STAT 835 Final Project

Code

J Kim

November 26, 2021

1 Introduction to the topic and motivation

Breast cancer is one of the common cancer among women and 1 in 8 women will be diagnosed with invasive breast cancer over the course of her lifetime. This means the average woman's breast cancer risk is 12~13%. Being a woman and aging are the two bigger risk factors, but there are many other things that can increase or decrease a person's breast cancer risk.

Here, a dataset of breast cancer are obtained from UCL ML repository (https://archive.ics.uci.edu/ml/datasets/Breast+Cancer+Wisconsin+%28Diagnostic%29). Breast cancer occurs as a result of abnormal growth of cells in the breast tissues which is called tumor. Tumors can develop benign or malignant (becomes breast cancer). MRI, mammogram, ultrasound and biopsy are utilized to diagnose breast cancer. Therefore, making use of this dataset would be beneficial for researchers or medical professionals go about breast cancers.

Dataset information.

1. ID number
2. Diagnosis - M = malignant, B = benign. This will be changed - 1 = malignant, 0 = benign.

Ten real-valued features are computed for each cell nucleus:

```
1. radius (mean of distances from center to points on the perimeter)
2. texture (standard deviation of gray-scale values)
3. perimeter
4. area
5. smoothness (local variation in radius lengths)
6. compactness (perimeter^2 / area - 1.0)
7. concavity (severity of concave portions of the contour)
8. symmetry
9. fractal dimension ("coastline approximation" - 1)
10. age
```

2 Method

Logistic regression is used to predict and describe association between explanatory variables and the response variable. Split the dataset into a test dataset and a traning dataset. Explanatory data analysis including visualization is performed. Initial model will be the following model and then will be modified.

Y_i Diagnosis
X_1 radius
X_2 texture
X_4 perimeter
X_5 area
X_6 smoothness
X_7 compactness
X_8 concavity
X_9 symmetry
X_10 fractal
X_11 age

$$logit(\hat{\pi_i}) = \hat{\beta_0} + \hat{\beta_1}X_1 + \hat{\beta_2}X_2 + \hat{\beta_3}X_3 + \hat{\beta_4}X_4 + \hat{\beta_5}X_5 + \hat{\beta_6}X_6 + \hat{\beta_7}X_7 + \hat{\beta_8}X_8 + \hat{\beta_9}X_9 + \hat{\beta_{10}}X_{10} + \hat{\beta_{11}}X_{11}$$

2.1 Explanatory data analysis

Exploring predictor variables and the response variable using plots and tables to discover any association between the predictor variables and the response variables. Even any relationship between the predictor variables is also important to the model.

