Logistic\_regression\_breast\_cancer\_diagnosis

Thandeka Chaka and Tinashe Chisora

11/18/2021

# Set global figure size  
knitr::opts\_chunk$set(fig.width=6, fig.height=3.5)

## Data Characteristics Section

## Read in the data

Breastcancer = read.csv("/Users/vladimirechisora/Downloads/data-3.csv")  
  
attach(Breastcancer)  
  
ftable(diagnosis)

## diagnosis 0 1  
##   
## 357 212

dim (Breastcancer)

## [1] 569 12

###Abstract

The data that we are using is a sample from Mayo clinic. The data include a number of patients that went through some tests before the doctors discovered whether they had breast cancer or not. The data includes a number of measures on the patient’s breast tumor. Later on some of the patients were diagnosed with breast cancer which is labelled as M for malignant and for some patients it was just a breast tumor which is labeled as B for benign. Some of the questions that we will try to during this research are as follows:

1. What are the most significant predictor variables when it comes to predicting the diagnosis of breast cancer? This would help the doctors to make it clear on which things to look out for when carrying out the diagnosis.
2. What is the residual standard error for this data since the data is measuring very small quantities like the size of a cell? This might have a reflection on the methodology that has been used to collect the data. How accurate are the instruments used to record these predictor variables?
3. What is the prediction of the model in diagnosing cancer given the other predictor variables? We will try to create a model that can predict whether the tumor is cancerous or not and see how accurate the model is.

In our data, we have one response variable that is the diagnosis of breast cancer. In our data people who have been diagnosed with breast cancer have been labeled M and when it is just a tumor it is labelled B.

ftable(diagnosis)

## diagnosis 0 1  
##   
## 357 212

The table above shows the number of patient that have been diagnosed with breast cancer. 212 patients have been diagnosed with breast cancer while for the other 357, it was just a tumor.

We have 10 predictor variables variables that are different measurements of the cells on the breast tumor.

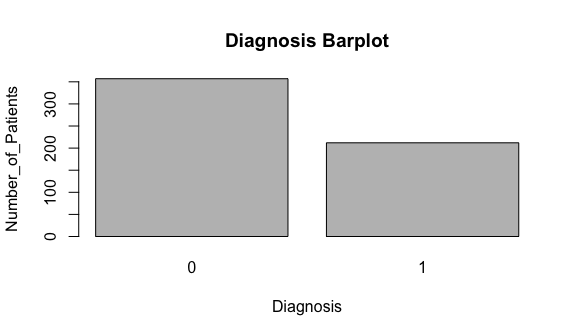
###Data Characteristics

##Variable characteristics

#Diagnosis

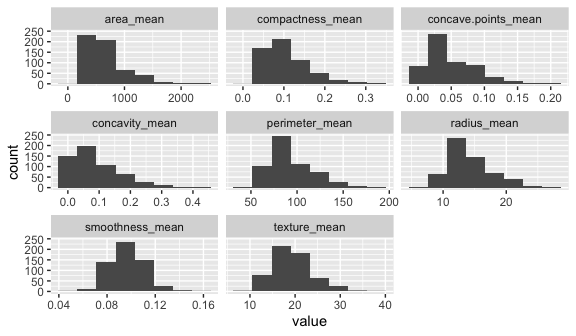
the diagnosis is non numeric. First of all we are going to make it binary that is 0 for Benign and 1 for M and then proceed to show the graphical representation.

barplot (table (diagnosis), xlab="Diagnosis", ylab="Number\_of\_Patients", main = "Diagnosis Barplot")



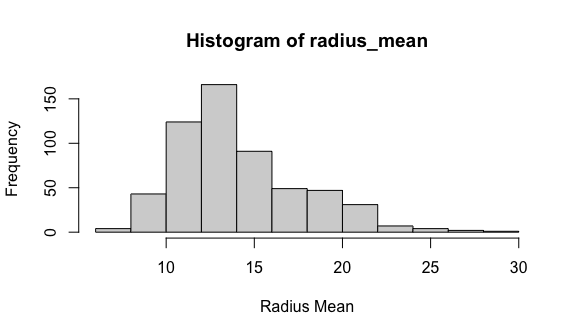
The plot above shows the distribution of patients who have been diagnosed with breast cancer and those one who have not.

library (ggplot2)  
library (tidyr)  
  
ggplot(gather(Breastcancer [, 3:10]), aes(value)) +   
 geom\_histogram(bins = 8) +   
 facet\_wrap(~key, scales = 'free\_x')

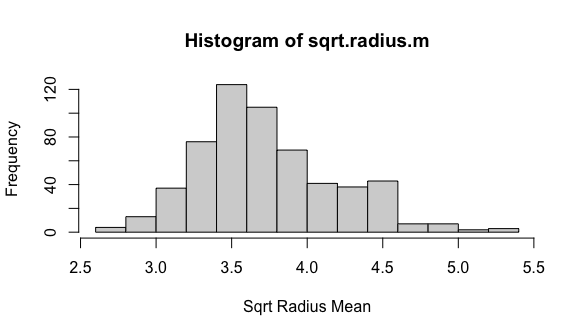


#Radius mean

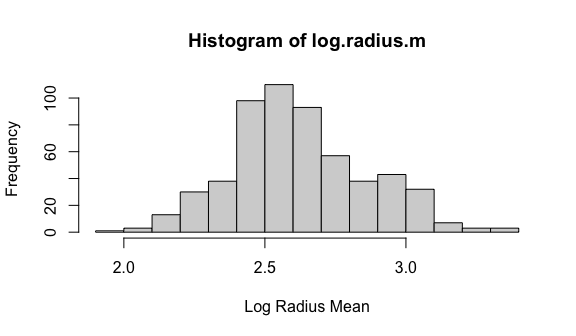
hist (radius\_mean, xlab="Radius Mean")



