SMU Data Science

Stats 6372

Final Project

Arnold Zhang [arnoldz@smu.edu](mailto:arnoldz@smu.edu)

Jeremy O. [jotsap@smu.edu](mailto:jotsap@smu.edu)

Tej Tenmattam [ttenmattam@smu.edu](mailto:ttenmattam@smu.edu)

Contents

[Data Description 3](#_Toc6418092)

[Overview of the Data Set 3](#_Toc6418093)

[Exploratory Data Analysis 3](#_Toc6418094)

[Data Assumption Checks 4](#_Toc6418095)

[Analysis of Data Correlation 4](#_Toc6418096)

[Overall Sample Behavior 6](#_Toc6418097)

[Principal Component Analysis 8](#_Toc6418098)

[Primary Logistic Regression 10](#_Toc6418099)

[APPENDIX 10](#_Toc6418100)

# Data Description

## Overview of the Data Set

The data is a collection of the Breast Cancer Diagnostic Data Set collected by the University of Wisconsin-Madison. 569 observations were taken using digitized imaging of breast mass biopsy samples taken between 1989 and 1991. Biopsies were acquired using FNA, Fine Needle Aspiration

The data set was taken from UC Irvine’s Machine Learning repository.

<https://archive.ics.uci.edu/ml/datasets/Breast+Cancer+Wisconsin+%28Diagnostic%29>

Below is a description of the data set

DEPENDENT VARIABLE

Diagnosis: Binomial – B [benign] or M [malignant]

INDEPENDENT VARIABLES

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

a) radius (distance 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)

For each of these metrics the data set included

* Mean
* Standard Error
* Worst [for asymmetrical masses this is the “longest” or “largest” value]

# Exploratory Data Analysis

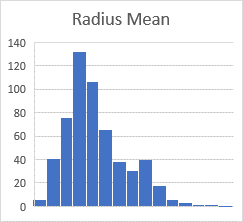
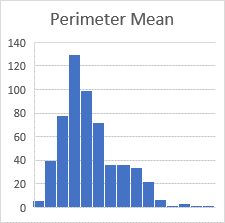
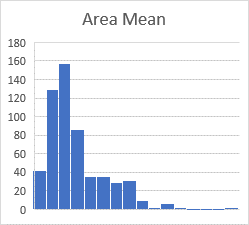
First validating assumptions of the data, then doing some visual analysis aided by a PCA to identify influential parameters

## Data Assumption Checks

**Sample Size**: The sample size is moderately sized with over 500 observations. Approximately 37.3% were Malignant and 62.7% were Benign.

**Independence**: Specific details as to patient selection were not available, so we will assume that independence were maintained. Points of possible concern would be a sample without proper diversity to represent the true population.

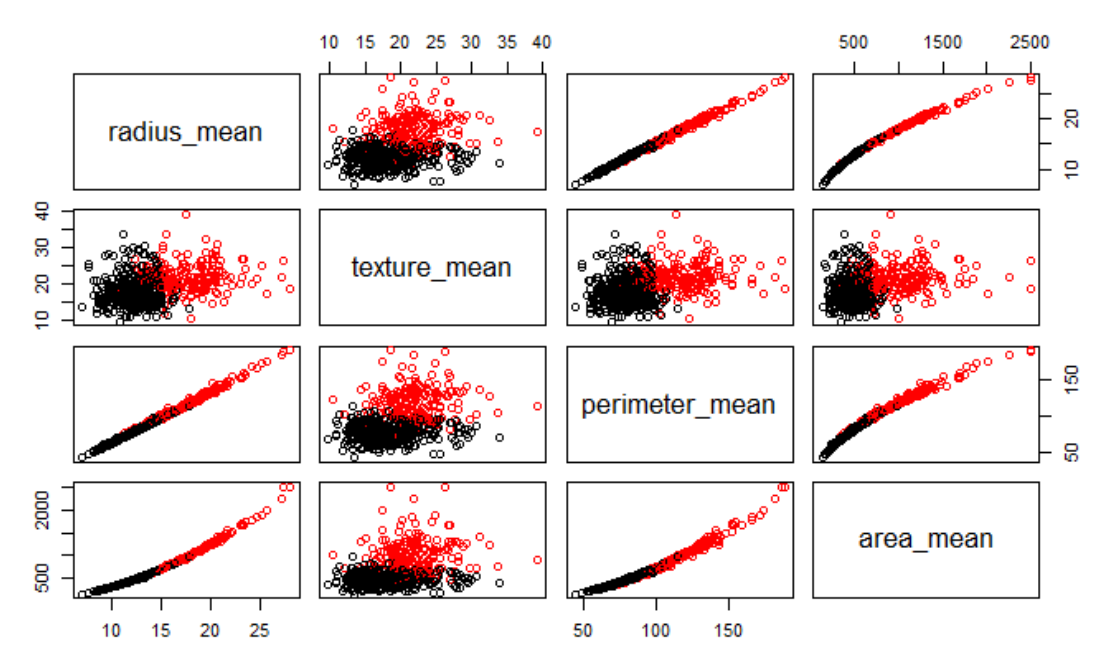
**Normality**: Normality assumptions were for the most part met. In cases where skew exists the CLT applied.