2.1.1 Summary statistics

Summary Statistics

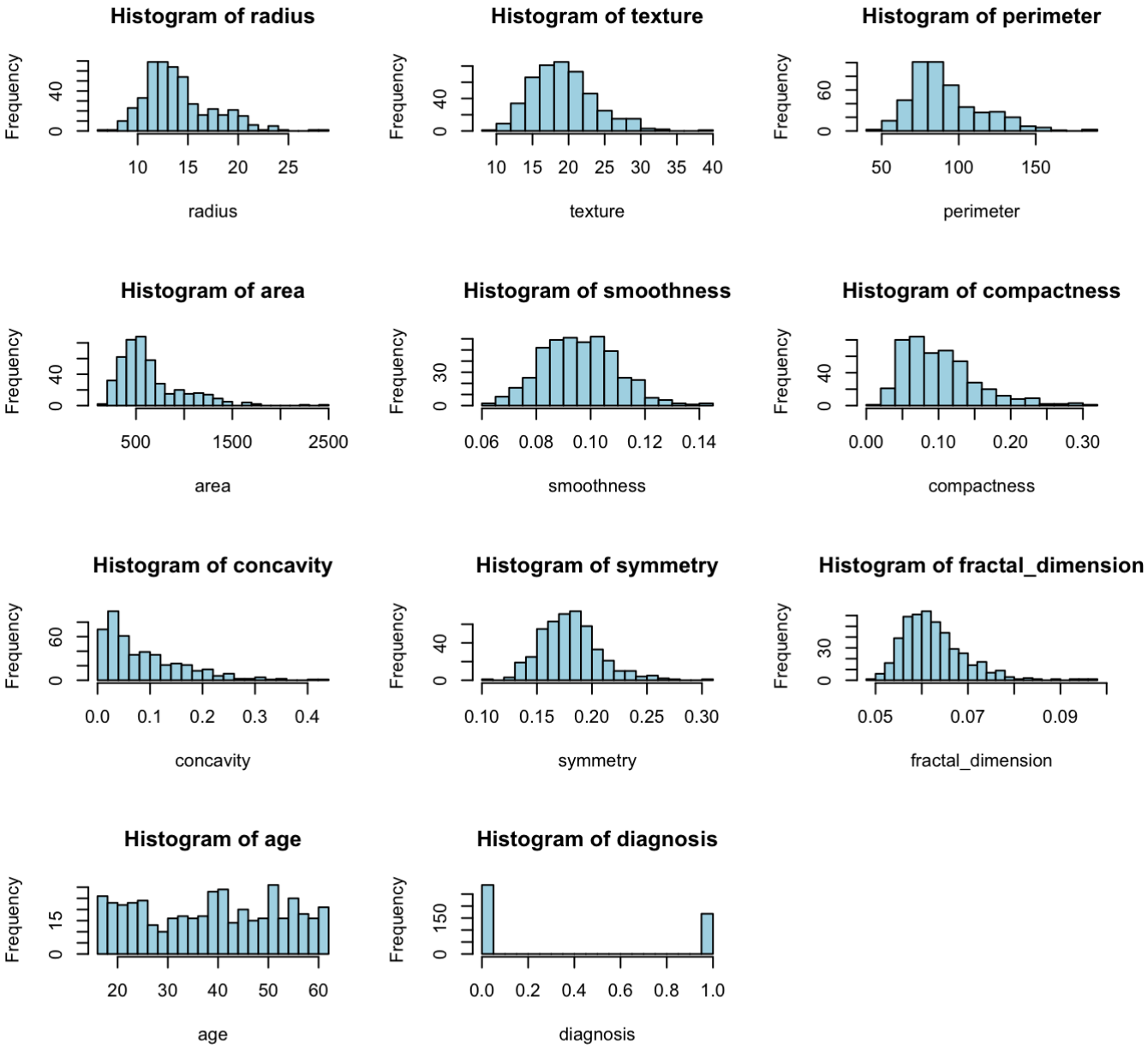
x
No. of observations = 456

Var. name	obs.	mean	median	s.d.	min.	max.
radius	456	14.04	13.39	3.35	6.98	28.11
texture	456	19.29	18.76	4.41	9.71	39.28
perimeter	456	91.29	86.29	23	43.79	188.5
area	456	643.12	552.95	329.91	143.5	2499
smoothness	456	0.1	0.1	0.01	0.06	0.14
compactness	456	0.1	0.09	0.05	0.02	0.31
concavity	456	0.09	0.06	0.08	0	0.43
symmetry	456	0.18	0.18	0.03	0.11	0.3
fractal_dimension	456	0.06	0.06	0.01	0.05	0.1
age	456	39.27	40	13.69	16	62
diagnosis	456	0.37	0	0.48	0	1

2.1.2 Histogram of explanatory variables

Some of the explanatory variables look right-tailed distribution - Symmetry variable seems close to normal distribution.

Code


Histogram of radius
Frequency
0 40
10 15 20 25
radius
Histogram of texture
Frequency
0 40
10 15 20 25 30 35 40
texture
Histogram of perimeter
Frequency
0 60
50 100 150
perimeter
Histogram of area
Frequency
0 40
500 1500 2500
area
Histogram of smoothness
Frequency
0 30
0.06 0.08 0.10 0.12 0.14
smoothness
Histogram of compactness
Frequency
0 40
0.00 0.10 0.20 0.30
compactness
Histogram of concavity
Frequency
0 60
0.0 0.1 0.2 0.3 0.4
concavity
Histogram of symmetry
Frequency
0 40
0.10 0.15 0.20 0.25 0.30
symmetry
Histogram of fractal_dimension
Frequency
0 30
0.05 0.07 0.09
fractal_dimension
Histogram of age
Frequency
0 15
20 30 40 50 60
age
Histogram of diagnosis
Frequency
0 150
0.0 0.2 0.4 0.6 0.8 1.0
diagnosis

