Benjamin Bian, Ashmeet Chhabra, Samuel Li

Mr. Andrews

Data Science

03 April 2024

Breast Cancer Analysis

Introduction

For this analysis we will be analyzing a dataset containing information about breast cancer cell nuclei extracted through the technique of fine needle aspiration, or FNA. All patients were examined in Wisconsin during the year 1993. The question we will be attempting to answer follows: what combination of quantitative variables produces the best model for predicting whether a breast cancer nuclei of a patient in Wisconsin is malignant or benign?

Dataset

The dataset is from the UI Machine Learning Repo, and context of each FNA sample's corresponding patient is provided above. The dataset collects ten parameters, and records the mean of each one. The parameters are as follows. If there is a [] that means that the definition is provided by us, otherwise it is the definition given by the repo.

- success (binary variable - Malignant: 1, Benign: 0)
- radius (mean of distances from center to points on the perimeter)
- texture (standard deviation of gray-scale values)
- perimeter [self explanatory]
- area [self explanatory]

- smoothness	(local variation in radius lengths)
- compactness	(perimeter^2 / area - 1.0)
- concavity	(severity of concave portions of the contour)
- concave points	(number of concave portions of the contour)
- symmetry	[no explanation given]
- fractal dimension	(“coastline approximation” - 1)

While normally, we would advance forward with all these variables, there’s a problem with this plan. This is because the quantities radius, perimeter, area, and concavity are all interdependent, and thus there could be significant interaction among these terms. To reduce this, it would be optimal to select the “best” factor out of these four quantities. This will be done through the model making process.

Conditions (Part 1)

The patients in the data were randomly sampled out of cancer patients in Wisconsin in 1993. Independence is a little bit more tricky, because cancer is a hereditary disease. However, because the sample size is relatively small, we can neglect this and assume independence. There is a final condition, linearity, but that will be checked following our selection of the best model. This is because to check the linearity condition we have to analyze each predictor individually rather than the entire model holistically, so it would be best to have N predictors chosen out of the set of 10 variables so we don’t need to check all 10 (hopefully).

Finding the Model

The first step in making the model is to determine which one of the four disputed predictors - radius, perimeter, area, compactness - we should use in addition to the other six predictors. This is done with the following.

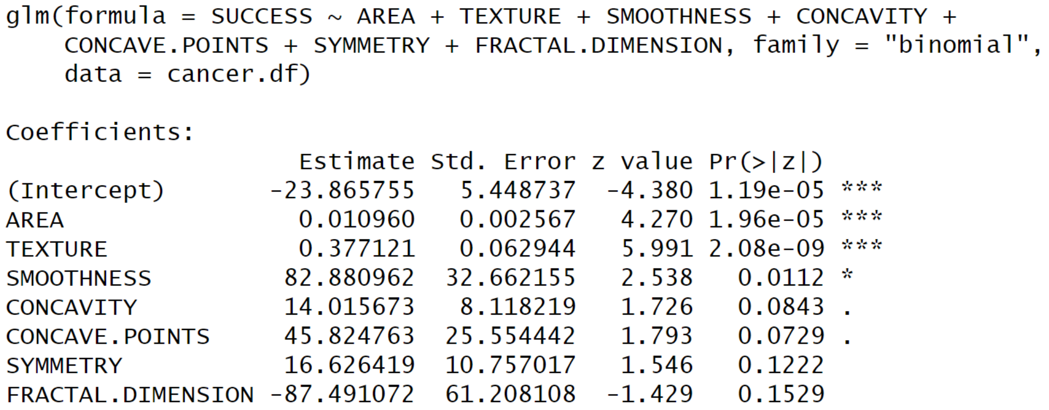
Success ~ Radius + Texture + Smoothness + Concavity + Concave.Points + Fractal.Dimension + Symmetry
Null deviance: 751.44 on 568 degrees of freedom
Residual deviance: 153.35 on 561 degrees of freedom