## Analysis of Data Correlation

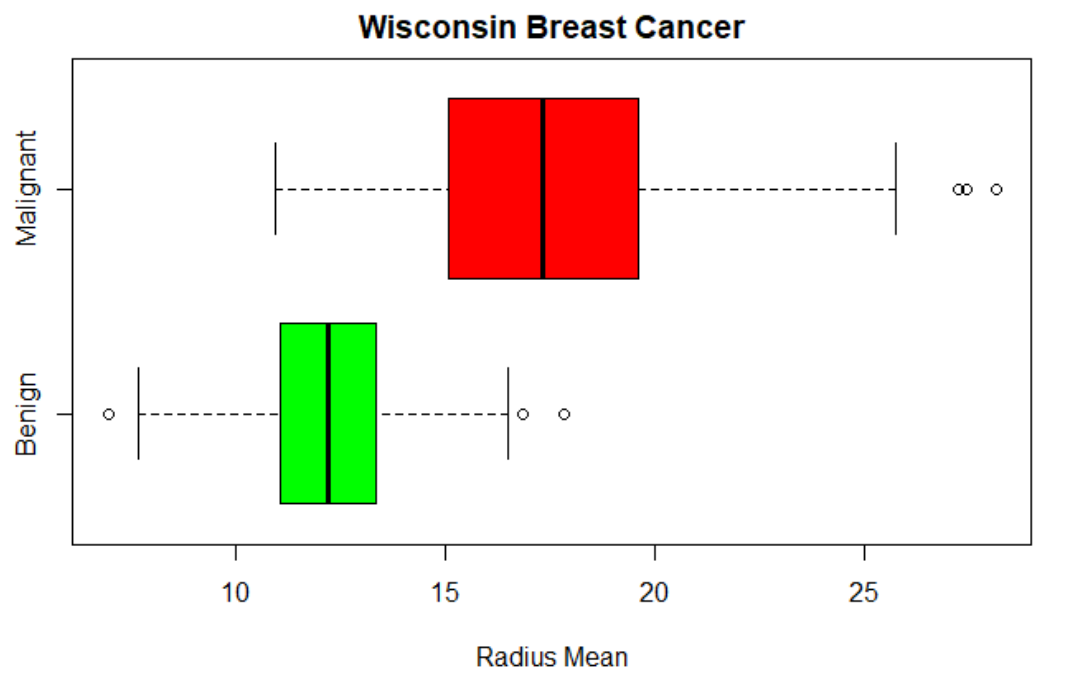
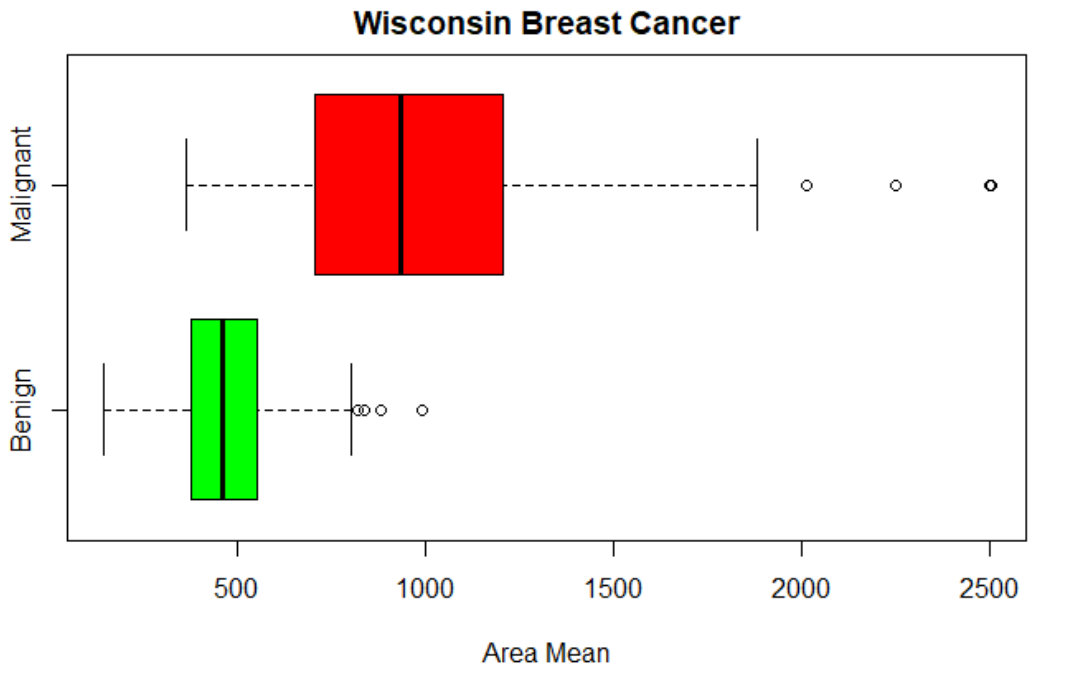
**Radius, Perimeter, Area**