2.1.2.1 Frequency of the response variable

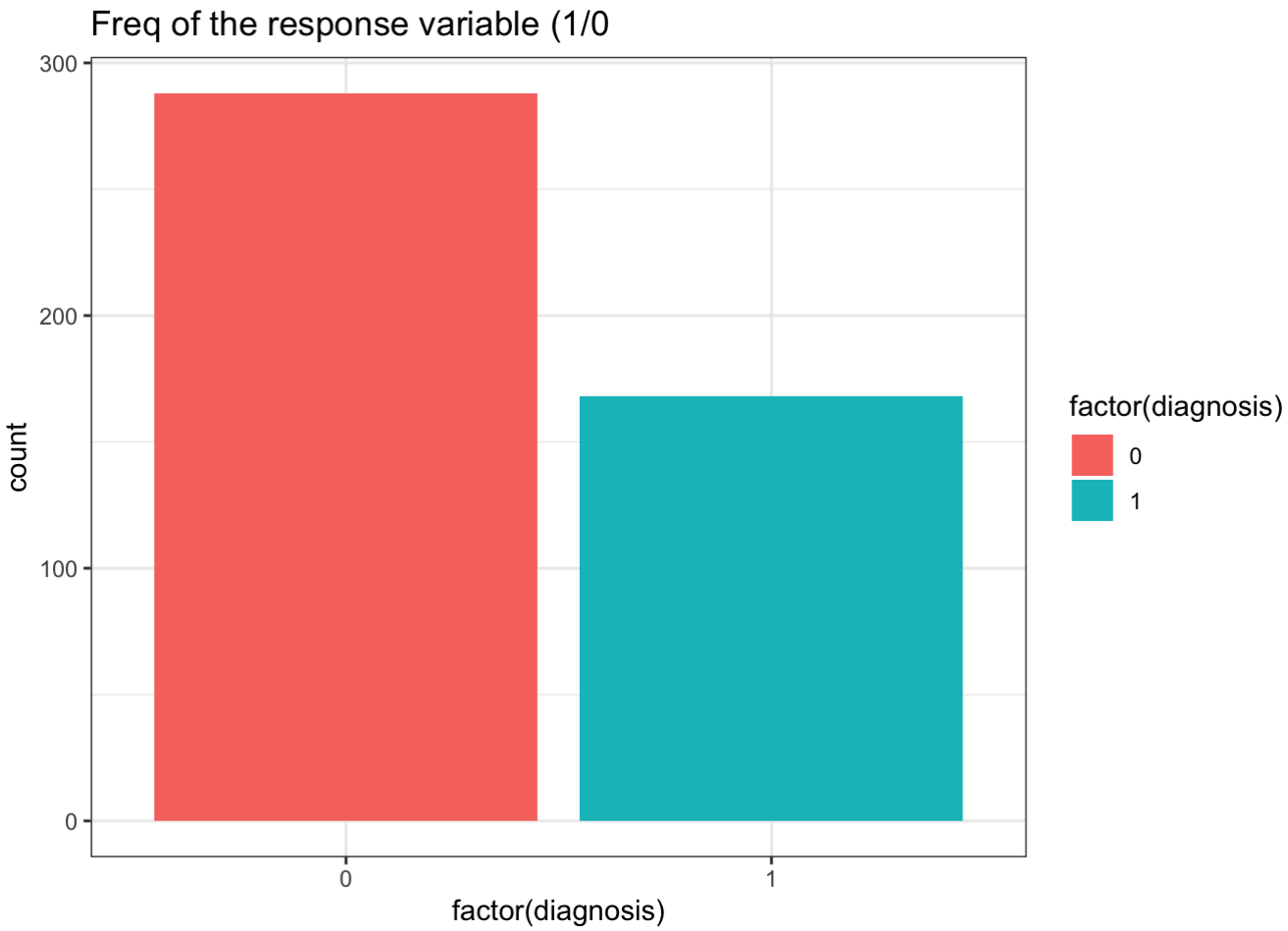
There doesn't seem to be a unbalanced issue in the response variable.

