MSCI 718 – INDIVIDUAL ASSIGNMENT 4

Problem statement and data used:

The dataset used is Breast cancer dataset which has various information such as the mean, standard error and the worst of radius, area, texture, perimeter, compactness, smoothness, symmetry, fractal dimension etc. This dataset contains 569 objects of 33 variables. This report is about the analysis] to find out the effect of **worst radius**(radius_worst; min: 7.93, max: 36.04, mean:16.27), **mean of symmetry**(symmetry_mean; min: 0.106 , max: 0.304, mean: 0.18) and **mean of fractal dimension**(fractal_dimension_mean; min: 0.049, max: 0.097, mean: 0.0628) in determining if the cancer cell is Benign or Malignant. I selected these variables based on past researches in the area.

As part of data cleaning, checked if there were any missing data and found none and plotted boxplots to check outliers and removed them using IQR method.

Planning

The regression analysis is to predict the **diagnosis** of a cancer cell(Malignant or Benign) by the **worst radius, mean of symmetry, mean of fractal dimension.** In order to draw conclusions from the regression analysis, several assumptions were checked

- It was checked if there were any incomplete information. Made sure that all the variables of interest had no missing values.
- Verified there is no complete separation between the variables of interest.

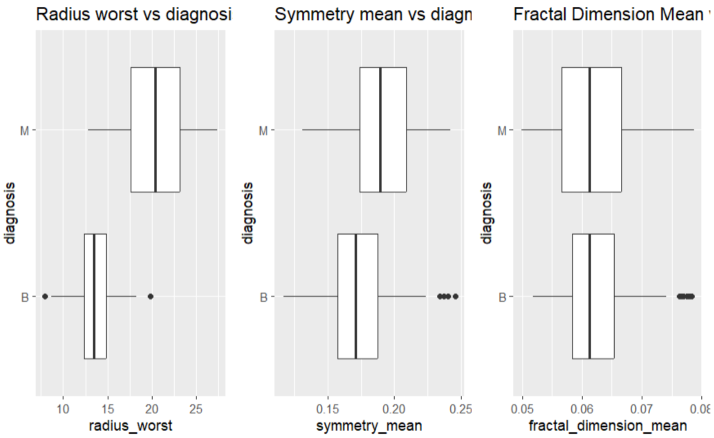


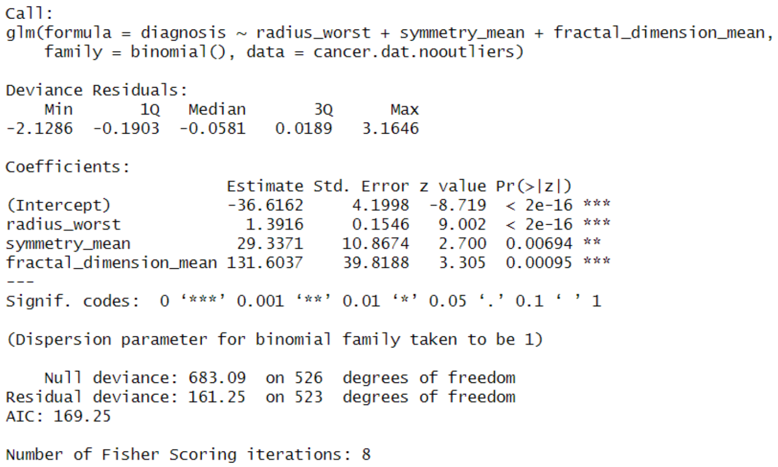
Fig 1: Boxplots to verify complete separation between variables

- The VIF and tolerance statistics were used to assess collinearity. The largest VIF is 1.49 which is much lesser than 10; the average VIF is 1.36 which is close to 1. The lowest tolerance(1/VIF) is 0.66, greater than 0.1(which would indicate serious problem) and 0.2(which would indicate potential problem). Thus, concluded that there is no collinearity in my dataset.

- The Durbin-Watson test for independent errors was not significant at the 5% level of significance (d=1.67, p=0). As d is very close to 2 (which would indicate no autocorrelation detected), we do not reject the null hypothesis that the errors are independent and continue with the assumption of independence met.
- Linearity of the logit has been assessed. All the four interactions have significance values greater than 0.05 indicating that assumption of linearity of the logit has been met for radius_worst, symmetry_mean and fractal_dimension_mean.

