Julian Politsch
1831710
Biomedical Statistics, Bioinformatics 2019/20

Biostatistical Analysis of Breast Cancer diagnosis in Wisconsin:

Predicting and modeling malignant vs benign tumors using R

Introduction:

The objective of this report will be a meta-analysis of the classification of tumors as malignant or benign based on 10 categories of observations. From the full data set of 569 observations entries, 300 will be randomly selected to preform analysis. The categories are broken into three subdivisions, for each variable, we have I. Mean II. Standard error III. Worst (In most cases largest), thus it is important to consider that while we have 30 change variables, they must be carefully differentiated in order to find meaningful analysis. For example, field 3 is Mean Radius, field 13 is Radius SE, field 23 is Worst Radius. No data is missing from the tables. Our target variable will be diagnosis, which is either ‘M’ or ‘B’ meaning dangerous, or likely not dangerous.

The variables are as follows:

a) radius (mean of distances from center to points on the perimeter)
b) texture (standard deviation of gray-scale values)
c) perimeter
d) area
e) smoothness (local variation in radius lengths)
f) compactness (perimeter^2 / area - 1.0)
g) concavity (severity of concave portions of the contour)
h) concave points (number of concave portions of the contour)
i) symmetry
j) fractal dimension ("coastline approximation" - 1)

Initial analysis:

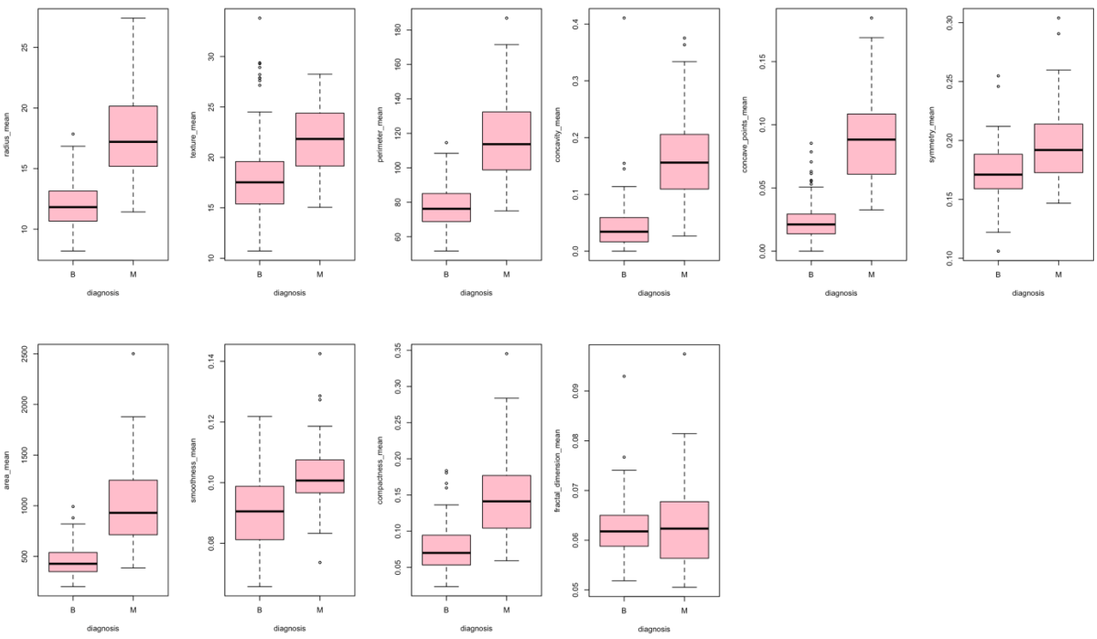
No data is missing from the tables. Our target variable will be diagnosis, which is either ‘M’ or ‘B’ meaning dangerous, or likely not dangerous. By setting the seed to 1831710 we return the following data, noting especially our target variable ‘diagnosis’.

```
 radius_mean      texture_mean   perimeter_mean     area_mean      smoothness_mean   compactness_mean
Min.   : 7.729   Min.   :10.72   Min.   : 47.98   Min.   : 178.8   Min.   :0.06429   Min.   :0.02344
1st Qu.:11.600   1st Qu.:16.33   1st Qu.: 74.34   1st Qu.: 414.5   1st Qu.:0.08640   1st Qu.:0.06252
Median :13.290   Median :18.70   Median : 86.25   Median : 545.6   Median :0.09477   Median :0.08592
Mean   :14.069   Mean   :19.19   Mean   : 91.50   Mean   : 649.4   Mean   :0.09575   Mean   :0.10177
3rd Qu.:16.080   3rd Qu.:21.73   3rd Qu.:104.80   3rd Qu.: 791.1   3rd Qu.:0.10355   3rd Qu.:0.12745
Max.   :27.420   Max.   :33.81   Max.   :186.90   Max.   :2501.0   Max.   :0.14250   Max.   :0.34540
concavity_mean    concave_points_mean symmetry_mean    fractal_dimension_mean   radius_se        texture_se
Min.   :0.00000   Min.   :0.00000     Min.   :0.1060   Min.   :0.05054        Min.   :0.1144   Min.   :0.3602
1st Qu.:0.02706   1st Qu.:0.01911     1st Qu.:0.1631   1st Qu.:0.05765        1st Qu.:0.2343   1st Qu.:0.8387
Median :0.05513   Median :0.03110     Median :0.1792   Median :0.06152        Median :0.3246   Median :1.0725
Mean   :0.08523   Mean   :0.04791     Mean   :0.1816   Mean   :0.06257        Mean   :0.4064   Mean   :1.2187
3rd Qu.:0.12075   3rd Qu.:0.07401     3rd Qu.:0.1954   3rd Qu.:0.06574        3rd Qu.:0.4889   3rd Qu.:1.4800
Max.   :0.42640   Max.   :0.19130     Max.   :0.3040   Max.   :0.09744        Max.   :2.5470   Max.   :4.8850
 perimeter_se       area_se        smoothness_se      compactness_se     concavity_se     concave_points_se
Min.   : 0.757   Min.   :  6.802   Min.   :0.001713   Min.   :0.00371   Min.   :0.00000   Min.   :0.000000
1st Qu.: 1.611   1st Qu.: 18.233   1st Qu.:0.005078   1st Qu.:0.01272   1st Qu.:0.01429   1st Qu.:0.007565
Median : 2.287   Median : 24.565   Median :0.006293   Median :0.01909   Median :0.02415   Median :0.010515
Mean   : 2.851   Mean   : 39.907   Mean   :0.006993   Mean   :0.02461   Mean   :0.03069   Mean   :0.011692
3rd Qu.: 3.384   3rd Qu.: 45.385   3rd Qu.:0.008279   3rd Qu.:0.03038   3rd Qu.:0.03924   3rd Qu.:0.014350
Max.   :18.650   Max.   :542.200   Max.   :0.021770   Max.   :0.10640   Max.   :0.39600   Max.   :0.052790
 symmetry_se       fractal_dimension_se  radius_worst    texture_worst   perimeter_worst    area_worst
Min.   :0.007882   Min.   :0.0008948    Min.   : 8.964   Min.   :12.49   Min.   : 57.17   Min.   : 242.2
1st Qu.:0.015360   1st Qu.:0.0021533    1st Qu.:12.848   1st Qu.:21.29   1st Qu.: 83.11   1st Qu.: 508.5
Median :0.018955   Median :0.0030285    Median :14.910   Median :25.21   Median : 97.14   Median : 684.0
Mean   :0.020764   Mean   :0.0036793    Mean   :16.179   Mean   :25.42   Mean   :106.37   Mean   : 871.6
3rd Qu.:0.023830   3rd Qu.:0.0042673    3rd Qu.:18.550   3rd Qu.:29.35   3rd Qu.:123.85   3rd Qu.:1038.8
Max.   :0.078950   Max.   :0.0298400    Max.   :36.040   Max.   :49.54   Max.   :251.20   Max.   :4254.0
smoothness_worst  compactness_worst concavity_worst  concave_points_worst symmetry_worst
Min.   :0.07117   Min.   :0.02729   Min.   :0.0000   Min.   :0.00000      Min.   :0.1565
1st Qu.:0.11650   1st Qu.:0.14527   1st Qu.:0.1050   1st Qu.:0.06294      1st Qu.:0.2513
Median :0.12950   Median :0.20060   Median :0.2069   Median :0.09287      Median :0.2824
Mean   :0.13136   Mean   :0.24164   Mean   :0.2583   Mean   :0.11184      Mean   :0.2905
3rd Qu.:0.14570   3rd Qu.:0.30635   3rd Qu.:0.3645   3rd Qu.:0.15573      3rd Qu.:0.3181
Max.   :0.20980   Max.   :1.05800   Max.   :1.1050   Max.   :0.29100      Max.   :0.6638
fractal_dimension_worst diagnosis
Min.   :0.05504         B:194
1st Qu.:0.07069         M:106
Median :0.07848
Mean   :0.08260
3rd Qu.:0.09026
Max.   :0.20750
```