Success ~ Area + Texture + Smoothness + Concavity + Concave.Points + Fractal.Dimension + Symmetry
Null deviance: 751.44 on 568 degrees of freedom
Residual deviance: 150.19 on 561 degrees of freedom

Success ~ Perimeter + Texture + Smoothness + Concavity + Concave.Points + Fractal.Dimension + Symmetry
Null deviance: 751.44 on 568 degrees of freedom
Residual deviance: 154.485 on 561 degrees of freedom

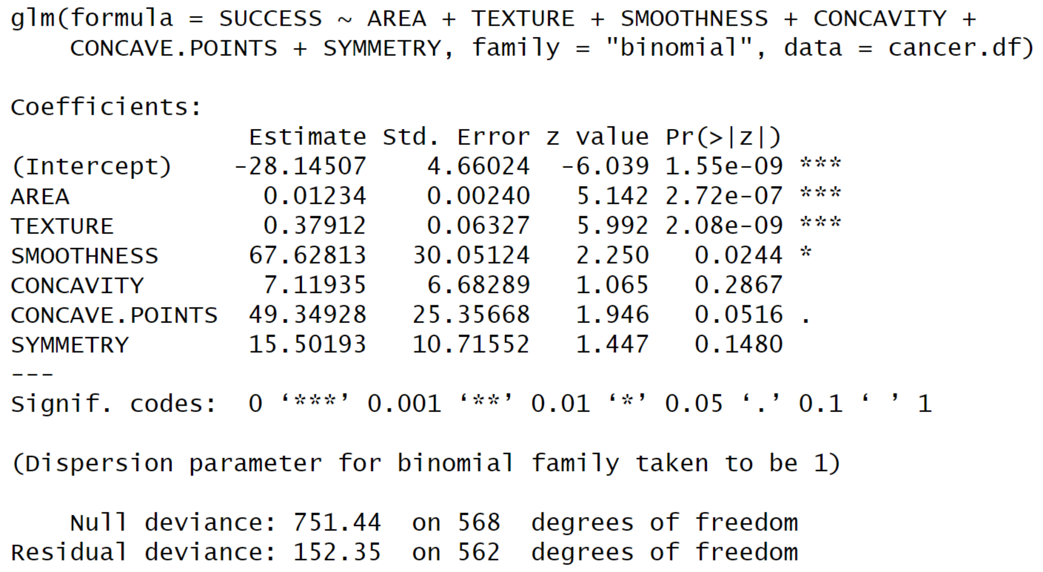
Success ~ Compactness + Texture + Smoothness + Concavity + Concave.Points + Fractal.Dimension + Symmetry
Null deviance: 751.44 on 568 degrees of freedom
Residual deviance: 173.6 on 561 degrees of freedom

Out of all these “full” models, AREA produces the lowest residual deviance, meaning that in this full form, AREA produces the model with the best fit. Now, we will see if all these predictors are really necessary with the backwards elimination procedure.