#transformation  
library ("MASS")  
  
#sqrt transformation  
sqrt.radius.m = sqrt(radius\_mean)  
hist (sqrt.radius.m, xlab="Sqrt Radius Mean")



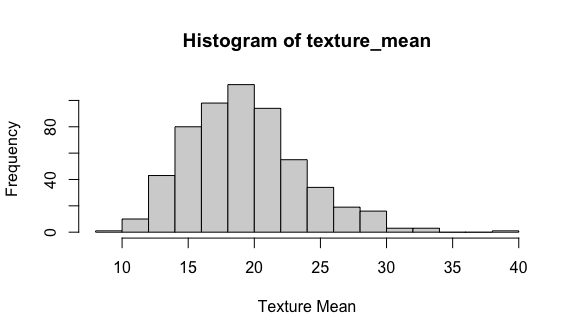
#log transformation  
log.radius.m = log(radius\_mean)  
hist (log.radius.m, xlab="Log Radius Mean")



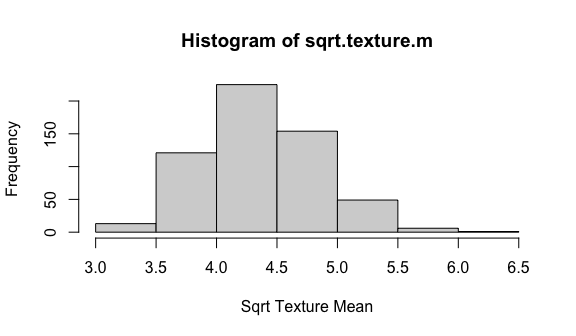
This is the graph representing the radius mean of all the patients. We tried a couple of transformations and the log transformation works better in this case.

#Texture Mean

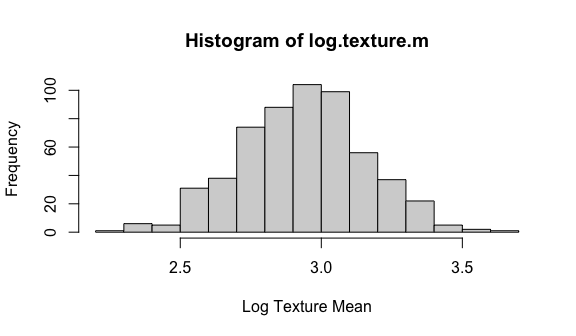
hist (texture\_mean, xlab="Texture Mean")



#transformation  
  
#sqrt transformation  
sqrt.texture.m = sqrt(texture\_mean)  
hist (sqrt.texture.m, xlab="Sqrt Texture Mean")



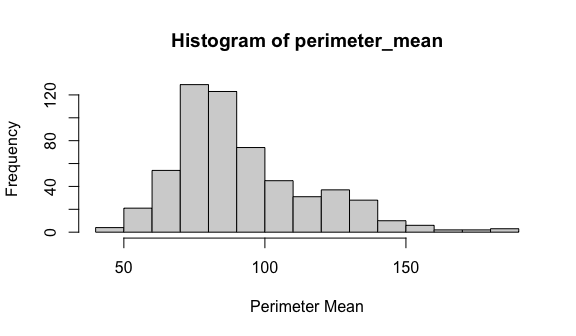
#log transformation  
log.texture.m = log(texture\_mean)  
hist (log.texture.m, xlab="Log Texture Mean")



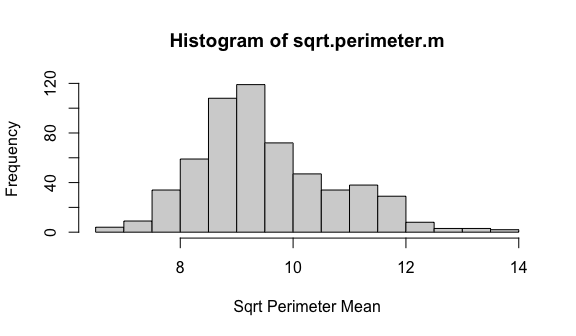
Our data is left skewed with one possible outlier. We used a square roottransformation which seemed to have worked better in dealing with the skewness and also got rid of the outlier.