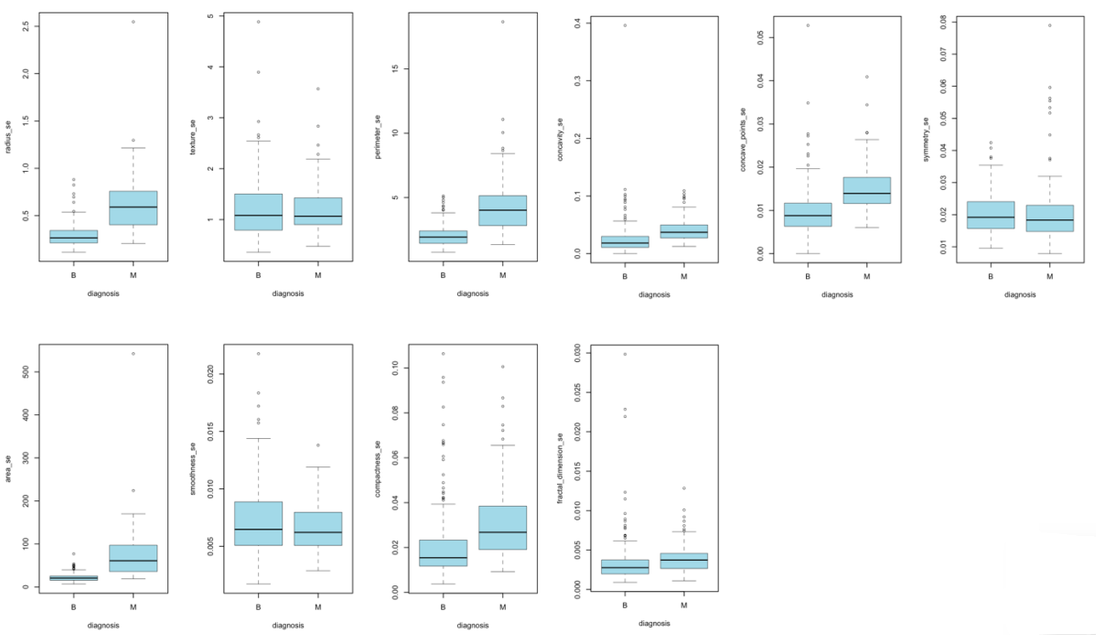
Descriptive analysis:

Boxplots show that for almost every category, malignant mean is higher than benign. This gives us a basis for analysis, but further studying is required.