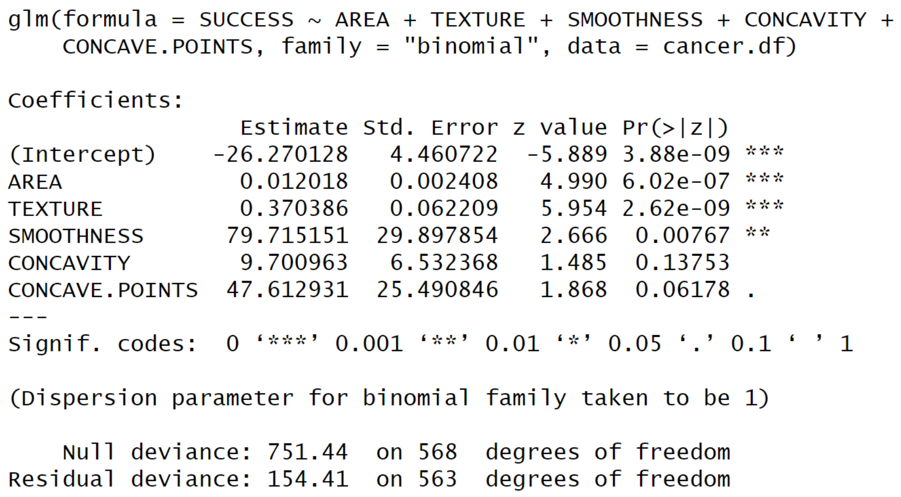
```
glm(formula = SUCCESS ~ AREA + TEXTURE + SMOOTHNESS + CONCAVITY +
    CONCAVE.POINTS + SYMMETRY + FRACTAL.DIMENSION, family = "binomial",
    data = cancer.df)

Coefficients:
                    Estimate Std. Error z value Pr(>|z|)
(Intercept)       -23.865755   5.448737  -4.380 1.19e-05 ***
AREA                0.010960   0.002567   4.270 1.96e-05 ***
TEXTURE             0.377121   0.062944   5.991 2.08e-09 ***
SMOOTHNESS         82.880962  32.662155   2.538   0.0112 *
CONCAVITY          14.015673   8.118219   1.726   0.0843 .
CONCAVE.POINTS     45.824763  25.554442   1.793   0.0729 .
SYMMETRY           16.626419  10.757017   1.546   0.1222
FRACTAL.DIMENSION -87.491072  61.208108  -1.429   0.1529
```

From the model above with the AREA predictor, FRACTAL.DIMENSION appears to be the least significant predictor. We will remove it and do a nested G test to evaluate whether or not the presence of this predictor causes a significant improvement in the model.



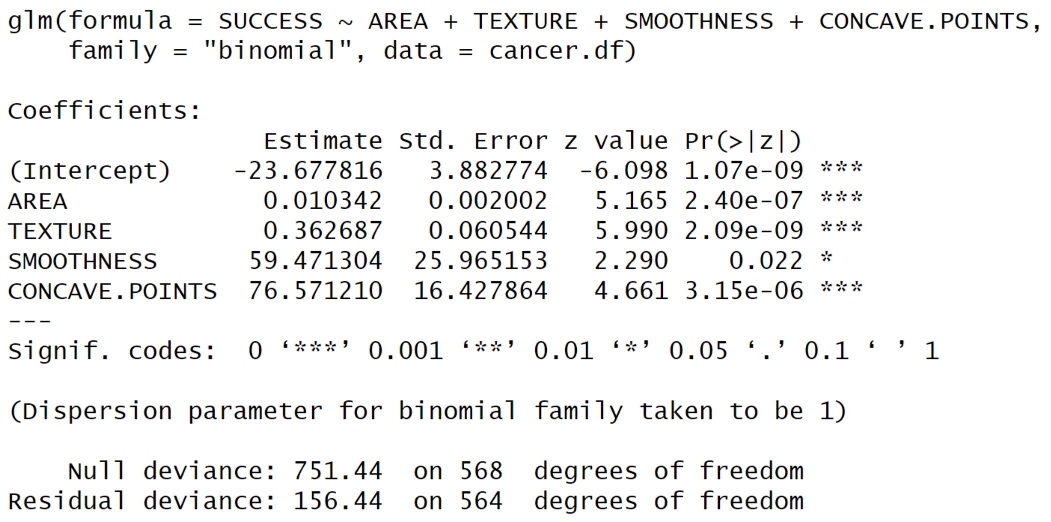
```
glm(formula = SUCCESS ~ AREA + TEXTURE + SMOOTHNESS + CONCAVITY +
    CONCAVE.POINTS + SYMMETRY, family = "binomial", data = cancer.df)

Coefficients:
                Estimate Std. Error z value Pr(>|z|)    
(Intercept)    -28.14507    4.66024  -6.039 1.55e-09 ***
AREA             0.01234    0.00240   5.142 2.72e-07 ***
TEXTURE          0.37912    0.06327   5.992 2.08e-09 ***
SMOOTHNESS      67.62813   30.05124   2.250   0.0244 *  
CONCAVITY        7.11935    6.68289   1.065   0.2867    
CONCAVE.POINTS  49.34928   25.35668   1.946   0.0516 .  
SYMMETRY        15.50193   10.71552   1.447   0.1480    
---
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: 152.35  on 562  degrees of freedom
```

