0.000353 ***
CompactnessW     19.934      8.124   2.454 0.014144 *
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 751.440  on 568  degrees of freedom
Residual deviance:  70.155  on 560  degrees of freedom
AIC: 88.155

Number of Fisher Scoring iterations: 10
```

The error continues.

Tried to remove the masking effect by the similar variables. When CompactnessW was removed, it unmasked PerimeterSE. When CompactnessM was removed, it unmasked CompactnessW. So, going ahead with removing CompactnessM.

```
Deviance Residuals:
     Min        1Q    Median        3Q       Max
```

```
-2.39788  -0.06463  -0.01205   0.00036   3.01770

Coefficients:
               Estimate Std. Error z value Pr(>|z|)
(Intercept)     -15.484      2.514  -6.159 7.33e-10 ***
RadiusM         -41.331     13.275  -3.113  0.00185 **
ConcavePointsM   16.913      4.281   3.951 7.78e-05 ***
PerimeterSE     -37.416     17.088  -2.190  0.02856 *
AreaSE           96.749     36.508   2.650  0.00805 **
TextureW         11.145      2.358   4.726 2.29e-06 ***
PerimeterW       71.504     17.488   4.089 4.34e-05 ***
CompactnessW     -1.584      3.429  -0.462  0.64417
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 751.440  on 568  degrees of freedom
Residual deviance:  88.346  on 561  degrees of freedom
AIC: 104.35

Number of Fisher Scoring iterations: 10
```

Dropping CompactnessW , as it is insignificant.

```
Deviance Residuals:
     Min        1Q    Median        3Q       Max
-2.51063  -0.06625  -0.01232   0.00042   3.04392

Coefficients:
               Estimate Std. Error z value Pr(>|z|)
(Intercept)     -15.563      2.499  -6.227 4.74e-10 ***
RadiusM         -37.776     10.515  -3.593 0.000327 ***
ConcavePointsM   15.871      3.535   4.490 7.12e-06 ***
PerimeterSE     -38.635     16.961  -2.278 0.022731 *
AreaSE          101.177     35.733   2.832 0.004633 **
TextureW         11.030      2.331   4.731 2.23e-06 ***
PerimeterW       66.731     13.579   4.914 8.90e-07 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 751.44  on 568  degrees of freedom
Residual deviance:  88.55  on 562  degrees of freedom
AIC: 102.55

Number of Fisher Scoring iterations: 10
```

The warning message still exists. Now trying similar approach with Perimeter values. When PerimeterW was removed, it exposed PerimeterSE. But otherwise not. However, Perimeter is highly correlated with RadiusM , so hence removed both perimeter values.

```
Deviance Residuals:
     Min        1Q    Median        3Q       Max
-2.59524  -0.11643  -0.02817   0.01164   2.72050

Coefficients:
               Estimate Std. Error z value Pr(>|z|)
```

```
(Intercept)     -15.408      1.810  -8.513  < 2e-16 ***
RadiusM          13.078      3.182   4.111 3.95e-05 ***
ConcavePointsM   19.929      2.866   6.954 3.56e-12 ***
AreaSE           21.833      9.134   2.390   0.0168 *
TextureW         11.774      1.855   6.346 2.21e-10 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 751.44  on 568  degrees of freedom
Residual deviance: 132.26  on 564  degrees of freedom
AIC: 142.26

Number of Fisher Scoring iterations: 8
```

The warning message remains, telling that the model is not converging. But trying to replace the SE values with M values , if it is making the model converge.
After few tries, the model did converge with few changes in parameters to M and then dropping redundant RadiusM , as it highly correlated with AreaM.

```
Deviance Residuals:
     Min        1Q    Median        3Q       Max
-2.27200  -0.15271  -0.04820   0.02016   2.80715

Coefficients:
              Estimate Std. Error z value Pr(>|z|)
(Intercept)     -12.472      1.356  -9.196  < 2e-16 ***
ConcavePointsM   20.443      2.641   7.740 9.91e-15 ***
AreaM            18.333      3.421   5.359 8.37e-08 ***
TextureM          9.624      1.646   5.847 4.99e-09 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 751.44  on 568  degrees of freedom
Residual deviance: 161.70  on 565  degrees of freedom
AIC: 169.7

Number of Fisher Scoring iterations: 8
```

Comparison of Models with these 3 variables in M, SE & W.
SE:

```
Deviance Residuals:
    Min       1Q   Median       3Q      Max
-3.3338  -0.4702  -0.2726   0.0548   2.6441

Coefficients:
               Estimate Std. Error z value Pr(>|z|)
(Intercept)     -3.4719     0.3868  -8.976  < 2e-16 ***
ConcavePointsSE  2.3707     1.4794   1.603    0.109
AreaSE          75.1988     7.5590   9.948  < 2e-16 ***
TextureSE       -6.7218     1.5751  -4.268 1.98e-05 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

```
(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 751.44  on 568  degrees of freedom
Residual deviance: 336.52  on 565  degrees of freedom
AIC: 344.52

Number of Fisher Scoring iterations: 7
```

High AIC values.

W:

```
Deviance Residuals:
    Min       1Q   Median       3Q      Max
-1.9881  -0.0687  -0.0089   0.0021   3.8495

Coefficients:
               Estimate Std. Error z value Pr(>|z|)
(Intercept)     -18.623      2.467  -7.549 4.39e-14 ***
ConcavePointsW   15.767      2.756   5.721 1.06e-08 ***
AreaW            47.644      7.665   6.216 5.10e-10 ***
TextureW         10.333      2.008   5.147 2.65e-07 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 751.440  on 568  degrees of freedom
Residual deviance:  97.987  on 565  degrees of freedom
AIC: 105.99

Number of Fisher Scoring iterations: 9
```

AIC is good and all the variables are significant, but with a warning message. So, when a ANOVA test run against Model_M and Model_W .

```
> anova(cfull,cfullw,"chisq")
Analysis of Deviance Table

Model 1: Diagnosis ~ ConcavePointsM + AreaM + TextureM
Model 2: Diagnosis ~ ConcavePointsW + AreaW + TextureW
  Resid. Df Resid. Dev Df Deviance
1       565    161.696
2       565     97.987  0   63.709
```

Telling us that the models are different.

As the AIC for the final model was little high, a trial & error method was applied to see if by bringing any new variable to model the AIC value improves.
When PerimeterW was introduced, the model came out little improved as follows.

```
Deviance Residuals:
    Min       1Q   Median       3Q      Max
-2.1076  -0.0901  -0.0165   0.0056   3.3372
```

```
Coefficients:
               Estimate Std. Error z value Pr(>|z|)
(Intercept)     -17.236      2.120  -8.131 4.27e-16 ***
ConcavePointsM   15.977      3.228   4.950 7.42e-07 ***
AreaM           -28.016      8.690  -3.224  0.00126 **
TextureM         10.328      2.028   5.093 3.53e-07 ***
PerimeterW       54.143     10.041   5.392 6.95e-08 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 751.44  on 568  degrees of freedom
Residual deviance: 113.46  on 564  degrees of freedom
AIC: 123.46

Number of Fisher Scoring iterations: 9
```

Comparing this with the old model (without PerimeterW)

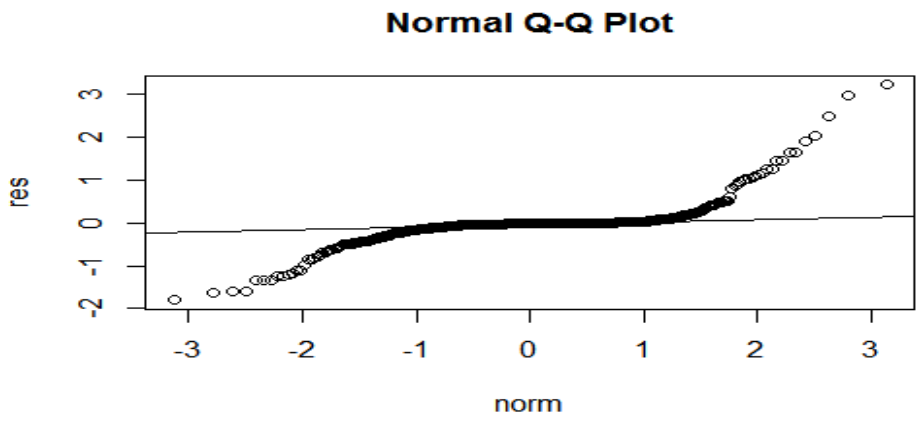
```
> anova(cfull,cfullold,"chi")
Analysis of Deviance Table

Model 1: Diagnosis ~ ConcavePointsM + AreaM + TextureM + PerimeterW
Model 2: Diagnosis ~ ConcavePointsM + AreaM + TextureM
  Resid. Df Resid. Dev Df Deviance
1       564     113.46
2       565     161.70 -1  -48.235
```

This tells us that the model with PerimeterW could be better compared with earlier one.