#Perimeter Mean

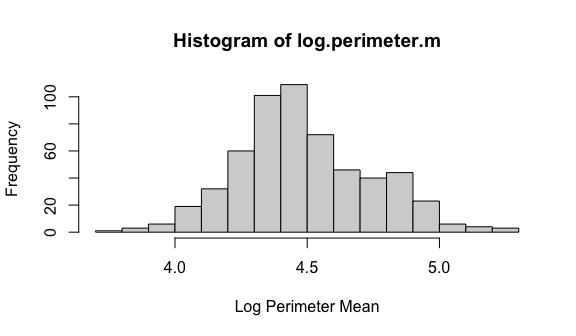
hist (perimeter\_mean, xlab="Perimeter Mean")



#transformation  
  
#sqrt transformation  
sqrt.perimeter.m = sqrt(perimeter\_mean)  
hist (sqrt.perimeter.m, xlab="Sqrt Perimeter Mean")



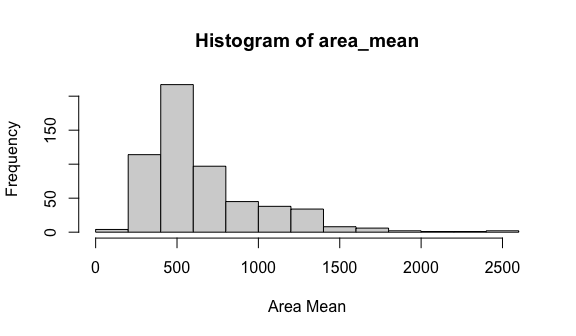
#log transformation  
log.perimeter.m = log(perimeter\_mean)  
hist (log.perimeter.m, xlab="Log Perimeter Mean")



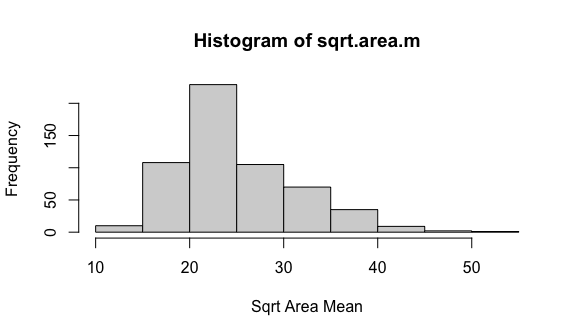
The perimeter mean is a liitle bit skewed to the left with no obvious outliers. We carried out a number of transformations and the log transformation is the best transformation we have at the moment.

#Area mean

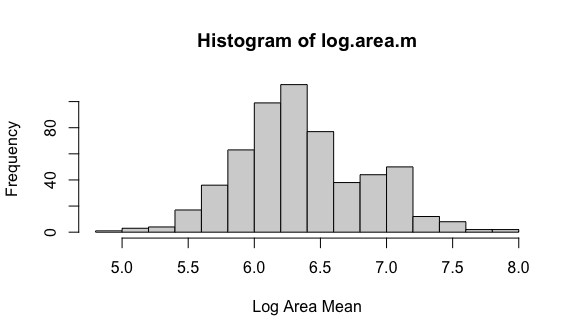
hist (area\_mean, xlab="Area Mean")



#transformation  
  
#sqrt transformation  
sqrt.area.m = sqrt(area\_mean)  
hist (sqrt.area.m, xlab="Sqrt Area Mean")



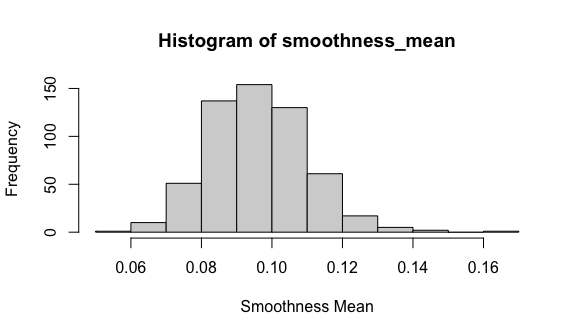
#log transformation  
log.area.m = log(area\_mean)  
hist (log.area.m, xlab="Log Area Mean")



The area mean was left skewed and we used a couple of transformations. The log one seems to be working better though it gives two peaks.

#Smoothness Mean

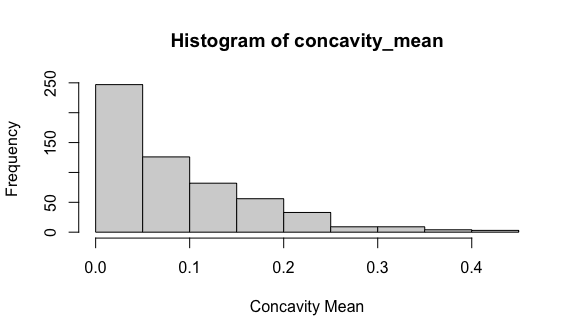
hist (smoothness\_mean, xlab="Smoothness Mean")



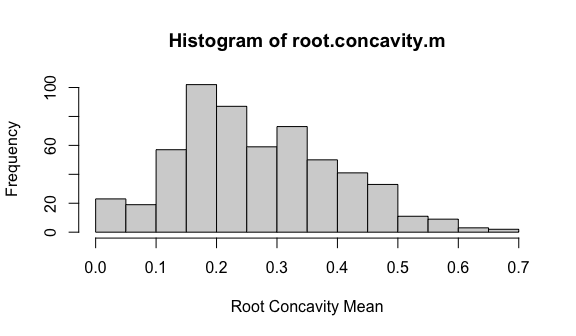
The smoothness mean data does not have an obvious skewness in it. There might be one possible outlier in the data set.