Code

Frequency of the response variable (1/0

Var1	Freq
0	357
1	212

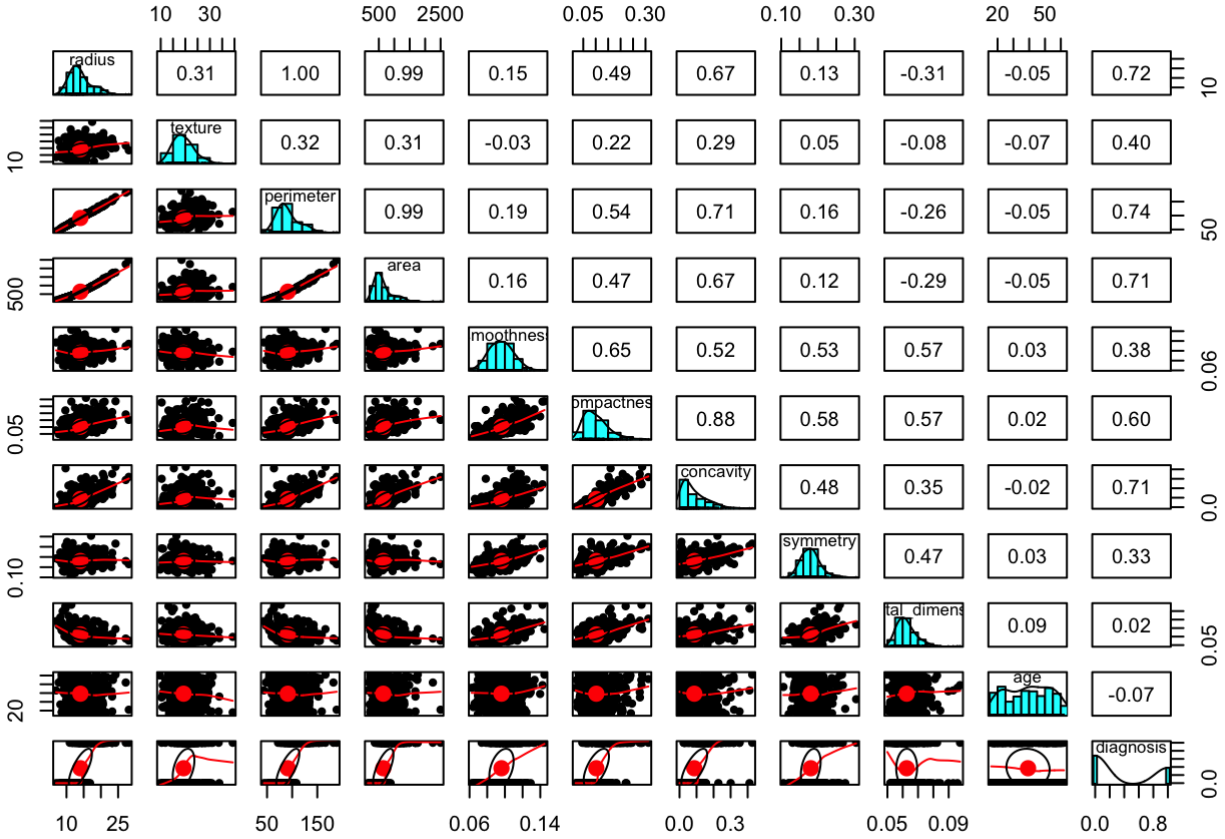
Code



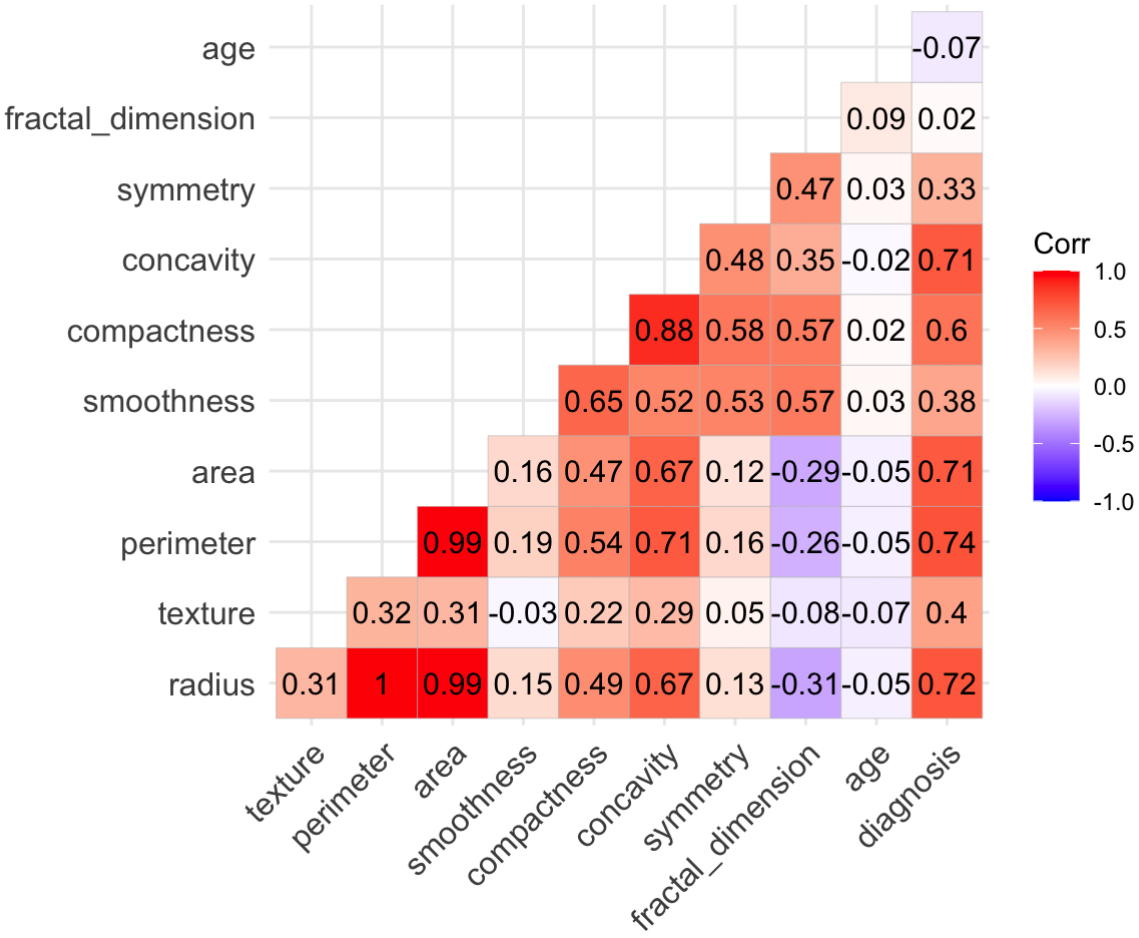
2.1.3 Correlation plot of explanatory variables

Some of the explanatory variables are highly correlated each other such as perimeter and area. Those variables cause Multicollinearity which bring unsuitability to the model.