The residual deviance went up by 152.35 - 150.19 = 2.16, which is to be expected, as adding predictors will always decrease the g value. What’s more important is whether or not this drop is significant. Plugging this into a chi squared distribution yields a p value of 0.14165, which is not a significant increase in model efficacy. Thus, we can discard the variable FRACTAL.DIMENSION. The next least significant predictor is SYMMETRY.



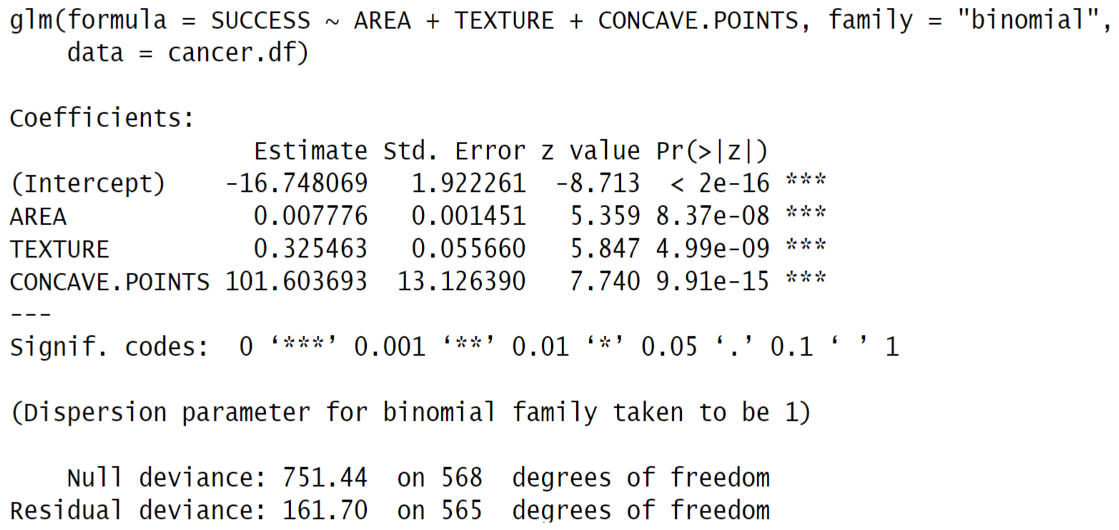
```
glm(formula = SUCCESS ~ AREA + TEXTURE + SMOOTHNESS + CONCAVITY +
    CONCAVE.POINTS, family = "binomial", data = cancer.df)

Coefficients:
                 Estimate Std. Error z value Pr(>|z|)    
(Intercept)    -26.270128   4.460722  -5.889 3.88e-09 ***
AREA             0.012018   0.002408   4.990 6.02e-07 ***
TEXTURE          0.370386   0.062209   5.954 2.62e-09 ***
SMOOTHNESS      79.715151  29.897854   2.666  0.00767 ** 
CONCAVITY        9.700963   6.532368   1.485  0.13753    
CONCAVE.POINTS  47.612931  25.490846   1.868  0.06178 .  
---
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: 154.41  on 563  degrees of freedom
```

The p value for the nested g test on SYMMETRY yields a value of 0.15121, so adding this predictor to the model does not improve the model by a significant amount. Repeating this for CONCAVITY:



```
glm(formula = SUCCESS ~ AREA + TEXTURE + SMOOTHNESS + CONCAVE.POINTS,
    family = "binomial", data = cancer.df)

Coefficients:
                 Estimate Std. Error z value Pr(>|z|)    
(Intercept)    -23.677816   3.882774  -6.098 1.07e-09 ***
AREA             0.010342   0.002002   5.165 2.40e-07 ***
TEXTURE          0.362687   0.060544   5.990 2.09e-09 ***
SMOOTHNESS      59.471304  25.965153   2.290    0.022 *  
CONCAVE.POINTS  76.571210  16.427864   4.661 3.15e-06 ***
---
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: 156.44  on 564  degrees of freedom
```