#Compactness Mean

hist (concavity\_mean, xlab="Concavity Mean")



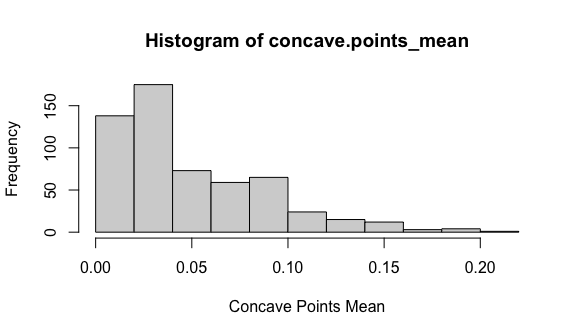
#transformation  
  
#sqrt transformation  
root.concavity.m = sqrt(concavity\_mean)  
hist (root.concavity.m, xlab="Root Concavity Mean")



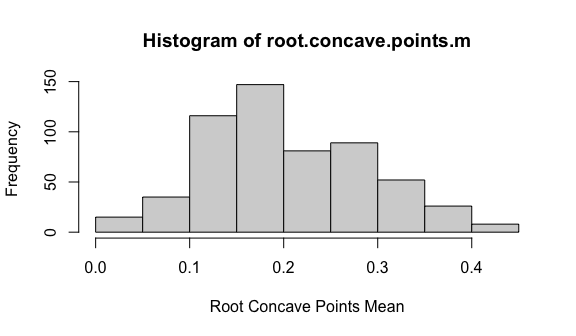
The compactness mean data is left skewed and we did a couple of transformations and the log transformation was our best option thought it also produced two peaks.

#Concavity Mean

hist (concave.points\_mean, xlab="Concave Points Mean")



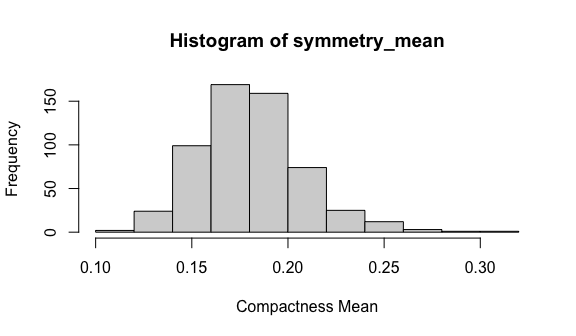
#transformation  
  
#sqrt transformation  
root.concave.points.m = sqrt(concave.points\_mean)  
hist (root.concave.points.m, xlab="Root Concave Points Mean")



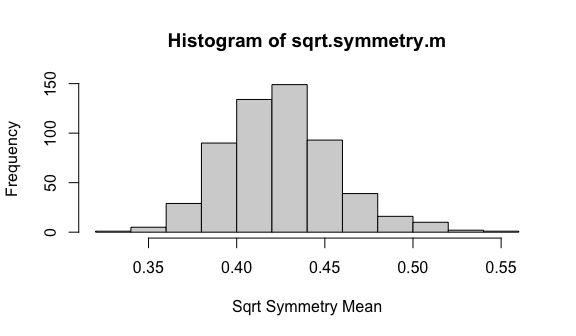
The initial data was obviously skewed to the left. The only transformation that worked better was transforming it to root 2.3.

#Symmetry mean

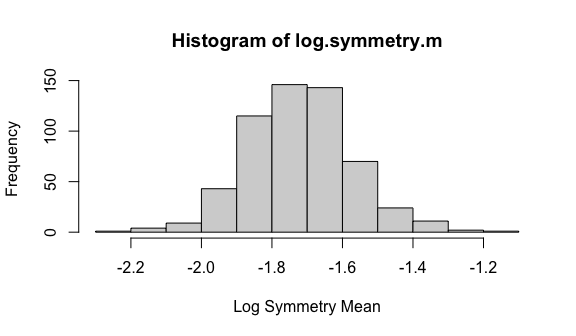
hist (symmetry\_mean, xlab="Compactness Mean")



#transformation  
  
#sqrt transformation  
sqrt.symmetry.m = sqrt(symmetry\_mean)  
hist (sqrt.symmetry.m, xlab="Sqrt Symmetry Mean")



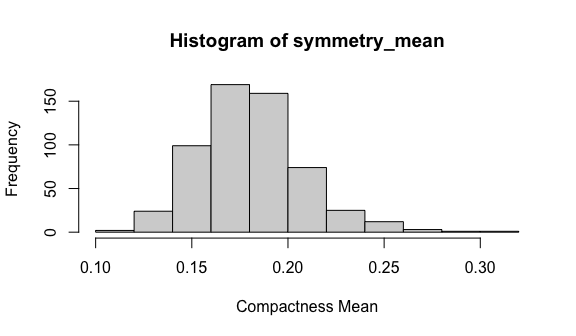
#log transformation  
log.symmetry.m = log(symmetry\_mean)  
hist (log.symmetry.m, xlab="Log Symmetry Mean")



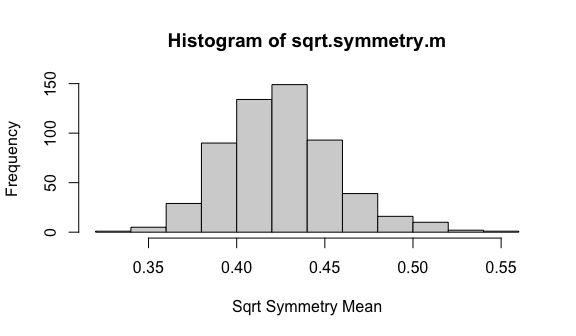
The initial data was left skewed and then we had to transform the variable by root 2.5. The variable now looks more normally distributed than what it looked before.