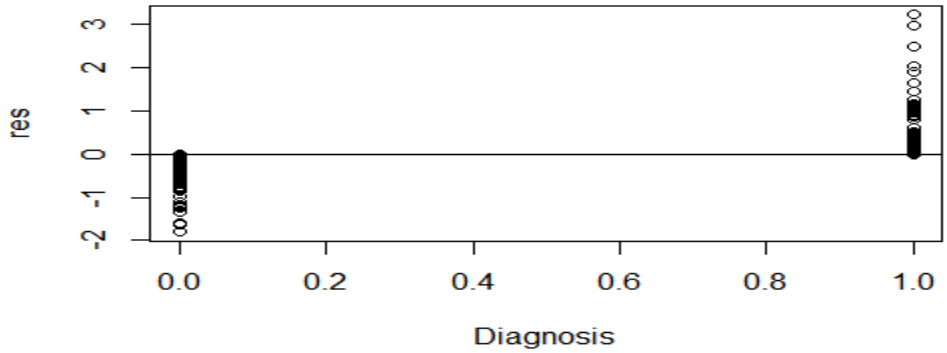
This somewhat matches our initial study with the boxplots.

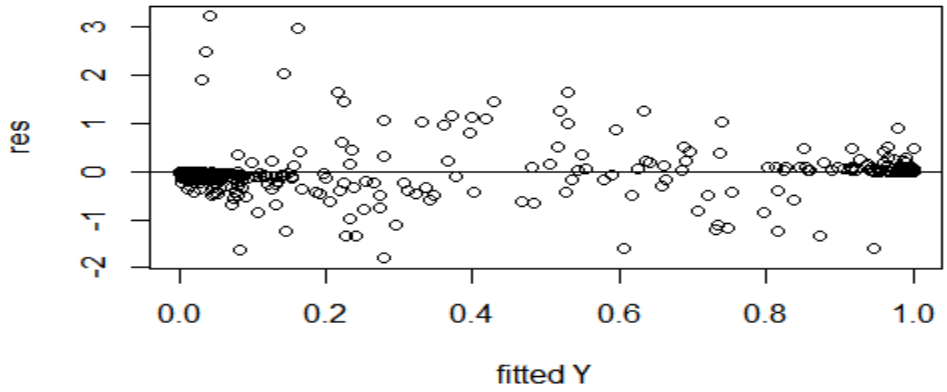
Checking the model with residual plots

```
residuals<-resid(cfull)
qqnorm(residuals,ylab="res",xlab="norm")
qqline(residuals)
plot(DiagB,residuals,ylab="res",xlab="Diagnosis")
abline(0,0)
fittedY<-fitted.values(stpfwd)
plot(fittedY,residuals,ylab="res",xlab="fitted Y")
abline(0,0)
```