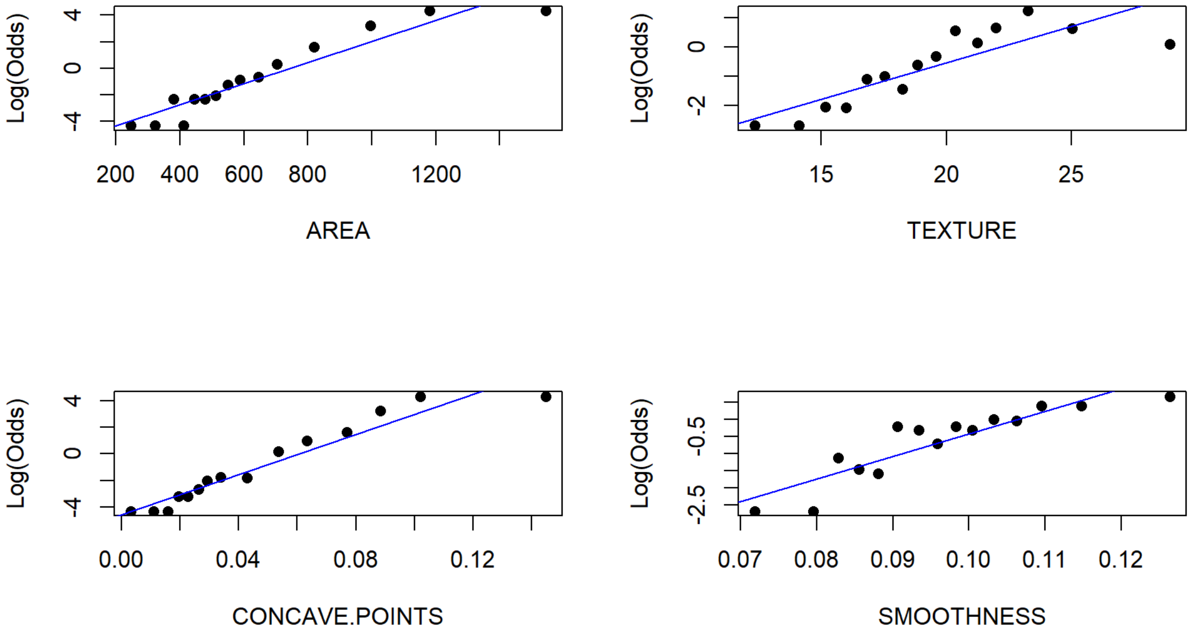
Yields a similar p value - 0.154221 - as the last few repetitions. However, now, a crucial distinction emerges. This is because now all the predictors are significant based on the Wald z test. This was not true for the last few iterations, so we might be able to stop this process now. Let’s try it for SMOOTHNESS.



```
glm(formula = SUCCESS ~ AREA + TEXTURE + CONCAVE.POINTS, family = "binomial",
    data = cancer.df)

Coefficients:
                 Estimate Std. Error z value Pr(>|z|)    
(Intercept)    -16.748069   1.922261  -8.713  < 2e-16 ***
AREA             0.007776   0.001451   5.359 8.37e-08 ***
TEXTURE          0.325463   0.055660   5.847 4.99e-09 ***
CONCAVE.POINTS 101.603693  13.126390   7.740 9.91e-15 ***
---
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
```