#Symmetry Mean

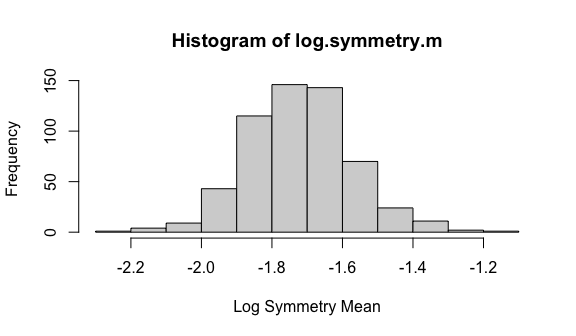
hist (symmetry\_mean, xlab="Compactness Mean")



#transformation  
  
#sqrt transformation  
sqrt.symmetry.m = sqrt(symmetry\_mean)  
hist (sqrt.symmetry.m, xlab="Sqrt Symmetry Mean")



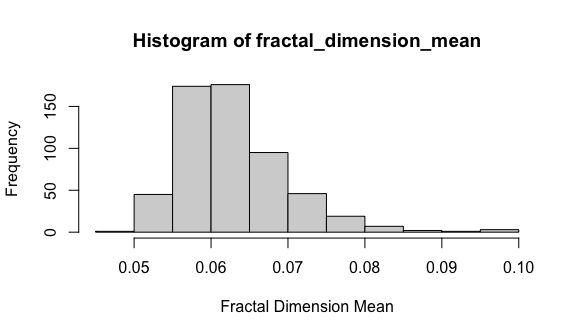
#log transformation  
log.symmetry.m = log(symmetry\_mean)  
hist (log.symmetry.m, xlab="Log Symmetry Mean")



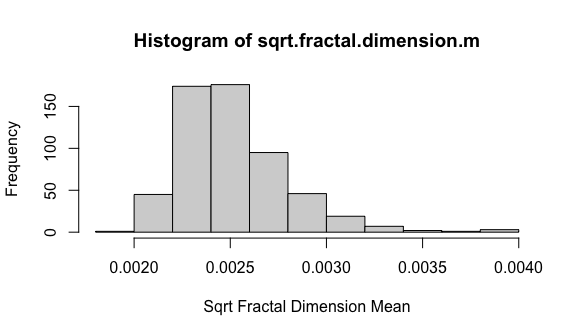
This variable also needed some log transformation in order to get rid of the left skewness.

#Fractal Dimension Mean

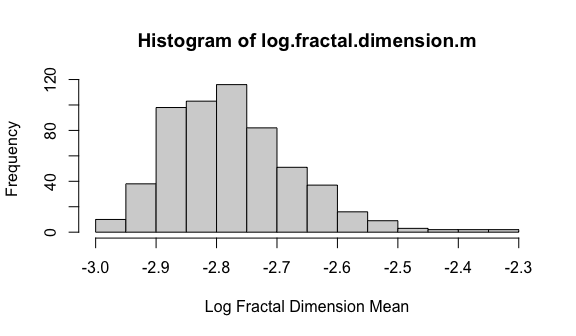
hist (fractal\_dimension\_mean, xlab="Fractal Dimension Mean")



#transformation  
  
#sqrt transformation  
sqrt.fractal.dimension.m = (fractal\_dimension\_mean)^1/25  
  
hist (sqrt.fractal.dimension.m, xlab="Sqrt Fractal Dimension Mean")