These are obviously highly correlated since they are all a function of the radius and Pi. This is confirmed with a visual examination via scatter plot that shows a strong positive correlation present between size-based measurements of the tumorous mass and the “malignant” vs “benign” diagnosis.



The line chart below shows the progression of measurements, from smallest to largest, for both radius and area, separated by diagnosis. Over all the malignant biopsies appear to be larger overall.

A box plot comparison also validates this initial observation

## Overall Sample Behavior

Looking at the dashboard below, its apparent that despite the majority of samples taken being benign, the malignant samples overall make up more than half the area, and the variance of the malignant samples is almost 7 times higher.



When we select just the Benign subset, we can see across all measurements of variance, standard deviation, and standard error for measurements of the samples size [i.e. radius, perimeter, and area] are lower across the board.



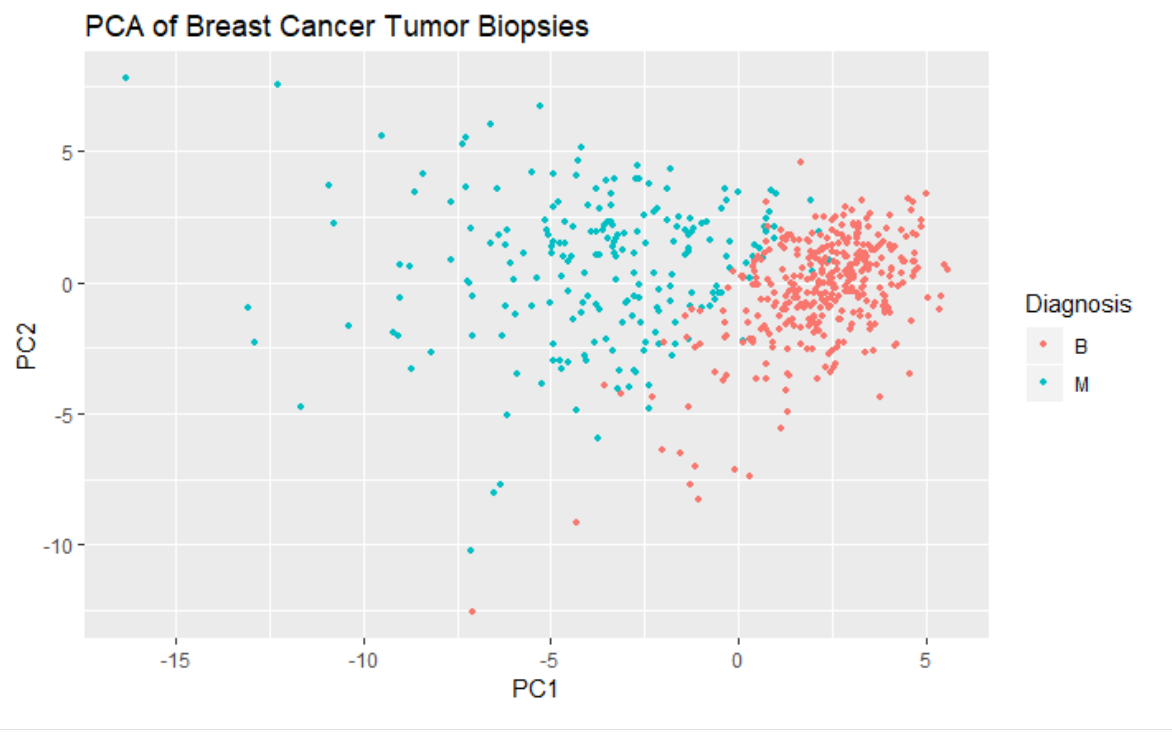
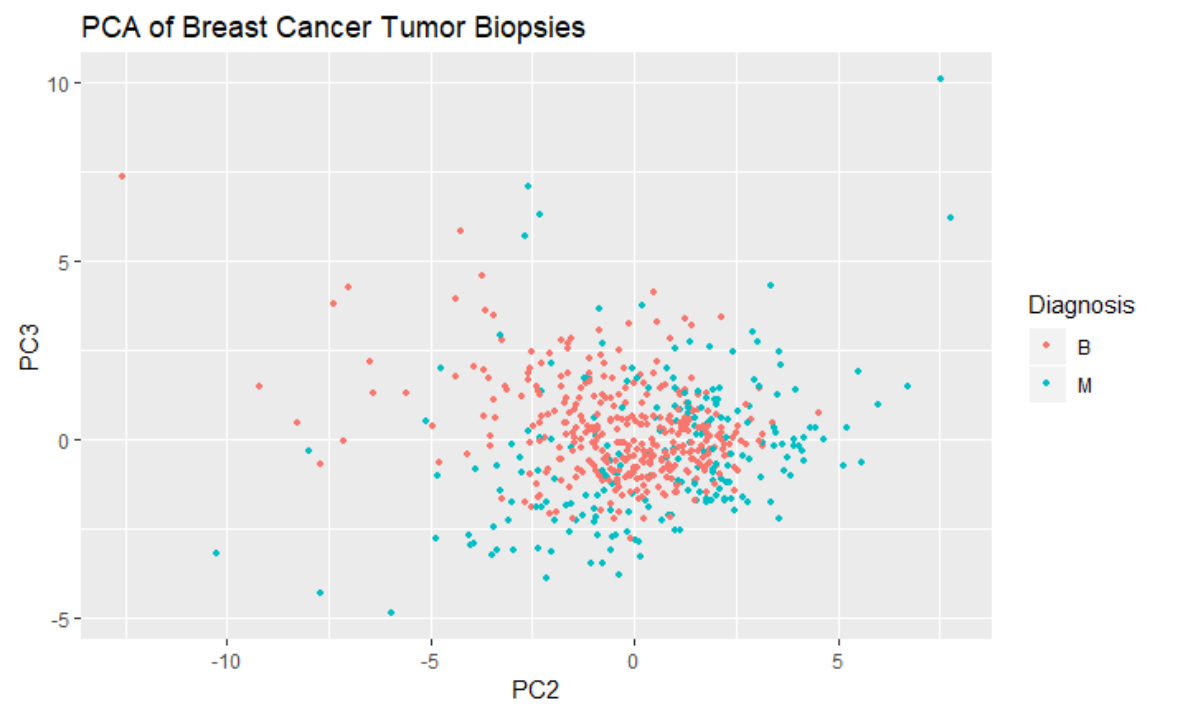
When selecting the Malignant sample, again the mean, standard error, and variance of size related measurements goes higher than the overall average of those categories.