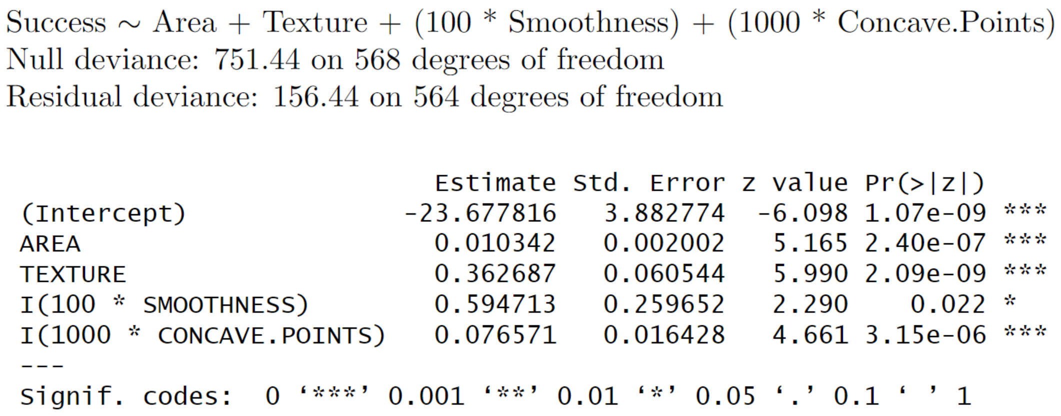
Our initial speculations are verified now: the g value has doubled in magnitude when compared to previous iterations, and the p value is now 0.021821, well under the significance level. This means that the presence of SMOOTHNESS in the model causes a significant increase in model efficacy. Therefore, the final model uses AREA, TEXTURE, CONCAVE.POINTS, and SMOOTHNESS to predict the log odds of whether or not a FNA tumor sample is malignant or benign.

Conditions (Part 2 Linearity)



All of the graphs show that data points are scattered roughly evenly around a line of best fit with equal variance throughout the x range and no significant outliers. Therefore, the linearity condition is met.