Analysis

Logistic regression is conducted on the dataset and model is created.



```
Call:
glm(formula = diagnosis ~ radius_worst + symmetry_mean + fractal_dimension_mean, 
    family = binomial(), data = cancer.dat.nooutliers)

Deviance Residuals: 
    Min       1Q   Median       3Q      Max  
-2.1286  -0.1903  -0.0581   0.0189   3.1646  

Coefficients:
                       Estimate Std. Error z value Pr(>|z|)    
(Intercept)            -36.6162     4.1998  -8.719  < 2e-16 ***
radius_worst             1.3916     0.1546   9.002  < 2e-16 ***
symmetry_mean           29.3371    10.8674   2.700  0.00694 ** 
fractal_dimension_mean 131.6037    39.8188   3.305  0.00095 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 683.09  on 526  degrees of freedom
Residual deviance: 161.25  on 523  degrees of freedom
AIC: 169.25

Number of Fisher Scoring iterations: 8
```

Fig 2: Summary of the logistic regression model

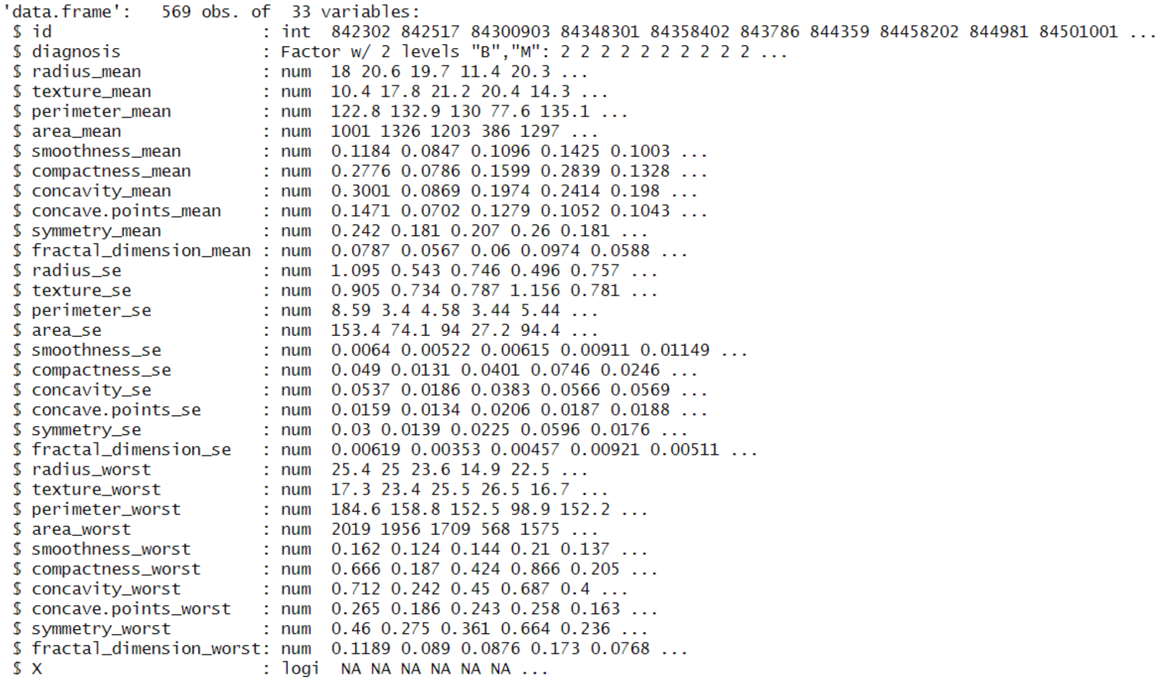
From the logistic regression model, we can see that the deviance of the model(161.25) is less than the value for null model(683.09). This indicates that the model is better at predicting whether the cancer is Malignant or Benign that it was before when the variables of interest were added. We can also see the change in the degrees of freedom of null model and our model is 4 which reflects the fact that we have only used 4 variables in the model. From the z-statistics we could say the following: (i) radius_worst is a significant predictor for diagnosis to be Benign or Malignnat, b = 1.4, z = 9.002, p <0.05 (ii) symmetry_mean is a significant predictor for diagnosis to be Benign or Malignant, b = 29.33, z = 2.7, p <0.05 (iv) fractal_dimension_mean is a significant predictor for diagnosis to be Benign or Malignant, b = 131.6, z = 3.35, p <0.05. The confidence intervals were calculated for all the variables of interest and all are greater than 1. This indicates that with these variables, the cancer is significantly more likely to be Malignant at 5% level of significance.