Other qualities like smoothness, symmetry, and fractal do not appear to favor either Malignant or Benign diagnosis.

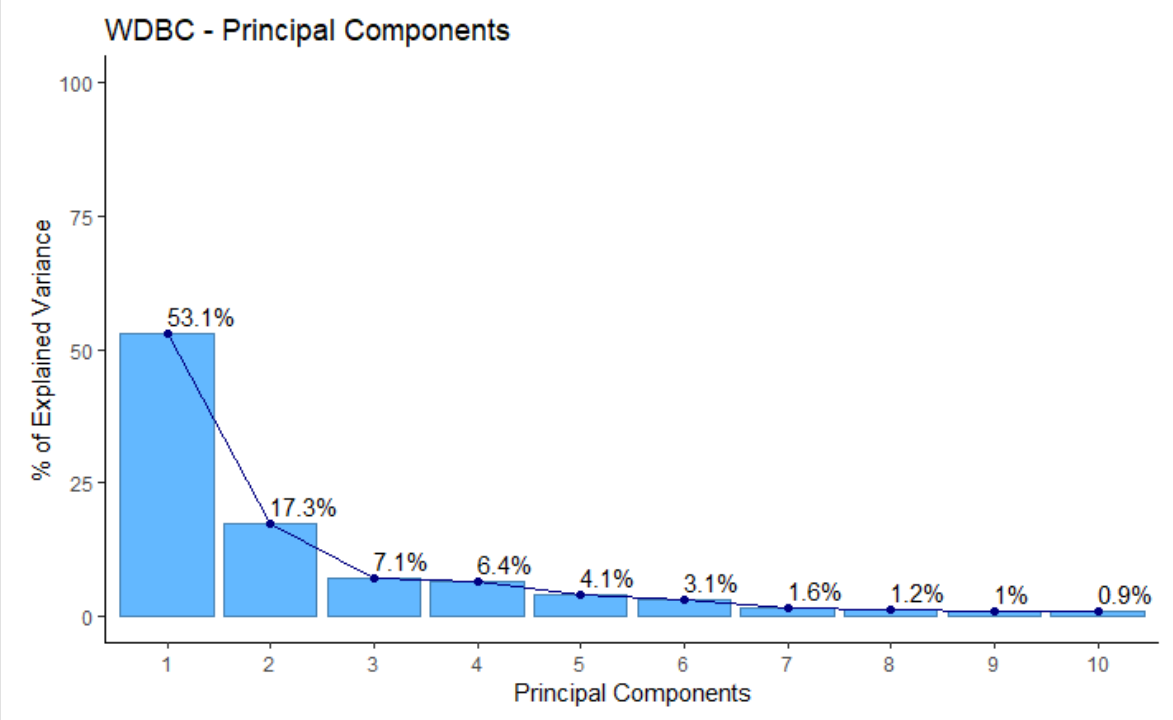


## Principal Component Analysis

Below we will conduct PCA on the predictors and plot the first few Principal Component’s [PC's], against each other and examine what level of separation they provide. The number of PCs to explore can be dictated by the scree plot. As part of the initial exploratory data analysis, this will allow us to see if there are clear patterns that immerge in the data.

We can see in the first graphic a clear separation exists between malignant and benign, which complements our initial visual examinations above. Because this clear separation exists in the PC's, this implies that a predictive model will probably do well.