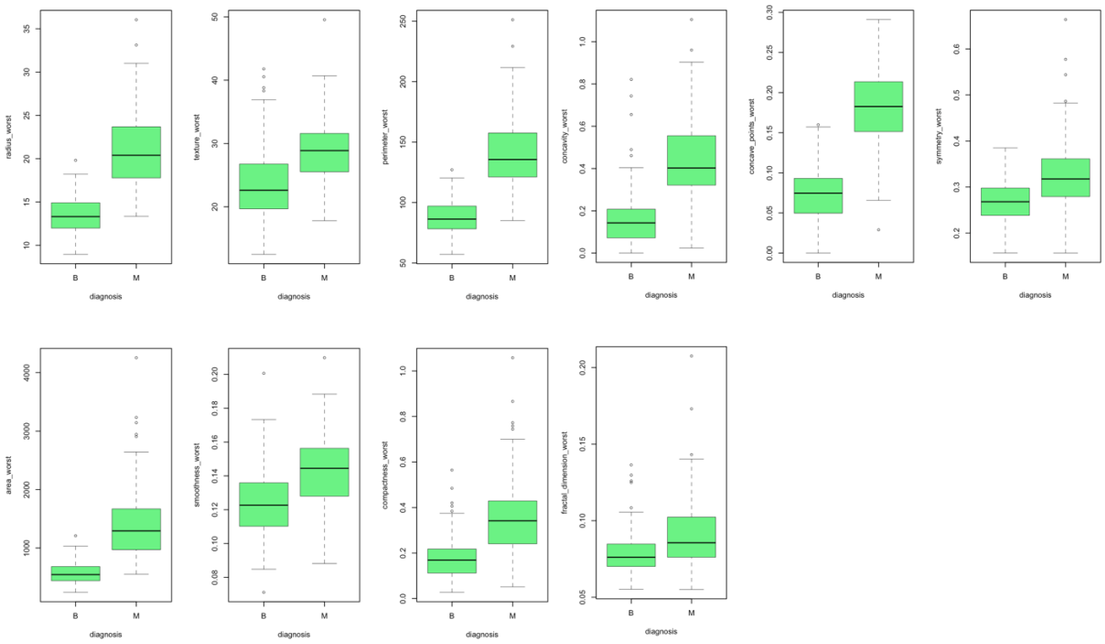
Mean box plots:



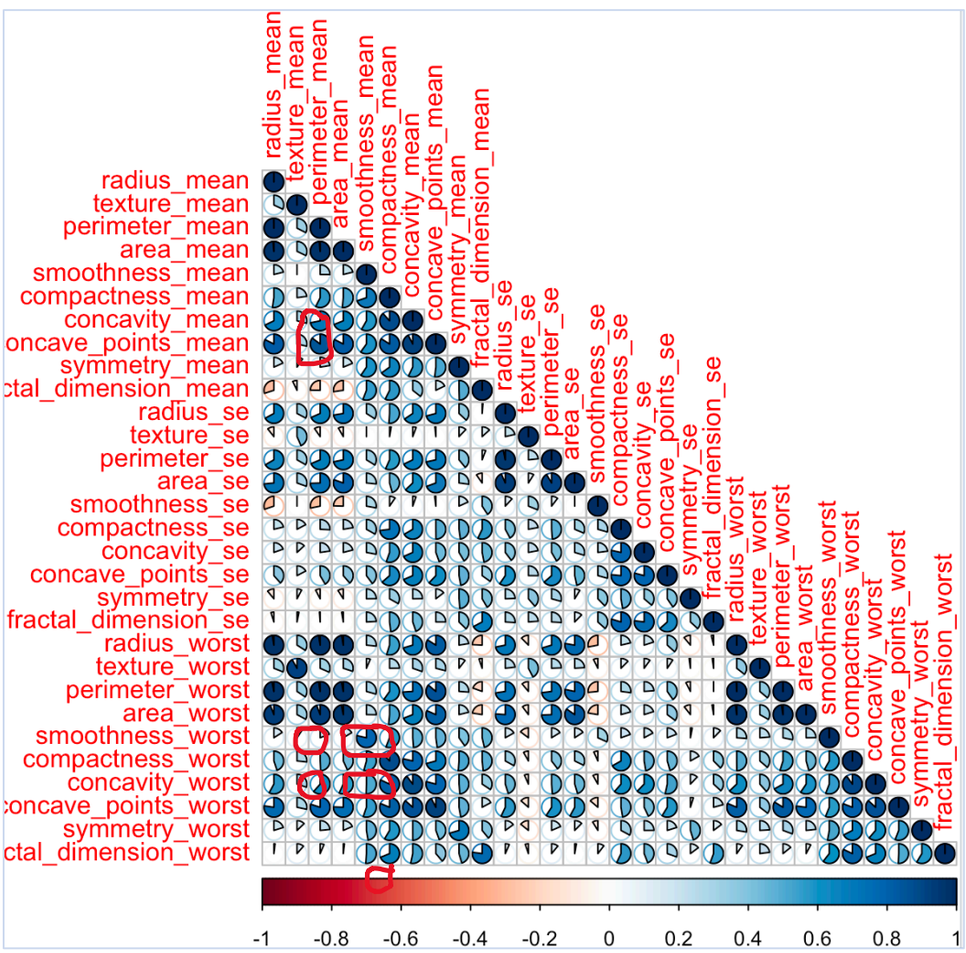
SE box plots:



Worst box plots:



Lastly we use this pie chart analysis to see which data points can be synthisized together, and which are too corilated to produce a meaningful model. Data couples approaching 1 can not be used together, some examples are perimeter_mean and radius_mean, area_mean and perimeter_worst.



Linear Regression Analysis:

```
Coefficients:
                          Estimate Std. Error z value Pr(>|z|)
(Intercept)             -1.085e+03  1.802e+06  -0.001    1.000
radius_mean             -3.000e+02  4.941e+05  -0.001    1.000
texture_mean             2.847e+00  4.142e+04   0.000    1.000
perimeter_mean           4.978e+01  5.529e+04   0.001    0.999
area_mean               -3.043e-01  4.409e+03   0.000    1.000
smoothness_mean          2.948e+03  6.642e+06   0.000    1.000
compactness_mean        -3.097e+03  2.308e+06  -0.001    0.999
concavity_mean           5.299e+02  1.951e+06   0.000    1.000
concave_points_mean     -4.257e+02  4.985e+06   0.000    1.000
symmetry_mean           -6.342e+02  4.143e+06   0.000    1.000
fractal_dimension_mean   3.327e+03  1.161e+07   0.000    1.000
radius_se                2.696e+02  1.107e+06   0.000    1.000
texture_se              -3.058e+01  1.899e+05   0.000    1.000
perimeter_se            -2.232e+01  8.352e+04   0.000    1.000
area_se                  1.674e+00  1.183e+04   0.000    1.000
smoothness_se           -8.358e+03  5.078e+07   0.000    1.000
compactness_se          -9.340e+01  8.091e+06   0.000    1.000
concavity_se             8.217e+02  6.530e+06   0.000    1.000
concave_points_se        7.389e+03  6.884e+07   0.000    1.000
symmetry_se              1.478e+03  2.286e+07   0.000    1.000
fractal_dimension_se    -3.262e+04  4.729e+07  -0.001    0.999
radius_worst             2.270e+00  2.755e+05   0.000    1.000
texture_worst            3.545e+00  3.150e+04   0.000    1.000
perimeter_worst          1.870e+00  1.908e+04   0.000    1.000
area_worst               5.520e-02  2.135e+03   0.000    1.000
smoothness_worst         4.371e+02  4.283e+06   0.000    1.000
compactness_worst       -2.777e+02  1.005e+06   0.000    1.000
concavity_worst          4.955e+01  6.283e+05   0.000    1.000
concave_points_worst     9.778e+01  6.623e+06   0.000    1.000
symmetry_worst          -4.915e+01  3.302e+06   0.000    1.000
fractal_dimension_worst  4.799e+03  8.314e+06   0.001    1.000

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 3.8969e+02  on 299  degrees of freedom
Residual deviance: 5.8270e-08  on 269  degrees of freedom
AIC: 62

Number of Fisher Scoring iterations: 25
```

Initial analysis of all variables provided shows an impossible to analysis correlation. This is because some of the variables are tremendously correlated to each other –as seen in the pie charts -- thus we must narrow down the field of variables to ones that will give us a better model

Pictured: Full Model Results

In the second attempt, I used a machine learning algorithm output that rates the analysis of the variables created by Kaggle.com to be used to analysis the data provided. The output rates the variables coordination, and therefore I was able to identify the 20 (of 30) most important, variables. In model1 we explore the accuracy of all of these 20 variables. It is clear we must continue to eliminate variables in order to make the model even more accurate, we still see some data is not correctly applied