Conclusion

A logistic regression model is created to predict the diagnosis of cancer cell depending on worst radius, mean of symmetry and mean of fractal dimension. From the analysis it is clear that our model does not violate any assumptions of logistic regression and our predictors are making significant contribution to the prediction of diagnosis.

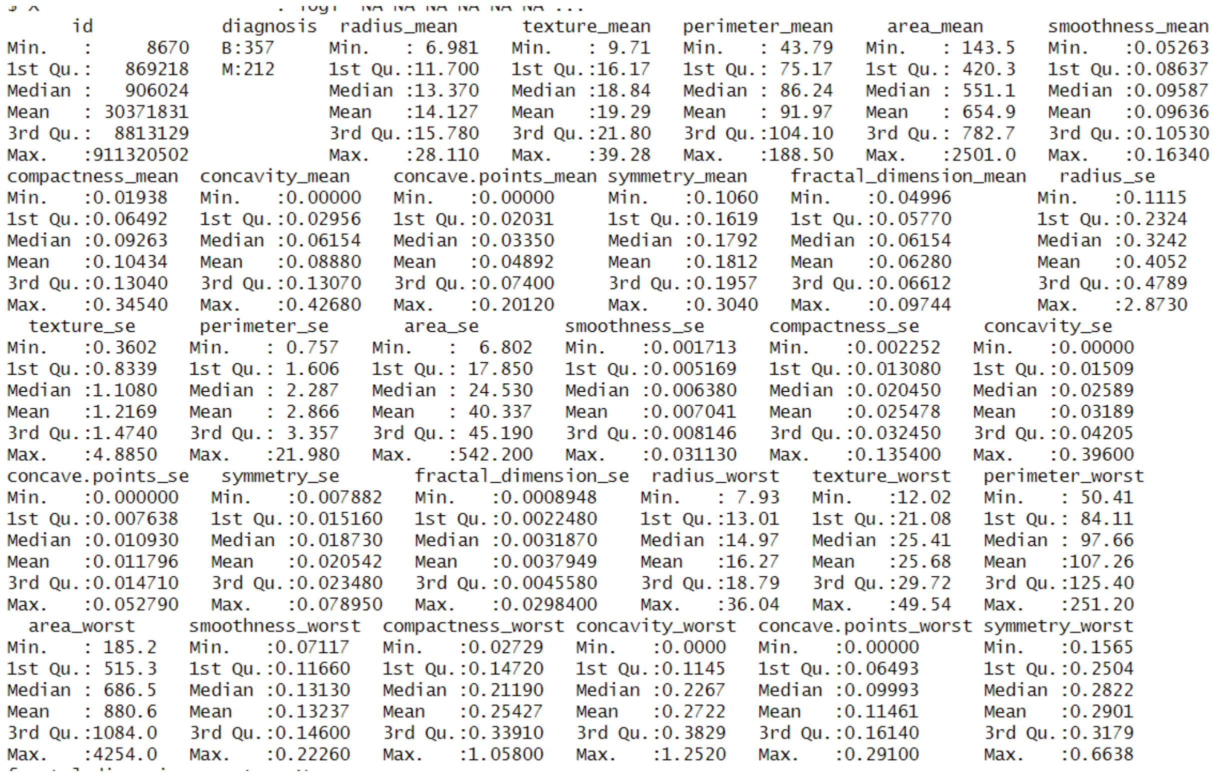
Appendices

Structure of Breast Cancer dataset:



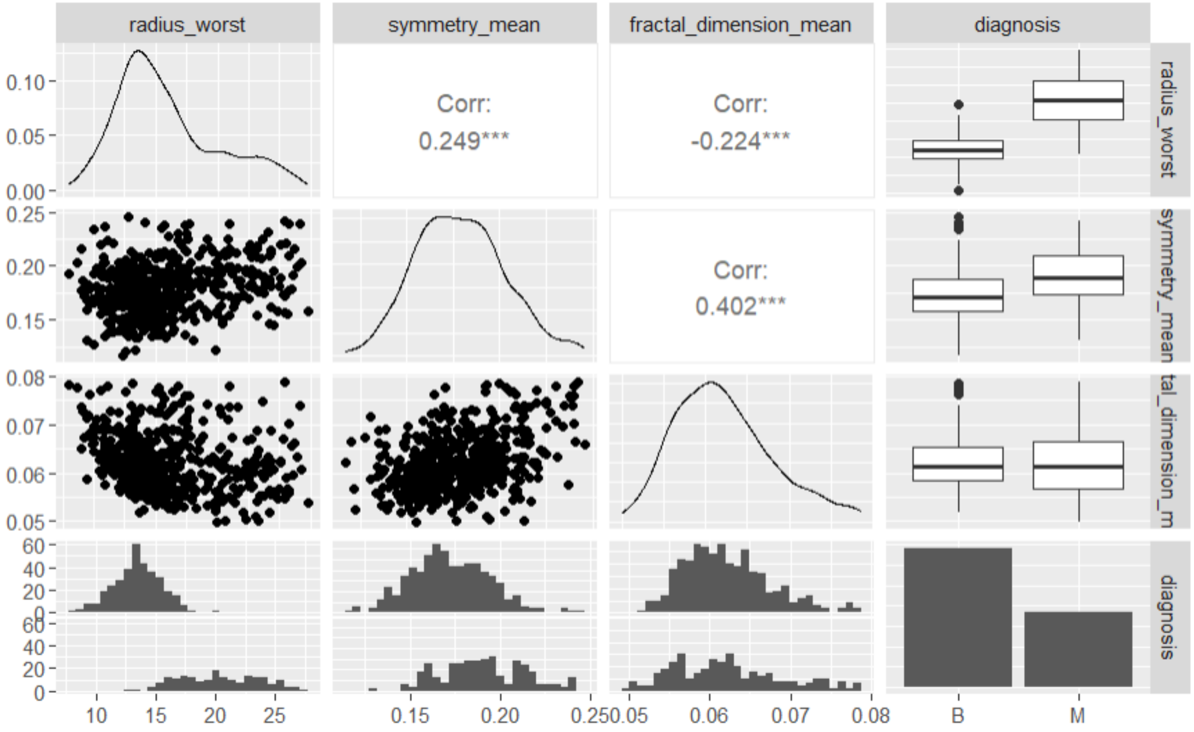
```
'data.frame':   569 obs. of  33 variables:
 $ id                     : int  842302 842517 84300903 84348301 84358402 843786 844359 84458202 844981 84501001 ...
 $ diagnosis              : Factor w/ 2 levels "B","M": 2 2 2 2 2 2 2 2 2 2 ...
 $ radius_mean            : num  18 20.6 19.7 11.4 20.3 ...
 $ texture_mean           : num  10.4 17.8 21.2 20.4 14.3 ...
 $ perimeter_mean         : num  122.8 132.9 130 77.6 135.1 ...
 $ area_mean              : num  1001 1326 1203 386 1297 ...
 $ smoothness_mean        : num  0.1184 0.0847 0.1096 0.1425 0.1003 ...
 $ compactness_mean       : num  0.2776 0.0786 0.1599 0.2839 0.1328 ...
 $ concavity_mean         : num  0.3001 0.0869 0.1974 0.2414 0.198 ...
 $ concave.points_mean    : num  0.1471 0.0702 0.1279 0.1052 0.1043 ...
 $ symmetry_mean          : num  0.242 0.181 0.207 0.26 0.181 ...
 $ fractal_dimension_mean : num  0.0787 0.0567 0.06 0.0974 0.0588 ...
 $ radius_se              : num  1.095 0.543 0.746 0.496 0.757 ...
 $ texture_se             : num  0.905 0.734 0.787 1.156 0.781 ...
 $ perimeter_se           : num  8.59 3.4 4.58 3.44 5.44 ...
 $ area_se                : num  153.4 74.1 94 27.2 94.4 ...
 $ smoothness_se          : num  0.0064 0.00522 0.00615 0.00911 0.01149 ...
 $ compactness_se         : num  0.049 0.0131 0.0401 0.0746 0.0246 ...
 $ concavity_se           : num  0.0537 0.0186 0.0383 0.0566 0.0569 ...
 $ concave.points_se      : num  0.0159 0.0134 0.0206 0.0187 0.0188 ...
 $ symmetry_se            : num  0.03 0.0139 0.0225 0.0596 0.0176 ...
 $ fractal_dimension_se   : num  0.00619 0.00353 0.00457 0.00921 0.00511 ...
 $ radius_worst           : num  25.4 25 23.6 14.9 22.5 ...
 $ texture_worst          : num  17.3 23.4 25.5 26.5 16.7 ...
 $ perimeter_worst        : num  184.6 158.8 152.5 98.9 152.2 ...
 $ area_worst             : num  2019 1956 1709 568 1575 ...
 $ smoothness_worst       : num  0.162 0.124 0.144 0.21 0.137 ...
 $ compactness_worst      : num  0.666 0.187 0.424 0.866 0.205 ...
 $ concavity_worst        : num  0.712 0.242 0.45 0.687 0.4 ...
 $ concave.points_worst   : num  0.265 0.186 0.243 0.258 0.163 ...
 $ symmetry_worst         : num  0.46 0.275 0.361 0.664 0.236 ...
 $ fractal_dimension_worst: num  0.1189 0.089 0.0876 0.173 0.0768 ...
 $ X                      : logi  NA NA NA NA NA NA ...
```