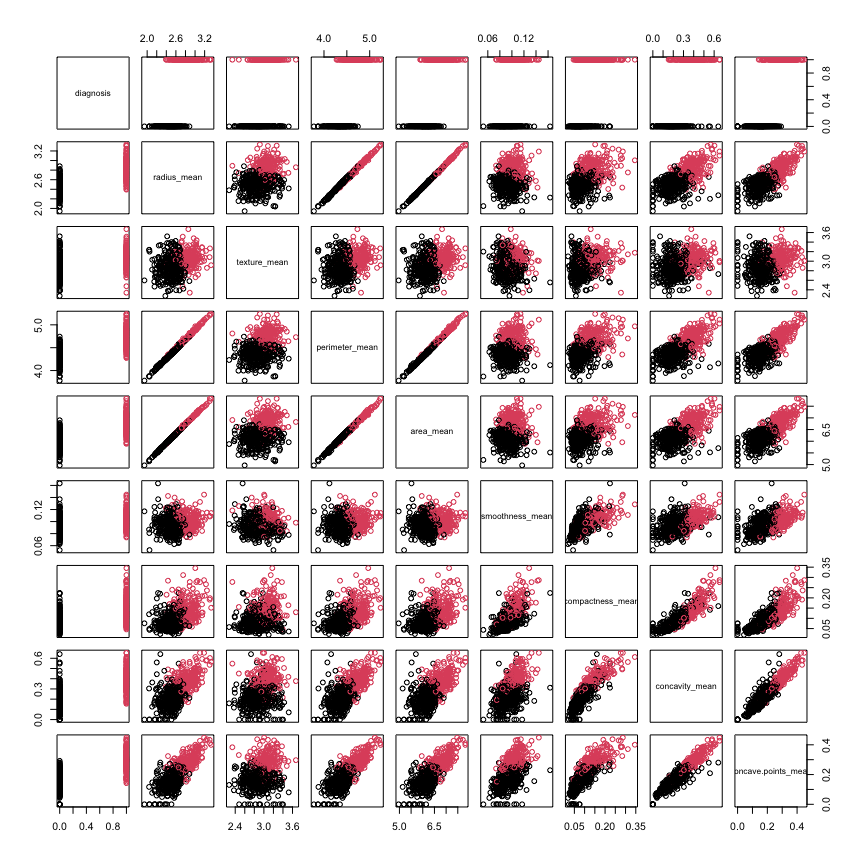


#log transformation  
log.fractal.dimension.m = log(fractal\_dimension\_mean)  
hist (log.fractal.dimension.m, xlab="Log Fractal Dimension Mean")



#Scatter plot matrix

Breastcancer$radius\_mean = log.radius.m  
Breastcancer$texture\_mean = log.texture.m  
Breastcancer$perimeter\_mean = log.perimeter.m  
Breastcancer$area\_mean = log.area.m  
Breastcancer$concavity\_mean = root.concavity.m  
Breastcancer$concave.points\_mean = root.concave.points.m  
Breastcancer$symmetry\_mean = log.symmetry.m  
Breastcancer$fractal\_dimension\_mean = log.fractal.dimension.m  
  
plot (Breastcancer[,2:10], col=Breastcancer$diagnosis+1)



This is the matrix plot from the scatter plot. There is some linear relationship between some of our predictor variables. For example, there is a strong linear relationship between radius mean and perimeter mean, radius mean and area mean and perimeter mean and area mean. These are variables that determine how big the tumor is and they mostly depend on how big radius is. The other linearity that is there is concavity mean and concave points mean. From the graph we can hypothesis that the greater the means of the predictor variables, the more chance of the tumor to be cancerous. However, we were curios on how the untransformed data would look on the matrix plot.

## Intial model: First model

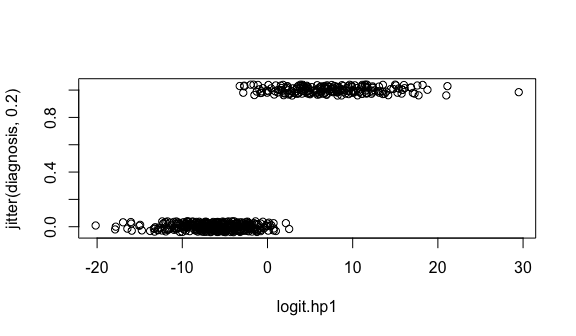
# Fitting the regression model:  
binary.logit = glm (data = Breastcancer, diagnosis ~ radius\_mean + texture\_mean + perimeter\_mean + area\_mean + smoothness\_mean + compactness\_mean + concavity\_mean + concave.points\_mean + symmetry\_mean + fractal\_dimension\_mean, family = 'binomial')  
summary(binary.logit)

##   
## Call:  
## glm(formula = diagnosis ~ radius\_mean + texture\_mean + perimeter\_mean +   
## area\_mean + smoothness\_mean + compactness\_mean + concavity\_mean +   
## concave.points\_mean + symmetry\_mean + fractal\_dimension\_mean,   
## family = "binomial", data = Breastcancer)  
##   
## Deviance Residuals:   
## Min 1Q Median 3Q Max   
## -2.28827 -0.09484 -0.01332 0.01205 2.56832   
##   
## Coefficients:  
## Estimate Std. Error z value Pr(>|z|)   
## (Intercept) 149.262 125.693 1.188 0.235026   
## radius\_mean -65.634 66.485 -0.987 0.323549   
## texture\_mean 7.795 1.514 5.149 2.62e-07 \*\*\*  
## perimeter\_mean -227.611 84.263 -2.701 0.006909 \*\*   
## area\_mean 154.027 41.087 3.749 0.000178 \*\*\*  
## smoothness\_mean 109.620 40.892 2.681 0.007347 \*\*   
## compactness\_mean 97.259 40.132 2.423 0.015373 \*   
## concavity\_mean 5.630 7.931 0.710 0.477737   
## concave.points\_mean 18.173 14.585 1.246 0.212761   
## symmetry\_mean 2.623 2.223 1.180 0.238076   
## fractal\_dimension\_mean -6.775 6.815 -0.994 0.320119   
## ---  
## 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: 118.70 on 558 degrees of freedom  
## AIC: 140.7  
##   
## Number of Fisher Scoring iterations: 8

From our first model, only three predictor variables we significant, that is texture mean, area mean and smoothness.