Code


radius
texture
perimeter
area
moothnes
ompactnes
concavity
symmetry
tal_dimens
age
diagnosis
0.31
1.00
0.99
0.15
0.49
0.67
0.13
-0.31
-0.05
0.72
0.32
0.31
-0.03
0.22
0.29
0.05
-0.08
-0.07
0.40
0.99
0.19
0.54
0.71
0.16
-0.26
-0.05
0.74
0.16
0.47
0.67
0.12
-0.29
-0.05
0.71
0.65
0.52
0.53
0.57
0.03
0.38
0.88
0.58
0.57
0.02
0.60
0.48
0.35
-0.02
0.71
0.47
0.03
0.33
0.09
0.02
-0.07
10 30
500 2500
0.05 0.30
0.10 0.30
20 50
10 25
50 150
0.06 0.14
0.0 0.3
0.05 0.09
0.0 0.8

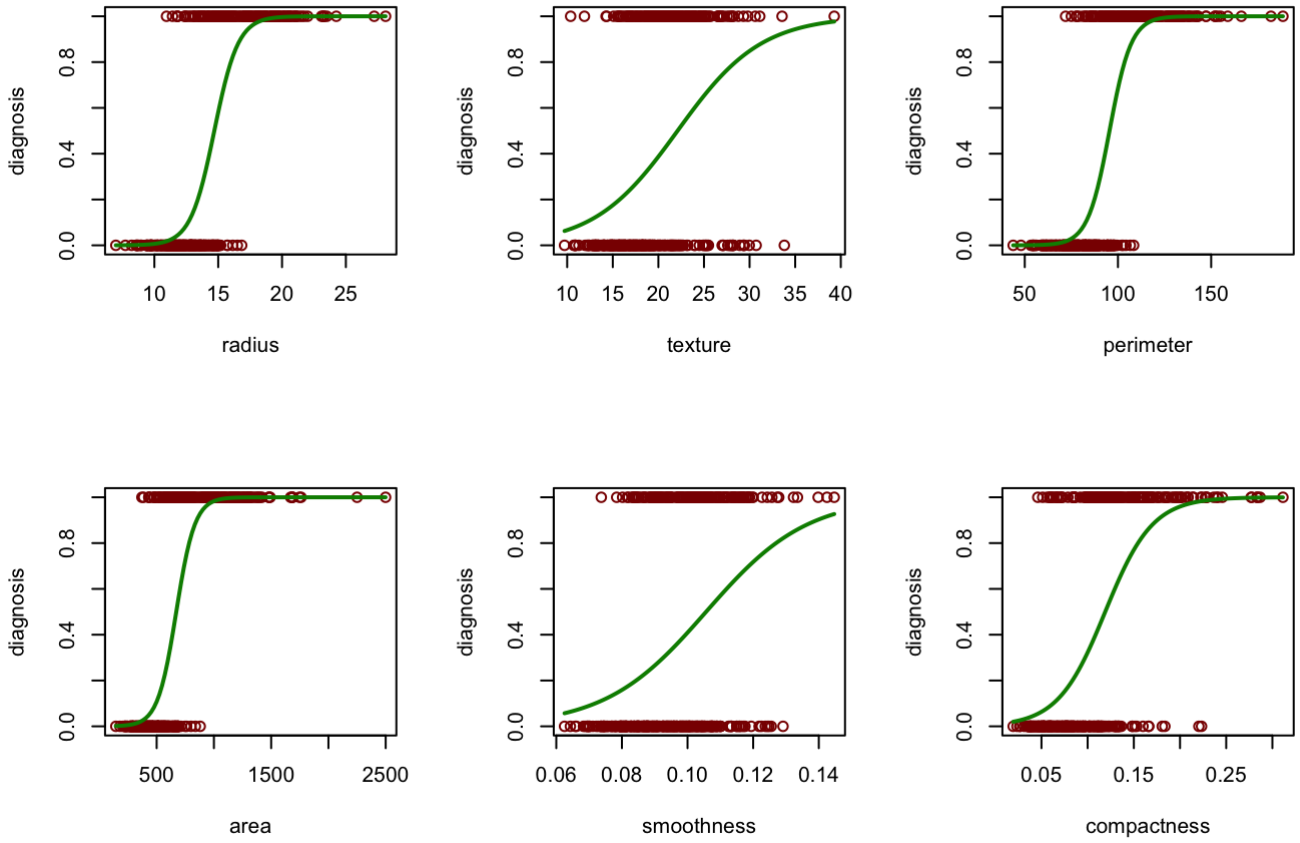
Code



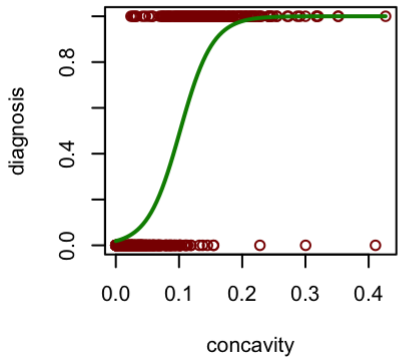
2.1.4 Univariate approach - simple regression

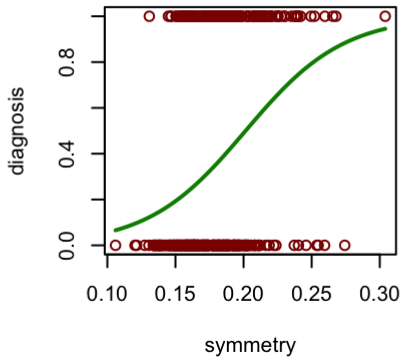
Some of the explanatory variables look promising such as radius and perimeter, while age and fractal_dimension don't appear to be beneficial to fit the data.