Summary of the dataset

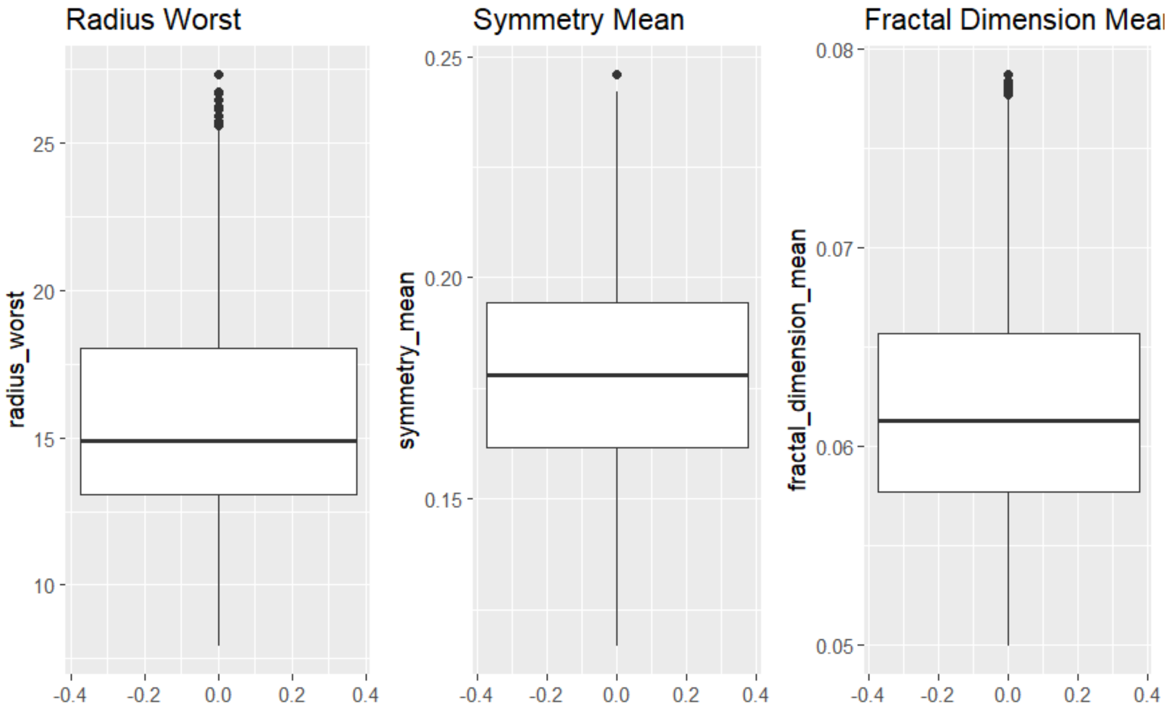


```
 $ X                      : logi  NA NA NA NA NA NA ...
       id             diagnosis  radius_mean      texture_mean   perimeter_mean     area_mean      smoothness_mean  
 Min.   :     8670   B:357     Min.   : 6.981   Min.   : 9.71   Min.   : 43.79   Min.   : 143.5   Min.   :0.05263  
 1st Qu.:   869218   M:212     1st Qu.:11.700   1st Qu.:16.17   1st Qu.: 75.17   1st Qu.: 420.3   1st Qu.:0.08637  
 Median :   906024             Median :13.370   Median :18.84   Median : 86.24   Median : 551.1   Median :0.09587  
 Mean   : 30371831             Mean   :14.127   Mean   :19.29   Mean   : 91.97   Mean   : 654.9   Mean   :0.09636  
 3rd Qu.:  8813129             3rd Qu.:15.780   3rd Qu.:21.80   3rd Qu.:104.10   3rd Qu.: 782.7   3rd Qu.:0.10530  
 Max.   :911320502             Max.   :28.110   Max.   :39.28   Max.   :188.50   Max.   :2501.0   Max.   :0.16340  
 compactness_mean  concavity_mean    concave.points_mean symmetry_mean    fractal_dimension_mean   radius_se     
 Min.   :0.01938   Min.   :0.00000   Min.   :0.00000     Min.   :0.1060   Min.   :0.04996        Min.   :0.1115  
 1st Qu.:0.06492   1st Qu.:0.02956   1st Qu.:0.02031     1st Qu.:0.1619   1st Qu.:0.05770        1st Qu.:0.2324  
 Median :0.09263   Median :0.06154   Median :0.03350     Median :0.1792   Median :0.06154        Median :0.3242  
 Mean   :0.10434   Mean   :0.08880   Mean   :0.04892     Mean   :0.1812   Mean   :0.06280        Mean   :0.4052  
 3rd Qu.:0.13040   3rd Qu.:0.13070   3rd Qu.:0.07400     3rd Qu.:0.1957   3rd Qu.:0.06612        3rd Qu.:0.4789  
 Max.   :0.34540   Max.   :0.42680   Max.   :0.20120     Max.   :0.3040   Max.   :0.09744        Max.   :2.8730  
   texture_se      perimeter_se       area_se         smoothness_se       compactness_se      concavity_se    
 Min.   :0.3602   Min.   : 0.757   Min.   :  6.802   Min.   :0.001713   Min.   :0.002252   Min.   :0.00000  
 1st Qu.:0.8339   1st Qu.: 1.606   1st Qu.: 17.850   1st Qu.:0.005169   1st Qu.:0.013080   1st Qu.:0.01509  