Pictured: model 1

```
Deviance Residuals:
       Min          1Q      Median          3Q         Max
-3.390e-04  -2.000e-08  -2.000e-08   2.000e-08   3.421e-04

Coefficients:
                         Estimate Std. Error z value Pr(>|z|)
(Intercept)            -2.011e+03  5.999e+05  -0.003    0.997
compactness_mean       -9.811e+03  8.729e+05  -0.011    0.991
area_worst              1.802e+00  4.674e+02   0.004    0.997
texture_mean            1.457e+01  2.284e+03   0.006    0.995
radius_mean            -1.577e+03  1.936e+05  -0.008    0.994
perimeter_mean          1.967e+02  1.656e+04   0.012    0.991
area_mean               1.321e+00  1.064e+03   0.001    0.999
concave_points_worst    8.967e+02  2.940e+05   0.003    0.998
compactness_se         -3.861e+03  9.169e+05  -0.004    0.997
texture_se              1.407e+02  1.910e+04   0.007    0.994
concavity_se            6.339e+02  6.857e+05   0.001    0.999
fractal_dimension_mean  3.816e+04  6.512e+06   0.006    0.995
compactness_worst       2.889e+02  1.652e+05   0.002    0.999
smoothness_mean         1.260e+03  1.523e+06   0.001    0.999
fractal_dimension_se   -7.959e+04  6.588e+06  -0.012    0.990
radius_worst            9.473e+01  5.185e+04   0.002    0.999
smoothness_se           2.209e+04  2.714e+06   0.008    0.994
concave_points_se       4.176e+04  2.699e+06   0.015    0.988
symmetry_worst          1.168e+03  2.295e+05   0.005    0.996
symmetry_se            -1.679e+04  2.019e+06  -0.008    0.993
symmetry_mean           4.162e+01  2.426e+05   0.000    1.000

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 3.8969e+02  on 299  degrees of freedom
Residual deviance: 9.0991e-07  on 279  degrees of freedom
AIC: 42

Number of Fisher Scoring iterations: 25
```

```
glm(formula = diagnosis ~ compactness_mean + area_worst + texture_mean +
    radius_mean + smoothness_mean, family = binomial, data = mydata)

Deviance Residuals:
     Min        1Q    Median        3Q       Max
-1.58779  -0.13766  -0.03020   0.00156   2.83641

Coefficients:
                 Estimate Std. Error z value Pr(>|z|)
(Intercept)      -23.24195    8.83998  -2.629 0.008559 **
compactness_mean   3.52227   11.86192   0.297 0.766513
area_worst         0.02376    0.00646   3.678 0.000236 ***
texture_mean       0.39076    0.10271   3.805 0.000142 ***
radius_mean       -1.28584    0.65561  -1.961 0.049845 *
smoothness_mean  139.54867   51.27777   2.721 0.006500 **
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 389.691  on 299  degrees of freedom
Residual deviance:  61.218  on 294  degrees of freedom
AIC: 73.218

Number of Fisher Scoring iterations: 9
```

In model 2 I propose a much more accurate level with more significance levels noted in the variables, however we must continue to eliminate variables that serve no significance. This is a *good* model, but not the best yet.

Pictured: Model 2

Model 3 is the best model I could find, and we will check the accuracy of all models in the next section. These variable all work independently at a high significance level to predict the risk of a tumor.

```
Deviance Residuals:
    Min       1Q   Median       3Q      Max
-1.4218  -0.1083  -0.0198   0.0043   3.2405

Coefficients:
                  Estimate Std. Error z value Pr(>|z|)
(Intercept)     -33.499530   6.436388  -5.205 1.94e-07 ***
area_worst        0.012269   0.002241   5.475 4.37e-08 ***
texture_mean      0.408958   0.099925   4.093 4.26e-05 ***
smoothness_mean 136.435837  39.342530   3.468 0.000525 ***
concavity_mean   16.530656   5.950295   2.778 0.005467 **
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 389.691  on 299  degrees of freedom
Residual deviance:  60.708  on 295  degrees of freedom
AIC: 70.708

Number of Fisher Scoring iterations: 9
```

<u>Analysis:</u>

We have found phenomenal accuracy with this model, false negatives and false positives compose less than 5% of trials. This accuracy level would be suited for a clinical model.

```
> pred_table

    FALSE TRUE
  B   189    5
  M     6  100
>
```

```
> Eva

                            [,1]
Overall Accuracy           0.963
Mis-classification Rate    0.037
Sensitivity                0.943
Specificity                0.974
```

What are the real variables that predict cancer?