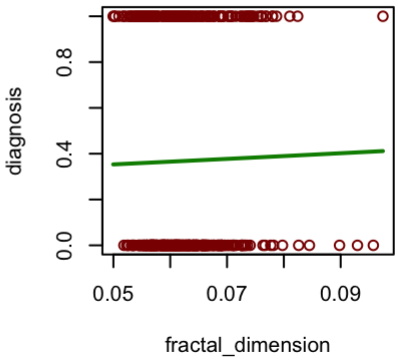
Code

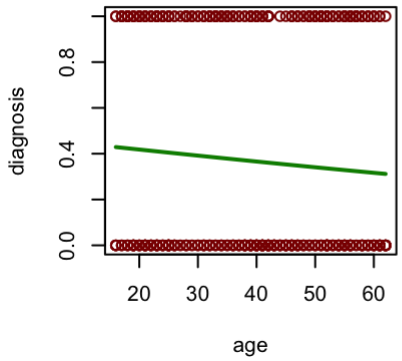

diagnosis
0.0
0.4
0.8
10
15
20
25
radius
10
15
20
25
30
35
40
texture
50
100
150
perimeter
500
1500
2500
area
0.06
0.08
0.10
0.12
0.14
smoothness
0.05
0.15
0.25
compactness

Code









2.2 Model selection

The stepwise method chooses texture, area, smoothness and concavity to fit the data using AIC score. This is the final model.

$$logit(\hat{\pi}) = -29.62423 + 0.38702X_1 + 0.01458X_2 + 101.10562X_3 + 30.31260X_4$$

$\hat{\pi}$: probability of being diagnosed by breast cancer X_1 : Texture
X_2 : Area
X_3 : Smoothness
X_4 : Concavity

```
## Start:  AIC=140.86
## diagnosis ~ radius + texture + perimeter + area + smoothness +
##     compactness + concavity + symmetry + fractal_dimension +
##     age
##
##                      Df Deviance    AIC
## - compactness         1   118.93 138.93
## - age                 1   119.04 139.04
## - perimeter           1   119.22 139.22
## - fractal_dimension   1   119.73 139.73
## - radius              1   120.49 140.49
## - symmetry            1   120.84 140.84
## <none>                    118.86 140.86
## - concavity           1   124.75 144.75
## - area                1   125.56 145.56
## - smoothness          1   149.02 169.02
## - texture             1   154.22 174.22
##
## Step:  AIC=138.93
## diagnosis ~ radius + texture + perimeter + area + smoothness +
##     concavity + symmetry + fractal_dimension + age
##
##                      Df Deviance    AIC
## - age                 1   119.07 137.07
## - perimeter           1   119.40 137.40
## - fractal_dimension   1   120.62 138.62
## - symmetry            1   120.84 138.84
## <none>                    118.93 138.93
## - radius              1   121.18 139.18
## - concavity           1   124.80 142.80
## - area                1   126.74 144.74
## - smoothness          1   151.06 169.06
## - texture             1   154.45 172.45
##
## Step:  AIC=137.07
## diagnosis ~ radius + texture + perimeter + area + smoothness +
##     concavity + symmetry + fractal_dimension
##
##                      Df Deviance    AIC
## - perimeter           1   119.58 135.57
## - fractal_dimension   1   120.94 136.94
## - symmetry            1   121.05 137.05
## <none>                    119.07 137.07
## - radius              1   121.38 137.38
## - concavity           1   125.00 141.00
## - area                1   126.88 142.88
## - smoothness          1   151.63 167.63
## - texture             1   155.33 171.33
##
## Step:  AIC=135.58
## diagnosis ~ radius + texture + area + smoothness + concavity +
##     symmetry + fractal_dimension
```

```
##
##                     Df Deviance    AIC
## - fractal_dimension  1   120.95 134.95
## - symmetry           1   121.54 135.54
## <none>                   119.58 135.57
## - radius             1   123.42 137.42
## - area               1   127.00 141.00
## - concavity          1   128.57 142.57
## - smoothness         1   151.88 165.88
## - texture            1   156.77 170.77
##
## Step:  AIC=134.95
## diagnosis ~ radius + texture + area + smoothness + concavity +
##     symmetry
##
##              Df Deviance    AIC
## - symmetry    1   122.63 134.63
## <none>            120.95 134.95
## - radius      1   125.16 137.16
## - area        1   129.64 141.64
## - concavity   1   130.22 142.22
## - smoothness  1   152.19 164.19
## - texture     1   159.06 171.06
##
## Step:  AIC=134.63
## diagnosis ~ radius + texture + area + smoothness + concavity
##
##              Df Deviance    AIC
## <none>            122.63 134.63
## - radius      1   127.22 137.22
## - area        1   131.90 141.90
## - concavity   1   138.29 148.29
## - texture     1   159.40 169.40
## - smoothness  1   161.45 171.45
```