 Median :1.1080   Median : 2.287   Median : 24.530   Median :0.006380   Median :0.020450   Median :0.02589  
 Mean   :1.2169   Mean   : 2.866   Mean   : 40.337   Mean   :0.007041   Mean   :0.025478   Mean   :0.03189  
 3rd Qu.:1.4740   3rd Qu.: 3.357   3rd Qu.: 45.190   3rd Qu.:0.008146   3rd Qu.:0.032450   3rd Qu.:0.04205  
 Max.   :4.8850   Max.   :21.980   Max.   :542.200   Max.   :0.031130   Max.   :0.135400   Max.   :0.39600  
 concave.points_se   symmetry_se        fractal_dimension_se  radius_worst    texture_worst   perimeter_worst 
 Min.   :0.000000   Min.   :0.007882   Min.   :0.0008948    Min.   : 7.93   Min.   :12.02   Min.   : 50.41  
 1st Qu.:0.007638   1st Qu.:0.015160   1st Qu.:0.0022480    1st Qu.:13.01   1st Qu.:21.08   1st Qu.: 84.11  
 Median :0.010930   Median :0.018730   Median :0.0031870    Median :14.97   Median :25.41   Median : 97.66  
 Mean   :0.011796   Mean   :0.020542   Mean   :0.0037949    Mean   :16.27   Mean   :25.68   Mean   :107.26  
 3rd Qu.:0.014710   3rd Qu.:0.023480   3rd Qu.:0.0045580    3rd Qu.:18.79   3rd Qu.:29.72   3rd Qu.:125.40  
 Max.   :0.052790   Max.   :0.078950   Max.   :0.0298400    Max.   :36.04   Max.   :49.54   Max.   :251.20  
   area_worst     smoothness_worst  compactness_worst concavity_worst  concave.points_worst symmetry_worst  
 Min.   : 185.2   Min.   :0.07117   Min.   :0.02729   Min.   :0.0000   Min.   :0.00000      Min.   :0.1565  
 1st Qu.: 515.3   1st Qu.:0.11660   1st Qu.:0.14720   1st Qu.:0.1145   1st Qu.:0.06493      1st Qu.:0.2504  
 Median : 686.5   Median :0.13130   Median :0.21190   Median :0.2267   Median :0.09993      Median :0.2822  
 Mean   : 880.6   Mean   :0.13237   Mean   :0.25427   Mean   :0.2722   Mean   :0.11461      Mean   :0.2901  
 3rd Qu.:1084.0   3rd Qu.:0.14600   3rd Qu.:0.33910   3rd Qu.:0.3829   3rd Qu.:0.16140      3rd Qu.:0.3179  
 Max.   :4254.0   Max.   :0.22260   Max.   :1.05800   Max.   :1.2520   Max.   :0.29100      Max.   :0.6638  
```