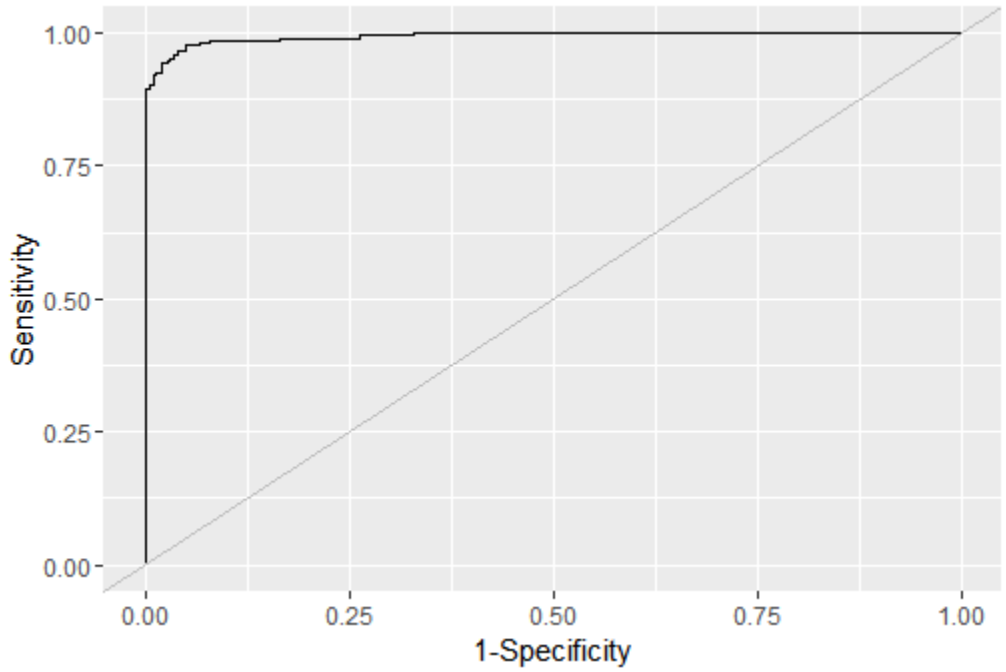


The normal plot show the bimodal nature.

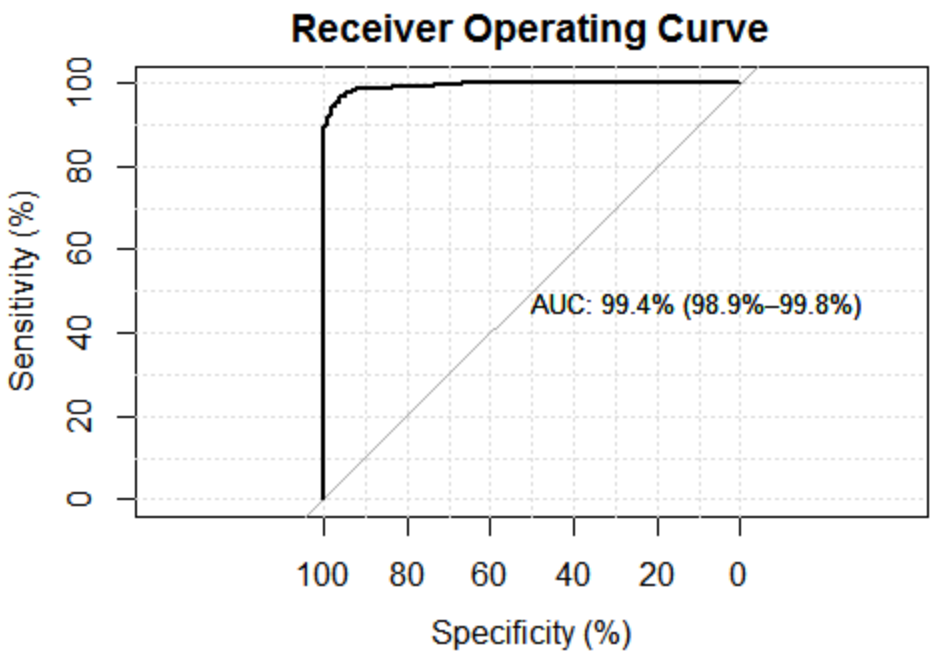




```
library("ROCR", lib.loc="~/R/win-library/3.3")
predicted<-predict(stpfwd)
prob<-prediction(predicted,hw3$DiagB)
tprfpr<-performance(prob,"tpr","fpr")
tpr<-unlist(slot(tprfpr,"y.values"))
fpr<-unlist(slot(tprfpr,"x.values"))
roc<-data.frame(tpr,fpr)
library("ggplot2", lib.loc="~/R/win-library/3.3")
ggplot(roc)+geom_line(aes(x=fpr,y=tpr))geom_abline(inter)
ggplot(roc)+geom_line(aes(x=fpr,y=tpr))geom_abline(intercept=0,slope=1,colour = "gray")+ylab("Sensitivity")+xlab("1-Specificity")
ggplot(roc)+geom_line(aes(x=fpr,y=tpr))+geom_abline(intercept=0,slope=1,colour = "gray")+ylab("Sensitivity")+xlab("1-
Specificity")
```



When plotted using Hosmer_Lemeshow method.



The model seems to be good with AIC 123.46

Diagnosis (Y) = -17.236 + 15.977 ConcavePointsM – 28.016 AreaM + 10.328 TextureM + 54.143 PerimeterW