**Importance of components:**

**Comp.1 Comp.2 Comp.3 Comp.4 Comp.5 Comp.6 Comp.7**

**Standard deviation** 0.5751101 0.3281171 0.21051526 0.19984110 0.15953874 0.13832097 0.09923315

**Proportion of Variance** 0.5309769 0.1728349 0.07114442 0.06411259 0.04086072 0.03071494 0.01580837

**Cumulative Proportion** 0.5309769 0.7038118 0.77495621 0.83906880 0.87992952 0.91064446 0.92645284

Looking at the performance of the PC’s we can see that 70% of the variance is explained by the first 2 components, and improvement significantly tapers off after that. This is further emphasized by the two scatterplots above showing significantly poorer separation using PC2 x PC2 compared to PC1 x PC2. We can see from this pairs plot of just the first few variables, separation between the malignant and benign groups are pretty well separated.

Given the above, we suspect an LDA analysis is appropriate to determine if the quantitative measurement parameters can be used to determine binomial categorical diagnosis response. However we will do multiple analysis models and compare the results.

# Primary Logistic Regression

Our initial model will be a logistic regression without modeling in the variable interactions. However the immediate concern is the covariance between multiple parameters. Running the full model through a VIF analysis shows significant covariance with most of the scores having VIF values well above 10.

radius\_mean texture\_mean perimeter\_mean area\_mean smoothness\_mean

3806.115296 11.884048 3786.400419 347.878657 8.194282

compactness\_mean concavity\_mean concave\_points\_mean symmetry\_mean fractal\_dimension\_mean

50.505168 70.767720 60.041733 4.220656 15.756977

radius\_se texture\_se perimeter\_se area\_se smoothness\_se

75.462027 4.205423 70.359695 41.163091 4.027923

compactness\_se concavity\_se concave\_points\_se symmetry\_se fractal\_dimension\_se

15.366324 15.694833 11.520796 5.175426 9.717987

radius\_worst texture\_worst perimeter\_worst area\_worst smoothness\_worst

799.105946 18.569966 405.023336 337.221924 10.923061

compactness\_worst concavity\_worst concave\_points\_worst symmetry\_worst fractal\_dimension\_worst

36.982755 31.970723 36.763714 9.520570 18.861533

This is not entirely suprising since:

1. Perimeter, Radius, and Area are all of a function of each other with respect to Pi
2. The other category of measurements are simply the standard error [i.e. variance] or the worst single measurement used to derive the

By simplifying the model to only use the Area measurements to represent the tumor size [which overall had the lower comparative VIF values than respective Radius and Perimeter measurements], and only keep the “mean” category of measurements, we were able to reduce the parameter VIF’s to values less than 10 across the board.

texture\_mean area\_mean smoothness\_mean compactness\_mean concavity\_mean

1.764829 3.250272 4.288467 5.763065 4.987014

concave\_points\_mean symmetry\_mean fractal\_dimension\_mean

5.744502 1.823926 8.311224

When comparing the two logistic regression models, the “full” model had a significantly worse AIC score than the simplified one.

OUTPUT FOR FULL LOGISTIC

glm(formula = diagnosis ~ ., family = binomial(link = "logit"),

data = bc.boolean, control = list(maxit = 50))

Deviance Residuals:

Min 1Q Median 3Q Max

-8.49 0.00 0.00 0.00 8.49

Coefficients:

Estimate Std. Error z value Pr(>|z|)