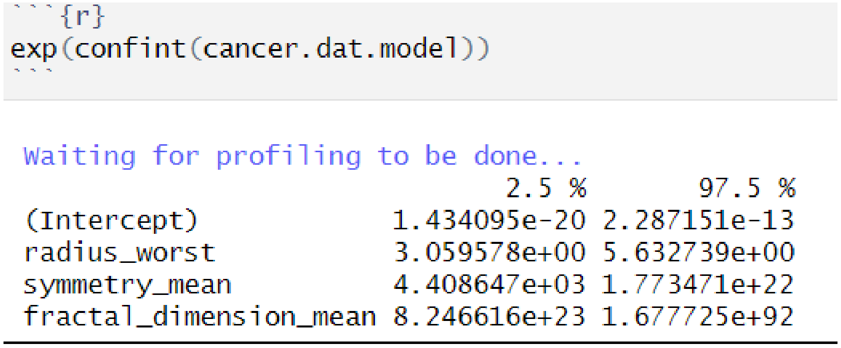
Pairplot of variables under analysis



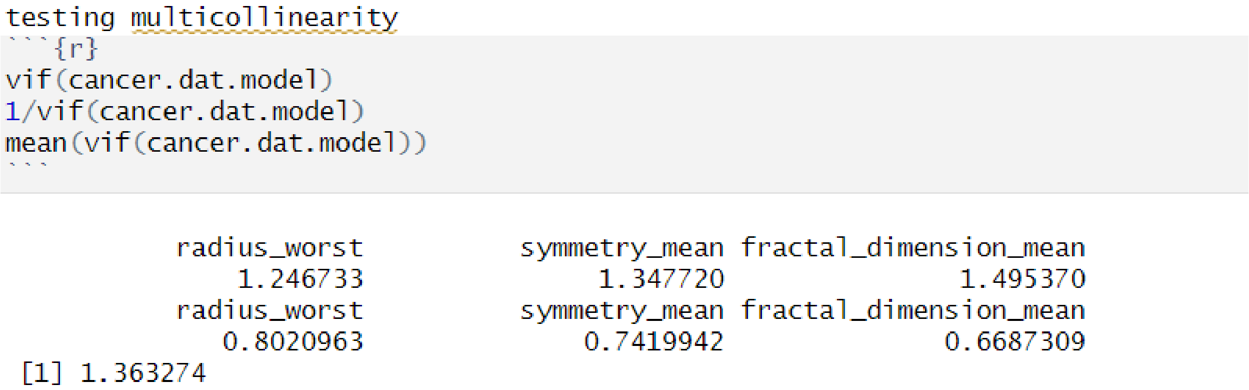
Boxplots after removing outliers:

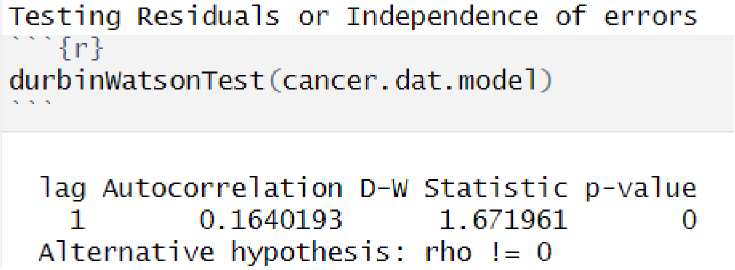


Confidence Intervals:



~~~
```{r}
exp(confint(cancer.dat.model))
~~~

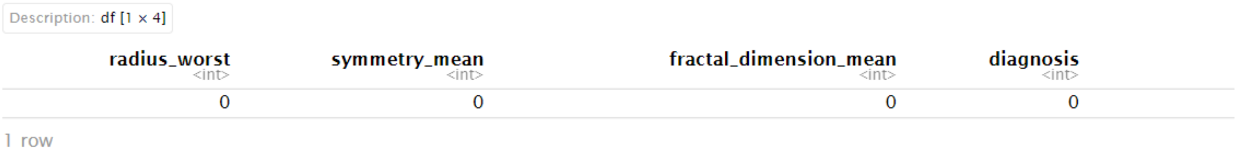
 Waiting for profiling to be done...
                              2.5 %   97.5 %
 (Intercept)   1.434095e-20 2.287151e-13
 radius_worst   3.059578e+00 5.632739e+00
 symmetry_mean   4.408647e+03 1.773471e+22
 fractal_dimension_mean 8.246616e+23 1.677725e+92
```

Result of testing multicollinearity



```
testing multicollinearity
```{r}
vif(cancer.dat.model)
1/vif(cancer.dat.model)
mean(vif(cancer.dat.model))
```

           radius_worst   symmetry_mean fractal_dimension_mean
               1.246733   1.347720   1.495370
           radius_worst   symmetry_mean fractal_dimension_mean
              0.8020963   0.7419942   0.6687309
 [1] 1.363274
```

Result of testing for independence of errors



```
Testing Residuals or Independence of errors
```{r}
durbinWatsonTest(cancer.dat.model)
```

  lag Autocorrelation D-W Statistic p-value
    1   0.1640193   1.671961   0
  Alternative hypothesis: rho != 0
```

Result of testing for linearity
```

```
Call:
glm(formula = diagnosis ~ radius_worst + symmetry_mean + fractal_dimension_mean +
    log.radius_worst + log.symmetry_mean + log.fractal_dimension_mean,
    family = binomial(), data = cancer.dat.nooutliers)

Deviance Residuals:
     Min        1Q    Median        3Q       Max
-2.08440  -0.19974  -0.06321   0.01151   3.11894

Coefficients:
                            Estimate Std. Error z value Pr(>|z|)
(Intercept)                 -46.2043    44.1586  -1.046    0.295
radius_worst                 -1.5482     6.8757  -0.225    0.822
symmetry_mean               -20.1762    78.6449  -0.257    0.798
fractal_dimension_mean     -110.6281  1060.6666  -0.104    0.917
log.radius_worst              0.7833     1.8261   0.429    0.668
log.symmetry_mean           -70.1834   110.1646  -0.637    0.524
log.fractal_dimension_mean -140.4961   607.7320  -0.231    0.817

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 683.09  on 526  degrees of freedom
Residual deviance: 160.65  on 520  degrees of freedom
AIC: 174.65

Number of Fisher Scoring iterations: 8
```

Result of testing for no information loss:



Description: df [1 × 4]

radius_worst <int>	symmetry_mean <int>	fractal_dimension_mean <int>	diagnosis <int>
0	0	0	0

1 row