Final Model



Success ~ Area + Texture + (100 * Smoothness) + (1000 * Concave.Points)
Null deviance: 751.44 on 568 degrees of freedom
Residual deviance: 156.44 on 564 degrees of freedom

```
                          Estimate Std. Error z value Pr(>|z|)    
(Intercept)             -23.677816   3.882774  -6.098 1.07e-09 ***
AREA                      0.010342   0.002002   5.165 2.40e-07 ***
TEXTURE                   0.362687   0.060544   5.990 2.09e-09 ***
I(100 * SMOOTHNESS)       0.594713   0.259652   2.290    0.022 *  
I(1000 * CONCAVE.POINTS)  0.076571   0.016428   4.661 3.15e-06 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

Our final model includes the predictors AREA, TEXTURE,SMOOTHNESS, and CONCAVE.POINTS to predict log(odds) that a nucleus is malignant. The transformation of multiplying all values in CONCAVE.POINTS by 1000 has no effect other than reducing the slope to conform to the other slopes better, the value does not change, only the unit. This is the same with the 100 times multiplier on SMOOTHNESS.

log(odds) = -23.677816 + 0.010342(AREA) + 0.362687(TEXTURE) + .59471304(100*SMOOTHNESS) + .076571210(1000*CONCAVE.POINTS)

For an increase in AREA by one, the log(odds) of a given tumor being malignant increase by .010342

For an increase in TEXTURE by one, the log(odds) of a given tumor being malignant increase by .362687

For an increase in SMOOTHNESS by one, the log(odds) of a given tumor being malignant increase by .59471304

For an increase in CONCAVE.POINTS by one, the log(odds) of a given tumor being malignant increase by .076571210

In this model, all four predictors have significant p-values at every common significance level.

The predicted probability of Success can be modeled by the following equation.

$$\pi = \frac{e^{-23.677816+0.010342(AREA)+0.362687(TEXTURE)+.59471304(100*SMOOTHNESS)+.076571210(1000*CONCAVE.POINTS)}}{1+e^{-23.677816+0.010342(AREA)+0.362687(TEXTURE)+.59471304(100*SMOOTHNESS)+.076571210(1000*CONCAVE.POINTS)}}$$

Multicollinearity

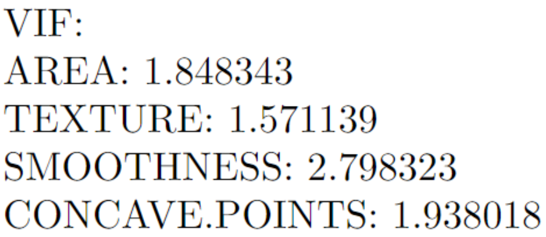
Variance Inflation Factor (VIF) is a measure of the strength of correlation between predictor variables in a model. It takes on a value between 1 and positive infinity.

VIF = 1: No correlation between predictors

1 < VIF< 5: moderate correlation but is fine

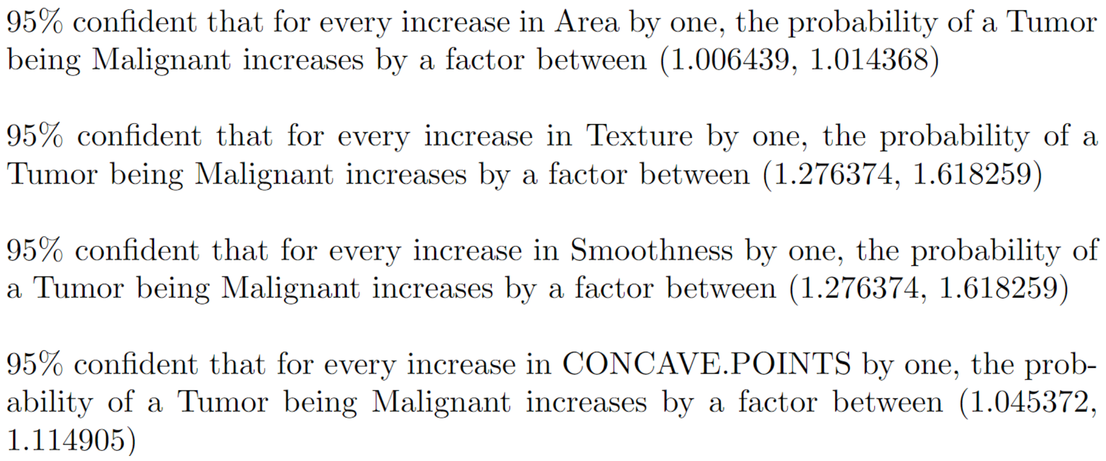
VIF> 5: strong correlation between predictors

Below is the VIF for each predictor respectively. Each VIF value lies between 1 and 5, so multicollinearity is negligible.



VIF:
AREA: 1.848343
TEXTURE: 1.571139
SMOOTHNESS: 2.798323
CONCAVE.POINTS: 1.938018

Confidence Intervals



95% confident that for every increase in Area by one, the probability of a Tumor being Malignant increases by a factor between (1.006439, 1.014368)

95% confident that for every increase in Texture by one, the probability of a Tumor being Malignant increases by a factor between (1.276374, 1.618259)

95% confident that for every increase in Smoothness by one, the probability of a Tumor being Malignant increases by a factor between (1.276374, 1.618259)

95% confident that for every increase in CONCAVE.POINTS by one, the probability of a Tumor being Malignant increases by a factor between (1.045372, 1.114905)

Conclusion

The original question we set out to answer is to find what combination of quantitative variables produces the best model for predicting whether a breast cancer nuclei of a patient in Wisconsin is malignant or benign. Through our analysis, we found that a four predictor model serves as the simplest model that predicts the Malignancy of a tumor with no significant difference than a full 8 predictor model. Our final model includes the variables Area, Texture, Smoothness, and Concave Points, which all have significant slopes in the logistic model. This model can be generalized to the population that this data was randomly sampled from to predict the probability that a given tumor is malignant if the required data is present as it met conditions for linearity, independence, and randomness.