(Intercept) -1.704e+16 1.216e+08 -140095174 <2e-16 \*\*\*

radius\_mean -4.508e+15 4.930e+07 -91446497 <2e-16 \*\*\*

texture\_mean -2.681e+13 2.257e+06 -11879545 <2e-16 \*\*\*

perimeter\_mean 5.592e+14 7.131e+06 78428443 <2e-16 \*\*\*

area\_mean 6.374e+12 1.492e+05 42711010 <2e-16 \*\*\*

smoothness\_mean 4.804e+16 5.731e+08 83827859 <2e-16 \*\*\*

compactness\_mean -2.931e+16 3.789e+08 -77357485 <2e-16 \*\*\*

concavity\_mean 3.110e+14 2.971e+08 1046677 <2e-16 \*\*\*

concave\_points\_mean 8.171e+15 5.623e+08 14531119 <2e-16 \*\*\*

symmetry\_mean -8.564e+15 2.110e+08 -40582198 <2e-16 \*\*\*

fractal\_dimension\_mean -1.030e+16 1.583e+09 -6507058 <2e-16 \*\*\*

radius\_se 8.161e+15 8.821e+07 92524229 <2e-16 \*\*\*

texture\_se -4.617e+14 1.047e+07 -44105313 <2e-16 \*\*\*

perimeter\_se -8.989e+14 1.168e+07 -76951480 <2e-16 \*\*\*

area\_se 6.591e+12 3.971e+05 16595576 <2e-16 \*\*\*

smoothness\_se -1.037e+16 1.882e+09 -5509676 <2e-16 \*\*\*

compactness\_se 4.729e+16 6.164e+08 76728075 <2e-16 \*\*\*

concavity\_se -2.689e+16 3.696e+08 -72764771 <2e-16 \*\*\*

concave\_points\_se 1.993e+17 1.549e+09 128680606 <2e-16 \*\*\*

symmetry\_se -4.806e+16 7.749e+08 -62015895 <2e-16 \*\*\*

fractal\_dimension\_se -4.364e+17 3.317e+09 -131546226 <2e-16 \*\*\*

radius\_worst 1.102e+15 1.647e+07 66911477 <2e-16 \*\*\*

texture\_worst 1.390e+14 1.974e+06 70396825 <2e-16 \*\*\*

perimeter\_worst 6.421e+13 1.686e+06 38076691 <2e-16 \*\*\*

area\_worst -9.950e+12 9.082e+04 -109559685 <2e-16 \*\*\*

smoothness\_worst -6.379e+15 4.076e+08 -15649766 <2e-16 \*\*\*

compactness\_worst -7.493e+15 1.088e+08 -68848424 <2e-16 \*\*\*

concavity\_worst 6.282e+15 7.632e+07 82312806 <2e-16 \*\*\*

concave\_points\_worst -7.308e+15 2.597e+08 -28137342 <2e-16 \*\*\*

symmetry\_worst 1.098e+16 1.404e+08 78221084 <2e-16 \*\*\*

fractal\_dimension\_worst 4.755e+16 6.771e+08 70226947 <2e-16 \*\*\*

---

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: 792.96 on 538 degrees of freedom

**AIC: 854.96**

OUTPUT FOR SIMPLIFIED LOGISTIC

glm(formula = diagnosis ~ texture\_mean + area\_mean + smoothness\_mean +

compactness\_mean + concavity\_mean + concave\_points\_mean +

symmetry\_mean + fractal\_dimension\_mean, family = binomial(link = "logit"),

data = bc.boolean, control = list(maxit = 50))

Deviance Residuals:

Min 1Q Median 3Q Max

-2.02338 -0.14079 -0.03572 0.01120 3.00158

Coefficients:

Estimate Std. Error z value Pr(>|z|)

(Intercept) -27.253090 6.400270 -4.258 2.06e-05 \*\*\*

texture\_mean 0.384407 0.063485 6.055 1.40e-09 \*\*\*

area\_mean 0.011787 0.002728 4.320 1.56e-05 \*\*\*

smoothness\_mean 79.598004 32.979533 2.414 0.0158 \*

compactness\_mean -13.731887 12.700676 -1.081 0.2796

concavity\_mean 13.249936 8.090709 1.638 0.1015

concave\_points\_mean 57.230227 28.042055 2.041 0.0413 \*

symmetry\_mean 17.779963 10.854712 1.638 0.1014

fractal\_dimension\_mean -26.454327 82.469431 -0.321 0.7484

---

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: 149.01 on 560 degrees of freedom