That is a complex question that relies on the model that one chooses to use. In my model the most important indicators are area_worst, texture_mean, smoothness_mean, and concavity_mean. These indicators give an accurate idea of the tumor's danger, but a false negative --even if unlikely-- can be the difference between life and death. For this reason, we should always follow up a test with additional models and especially lab monitoring in order to reduce the chances that we give bad medical information.

The significance of these four factors can be seen by the significance codes, *** being the highest. In my model, it is clear that area_worst, texture_mean, smoothness_mean are the most important in the analysis, but those alone yielded an overall accuracy of .95 while when including concavity_mean (**), we get to .963, which in medicine is a very important accuracy.

Additional Points:

Predict the Breast Cancer diagnosis of the Z patient[1]:

To predict the cancer diagnosis of patient Z, we use the r function :

```
dataz <- read_xlsx("Z_patient.xlsx")
predict(model3, dataz, type = "response")
```

Where the patient's data has been entered into an excel file and bound to the variable dataz. The result using my best model is:

```
> dataz <- read_xlsx("Z_patient.xlsx")
> predict(model3, dataz, type = "response")
        1
0.9945483
```

In which case we classify the tumor as benign, with a high specificity from model3

Below is the script I used in R-studio

```
# julian politsch 1831710 biostats I, bioinformatics 2019/20

install.packages("corrplot")
install.packages("MASS")
install.packages("ISLR")
install.packages("matlib")

library(corrplot)
library(MASS)
library(ISLR)
library(matlib)
library(Matrix)


data <- read.table(file = "cancer_data.txt", header = TRUE)
dim(data)
str(data)
data[,31] <- as.factor(data[,31])
data[,31]

set.seed(1831710)
idx <- sample(x = (1:dim(data)[1]), size = 300, replace = FALSE)
mydata <- data[idx,]
colnames(mydata) <- colnames(data)
dim(mydata)
summary(mydata)
attach(mydata)


# Logistic regression model ------------------------------------------------

par(mfrow=c(2.5,3))
for (i in 1:30){
  boxplot(mydata[,i] ~ diagnosis , lwd = .5, col = "light green", ylab = colnames(mydata[i]))}

modelfull <- glm(formula = diagnosis ~., family = binomial, data = mydata)
summary(modelfull)

model1 <- glm(formula = diagnosis ~ compactness_mean + area_worst + texture_mean + radius_mean +
perimeter_mean + area_mean + concave_points_worst + compactness_se + texture_se + concavity_se +
fractal_dimension_mean + compactness_worst + smoothness_mean + fractal_dimension_se + radius_worst +
smoothness_se + concave_points_se + symmetry_worst + symmetry_se + symmetry_mean , family = binomial, data
= mydata)
summary(model1)

model2 <- glm(formula = diagnosis ~ compactness_mean + area_worst + texture_mean + radius_mean +
smoothness_mean , family = binomial, data = mydata)
summary(model2)

model3 <- glm(formula = diagnosis ~ area_worst + texture_mean + smoothness_mean + concavity_mean, family =
binomial, data = mydata)
summary(model3)


# Assessing the accuracy of the model -------------------------------------

predict <- model3$fitted.values

pred_table <- table(mydata$diagnosis, predict > 0.5)

Acc <- sum(diag(pred_table))/sum(pred_table)
Mis <- 1-Acc

Sen <- pred_table[2,2]/sum(pred_table[2,])
Spe <- pred_table[1,1]/sum(pred_table[1,])

Eva <- round(matrix(data = c(Acc, Mis, Sen, Spe), nrow = 4, ncol = 1),3)
row.names(Eva) <- c("Overall Accuracy", "Mis-classification Rate", "Sensitivity","Specificity")
```