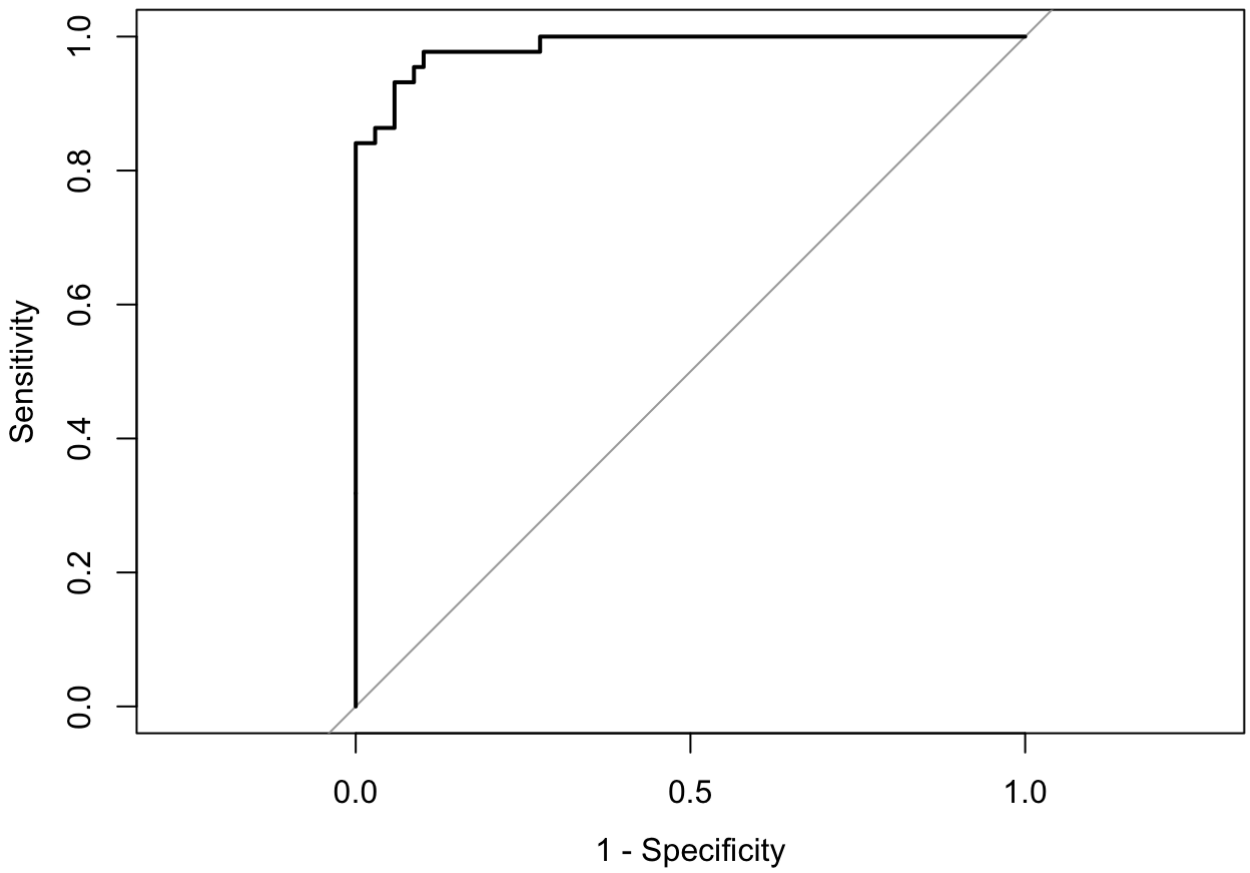
```
##
## Call:  glm(formula = diagnosis ~ radius + texture + area + smoothness +
##     concavity, family = binomial, data = train_df)
##
## Coefficients:
## (Intercept)       radius      texture         area   smoothness    concavity
##   -12.08228     -2.98142      0.35252      0.05068    140.48387     16.18013
##
## Degrees of Freedom: 455 Total (i.e. Null);  450 Residual
## Null Deviance:       600.2
## Residual Deviance: 122.6     AIC: 134.6
```

```
##
## Call:
## glm(formula = diagnosis ~ texture + area + smoothness + concavity,
##     family = binomial, data = train_df)
##
## Deviance Residuals:
##      Min        1Q    Median        3Q       Max
## -1.95048  -0.16211  -0.03884   0.01819   3.11619
##
## Coefficients:
##               Estimate Std. Error z value Pr(>|z|)
## (Intercept) -31.131502   4.375815  -7.114 1.12e-12 ***
## texture       0.353013   0.066002   5.349 8.87e-08 ***
## area          0.015127   0.002152   7.028 2.09e-12 ***
## smoothness  129.234770  25.362328   5.096 3.48e-07 ***
## concavity    17.757038   4.128404   4.301 1.70e-05 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 600.20  on 455  degrees of freedom
## Residual deviance: 127.22  on 451  degrees of freedom
## AIC: 137.22
##
## Number of Fisher Scoring iterations: 8
```