**AIC: 167.01**

CONFIDENCE INTERVALS FOR PARAMETERS OF REDUCED MODEL

2.5 % 97.5 %

(Intercept) -4.053919e+01 -15.31563155

texture\_mean 2.674975e-01 0.51837370

area\_mean 6.757782e-03 0.01749949

smoothness\_mean 1.771586e+01 147.27093557

compactness\_mean -3.951821e+01 11.10875994

concavity\_mean -2.529860e+00 29.24616163

concave\_points\_mean 3.911450e+00 114.21970795

symmetry\_mean -3.644306e+00 39.25139134

fractal\_dimension\_mean -1.925487e+02 132.60062490

USING STEPWISE SELECTION

# APPENDIX

**Code: Loading & Cleaning Data**

bc<-read.table("https://archive.ics.uci.edu/ml/machine-learning-databases/breast-cancer-wisconsin/wdbc.data",header=F,sep=",")

names(bc)<- c('id\_number', 'diagnosis', 'radius\_mean',

'texture\_mean', 'perimeter\_mean', 'area\_mean',

'smoothness\_mean', 'compactness\_mean',

'concavity\_mean','concave\_points\_mean',

'symmetry\_mean', 'fractal\_dimension\_mean',

'radius\_se', 'texture\_se', 'perimeter\_se',

'area\_se', 'smoothness\_se', 'compactness\_se',

'concavity\_se', 'concave\_points\_se',

'symmetry\_se', 'fractal\_dimension\_se',

'radius\_worst', 'texture\_worst',

'perimeter\_worst', 'area\_worst',

'smoothness\_worst', 'compactness\_worst',

'concavity\_worst', 'concave\_points\_worst',

'symmetry\_worst', 'fractal\_dimension\_worst')

# Data Summary

summary(bc)

# Normalize Data

bc.clean <- bc[,-c(1)]

normalize <- function(x){

return (( x - min(x))/(max(x) -min(x)))

}

bc.clean.normalized <- as.data.frame(

lapply(bc.clean[,2:31],normalize)

)

bc.clean.normalized <- cbind(

bc.clean[,1],

bc.clean.normalized

)

names(bc.clean.normalized)[1] <- "diagnosis"

summary(bc.clean.normalized)

**Code: Box Plot Analysis**

#Box Plot: Area Mean

boxplot(area\_mean ~ diagnosis,data=wdbc,

horizontal=TRUE,

names=c("Benign","Malignant"),

col=c("green","red"),

xlab="Area Mean", main="Wisconsin Breast Cancer")

#Box Plot: Radius Mean

boxplot(radius\_mean ~ diagnosis,data=wdbc,

horizontal=TRUE,

names=c("Benign","Malignant"),

col=c("green","red"),

xlab="Radius Mean",

main="Wisconsin Breast Cancer")

**Code: Principal Component Analysis**

pc.bc<-prcomp(bc[,-c(1,2)],scale.=TRUE)

pc.bc.scores<-pc.bc$x

#Adding the response column to the PC's data frame

pc.bc.scores<-data.frame(pc.bc.scores)

pc.bc.scores$Diagnosis<-bc$diagnosis

#Use ggplot2 to plot the first few pc's

library(ggplot2)

ggplot(data = pc.bc.scores, aes(x = PC1, y = PC2)) +

geom\_point(aes(col=Diagnosis), size=1)+

ggtitle("PCA of Breast Cancer Tumor Biopsies")

ggplot(data = pc.bc.scores, aes(x = PC2, y = PC3)) +

geom\_point(aes(col=Diagnosis), size=1)+

ggtitle("PCA of Breast Cancer Tumor Biopsies")

**Code: Logistic Regression Full & VIF**

main.glm <- glm(diagnosis ~ . , data=bc.clean, family = binomial(link = "logit") , control = list(maxit = 50))

summary(main.glm)

# VIF for covariance between Radius, Perimeter, Area

vif(main.glm) -> main.glm.vif

main.glm.vif

**Code: Logistic Regression Simplified & VIF**

# REDUCED model removing all "SE"" measurements, all "Worst"", and only using "Area" in place of 'perimeter' and 'radius'

redux.glm <- glm(diagnosis ~ texture\_mean + area\_mean + smoothness\_mean + compactness\_mean + concavity\_mean + concave\_points\_mean + symmetry\_mean + fractal\_dimension\_mean , data=bc.clean, family = binomial(link = "logit") , control = list(maxit = 60))

summary(redux.glm)

# VIF for covariance between Radius, Perimeter, Area

vif(redux.glm) -> redux.glm.vif

redux.glm.vif

#95% CONFIDENCE INTERVALS

confint(redux.glm, level = 0.95)