3 Result

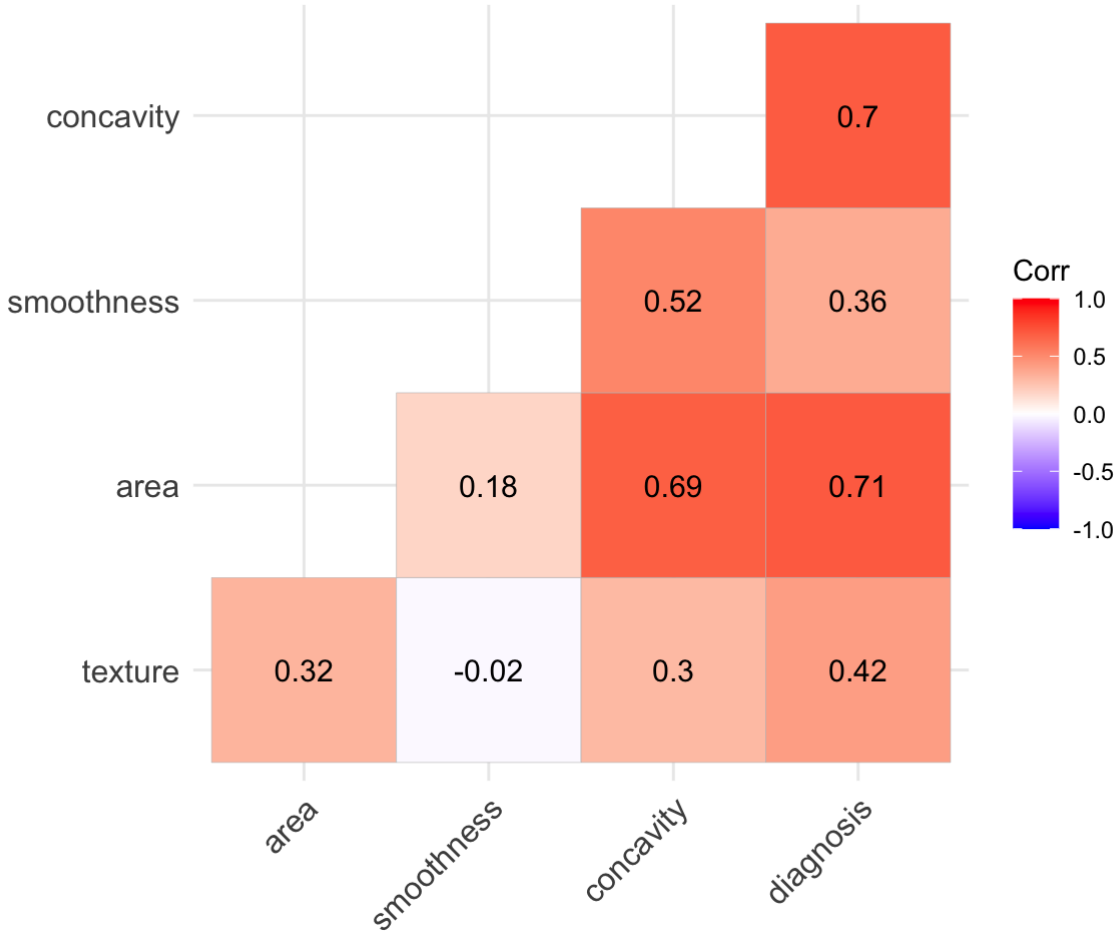
3.1 Evaluate the final model

The final model fits to a test data set which is not used to build the final model. AUC is 0.98 which is a quite high score and the ROC curve looks promising. The variance inflation factor(VIF) shows that the chosen explanatory variables don’t appear to have multicollinearity.




```
## Area under the curve: 0.9848
```

```
##    texture       area smoothness  concavity
##   1.624707   1.540884   1.942022   1.110075
```




```
## Analysis of Deviance Table (Type II tests)
##
## Response: diagnosis
##            LR Chisq Df Pr(>Chisq)
## texture      38.347  1  5.923e-10 ***
## area        136.710  1  < 2.2e-16 ***
## smoothness   34.709  1  3.829e-09 ***
## concavity    16.664  1  4.463e-05 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

4 Discussions

The final model is introduced and the model fits the test dataset which displays in the ROC curve plot. However, without the domain knowledge, letting the algorithm selects the explanatory variables is not a great way of building the model. If possible, going over with medical professionals would be a proper way of analyzing this dataset.

5 Appendix

Code