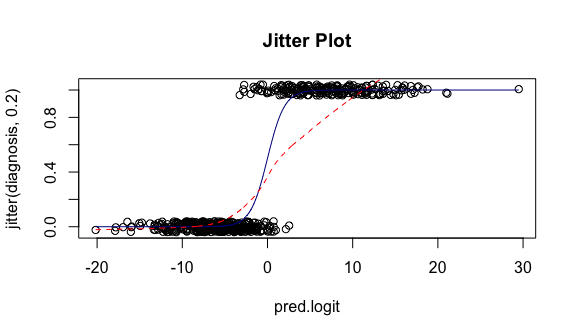
## Plot jittered response vs. predicted values with the fitted logistic curve and a lowess fit

logit.hp1 = predict (binary.logit)  
plot (jitter (diagnosis, 0.2) ~ logit.hp1, data=Breastcancer)



From the above jitter plot the points are more concentrated on the benign diagnosis which is 0. The concentration is less concentrated for the malignant points on the jitter plot.

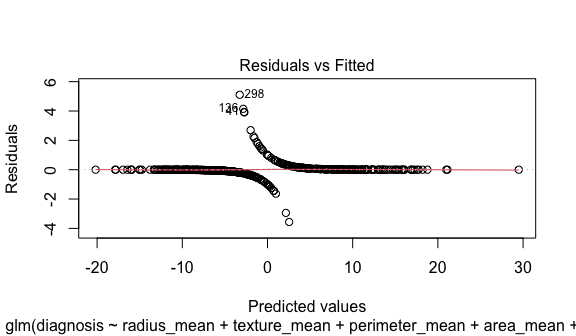
pred.logit = predict (binary.logit)  
plot(jitter(diagnosis,0.2) ~ pred.logit, data = Breastcancer, main = 'Jitter Plot')  
  
  
pred.r = predict(binary.logit, type='response')  
pred.order =order(pred.logit)  
  
lines(pred.logit[pred.order],pred.r[pred.order],col='darkblue')  
  
lines(lowess(pred.logit[pred.order], pred.r[pred.order]),col='red',lty=2)



There is a deviation from the blue and red line in the jitter plot. This can be explained by the concentration of the points in the benign diagnosis as compared to the malignant points in the data. Therefore, we reevaluate the first model to see if there are points we included that are not as significant and could be contributing to the deviation of the blue and red line in the jitter plot.

## Plot residuals vs. fitted values: plot (myfit, which=1)

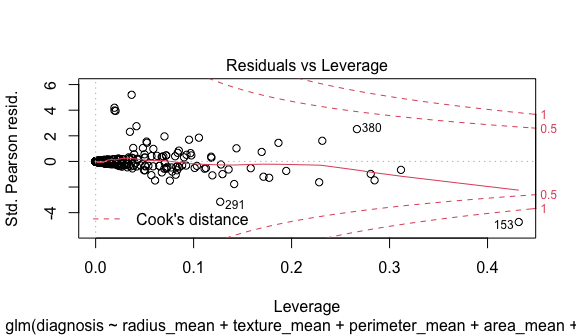
plot(binary.logit, which=1)



From the residual vs fitted plot most of the predicted values are along the zero mean of the residuals with a few deviation as 3 points are evidently the outliers.

##Check for points with high leverage or high Cook’s Distance

plot(binary.logit, which=5)



From the plot above only one point, the one marked 153 is a high leverage point.

##Check for multicollinearity

car::vif (binary.logit)

## radius\_mean texture\_mean perimeter\_mean   
## 905.624889 1.671990 1460.909111   
## area\_mean smoothness\_mean compactness\_mean   
## 1403.695351 4.644819 40.753595   
## concavity\_mean concave.points\_mean symmetry\_mean   
## 6.858544 6.642402 1.655188   
## fractal\_dimension\_mean   
## 7.611831

From the table above most of the points are correlated as they have vif value above 5 except texture\_mean, smoothness\_mean and the symmetry\_mean.

## Model Selection

#Stepwise regression on the First model

binary.logit.sw = step (binary.logit, directon='both')

## Start: AIC=140.7  
## diagnosis ~ radius\_mean + texture\_mean + perimeter\_mean + area\_mean +   
## smoothness\_mean + compactness\_mean + concavity\_mean + concave.points\_mean +   
## symmetry\_mean + fractal\_dimension\_mean

## Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred

## Df Deviance AIC  
## - concavity\_mean 1 119.21 139.21  
## - radius\_mean 1 119.70 139.70  
## - fractal\_dimension\_mean 1 119.70 139.70  
## - symmetry\_mean 1 120.11 140.10  
## - concave.points\_mean 1 120.29 140.29  
## <none> 118.70 140.70  
## - compactness\_mean 1 125.29 145.29  
## - smoothness\_mean 1 127.26 147.26  
## - perimeter\_mean 1 127.55 147.55  
## - area\_mean 1 139.57 159.57  
## - texture\_mean 1 156.78 176.78  
##   
## Step: AIC=139.21  
## diagnosis ~ radius\_mean + texture\_mean + perimeter\_mean + area\_mean +   
## smoothness\_mean + compactness\_mean + concave.points\_mean +   
## symmetry\_mean + fractal\_dimension\_mean

## Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred

## Df Deviance AIC  
## - fractal\_dimension\_mean 1 120.17 138.17  
## - symmetry\_mean 1 120.72 138.72  
## <none> 119.21 139.21  
## - radius\_mean 1 121.23 139.23  
## - concave.points\_mean 1 123.94 141.94  
## - smoothness\_mean 1 127.30 145.30  
## - compactness\_mean 1 127.58 145.58  
## - perimeter\_mean 1 128.79 146.79  
## - area\_mean 1 147.96 165.96  
## - texture\_mean 1 156.79 174.79  
##   
## Step: AIC=138.17  
## diagnosis ~ radius\_mean + texture\_mean + perimeter\_mean + area\_mean +   
## smoothness\_mean + compactness\_mean + concave.points\_mean +   
## symmetry\_mean

## Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred

## Df Deviance AIC  
## - radius\_mean 1 121.91 137.91  
## - symmetry\_mean 1 122.10 138.10  
## <none> 120.17 138.17  
## - concave.points\_mean 1 125.72 141.72  
## - smoothness\_mean 1 127.35 143.35  
## - compactness\_mean 1 128.19 144.19  
## - perimeter\_mean 1 128.91 144.91  
## - area\_mean 1 148.19 164.19  
## - texture\_mean 1 159.45 175.45  
##   
## Step: AIC=137.91  
## diagnosis ~ texture\_mean + perimeter\_mean + area\_mean + smoothness\_mean +   
## compactness\_mean + concave.points\_mean + symmetry\_mean  
##   
## Df Deviance AIC  
## - symmetry\_mean 1 123.88 137.88  
## <none> 121.91 137.91  
## - smoothness\_mean 1 127.41 141.41  
## - concave.points\_mean 1 131.49 145.49  
## - compactness\_mean 1 136.72 150.72  
## - perimeter\_mean 1 145.78 159.78  
## - area\_mean 1 148.21 162.21  
## - texture\_mean 1 165.02 179.02  
##   
## Step: AIC=137.88  
## diagnosis ~ texture\_mean + perimeter\_mean + area\_mean + smoothness\_mean +   
## compactness\_mean + concave.points\_mean  
##   
## Df Deviance AIC  
## <none> 123.88 137.88  
## - smoothness\_mean 1 130.22 142.22  
## - concave.points\_mean 1 133.56 145.56  
## - compactness\_mean 1 142.92 154.92  
## - perimeter\_mean 1 150.82 162.82  
## - area\_mean 1 153.25 165.25  
## - texture\_mean 1 165.87 177.87

#Retained model by the stepwise regression

binary.logit2 = glm (data = Breastcancer, diagnosis ~ texture\_mean + perimeter\_mean + area\_mean + smoothness\_mean + compactness\_mean + concave.points\_mean, family = 'binomial')  
summary(binary.logit2)

##   
## Call:  
## glm(formula = diagnosis ~ texture\_mean + perimeter\_mean + area\_mean +   
## smoothness\_mean + compactness\_mean + concave.points\_mean,   
## family = "binomial", data = Breastcancer)  
##   
## Deviance Residuals:   
## Min 1Q Median 3Q Max   
## -2.43744 -0.10684 -0.01919 0.01436 2.53444   
##   
## Coefficients:  
## Estimate Std. Error z value Pr(>|z|)   
## (Intercept) 276.462 88.202 3.134 0.001722 \*\*   
## texture\_mean 7.730 1.434 5.390 7.05e-08 \*\*\*  
## perimeter\_mean -286.885 72.095 -3.979 6.91e-05 \*\*\*  
## area\_mean 151.170 36.599 4.130 3.62e-05 \*\*\*  
## smoothness\_mean 80.471 33.671 2.390 0.016853 \*   
## compactness\_mean 109.651 31.620 3.468 0.000525 \*\*\*  
## concave.points\_mean 31.731 10.870 2.919 0.003512 \*\*   
## ---  
## 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: 123.88 on 562 degrees of freedom  
## AIC: 137.88  
##   
## Number of Fisher Scoring iterations: 8

From the step- wise regression performed on the initial model, we got the model with the lowest AIC value of 137.88. The predictor variables in this model are all significant with their p-values below 0.05 except for symmetry mean. Removing the symmetry mean would increase the AIC value so we decided to keep it there.

#Centering and interactions

#Centering quantitative predictor variables  
texture.c = scale(Breastcancer$texture\_mean, center = TRUE, scale = FALSE)  
perimeter.c = scale(Breastcancer$perimeter\_mean, center = TRUE, scale = FALSE)  
area.c = scale(Breastcancer$area\_mean, center = TRUE, scale = FALSE)  
smoothness.c = scale(Breastcancer$smoothness\_mean, center = TRUE, scale = FALSE)  
compactness.c = scale(Breastcancer$compactness\_mean, center = TRUE, scale = FALSE)  
concave.points.c = scale(Breastcancer$concave.points\_mean, center = TRUE, scale = FALSE)  
  
#Interactions  
  
binary.logit3 = glm (data = Breastcancer, diagnosis ~ (texture.c + perimeter.c + area.c + smoothness.c  
 + compactness.c + concave.points.c)^2, family = 'binomial')

## Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred

summary(binary.logit3)

##   
## Call:  
## glm(formula = diagnosis ~ (texture.c + perimeter.c + area.c +   
## smoothness.c + compactness.c + concave.points.c)^2, family = "binomial",   
## data = Breastcancer)  
##   
## Deviance Residuals:   
## Min 1Q Median 3Q Max   
## -2.35537 -0.06229 -0.00356 0.00321 2.53195   
##   
## Coefficients:  
## Estimate Std. Error z value Pr(>|z|)   
## (Intercept) -9.788e-01 5.234e-01 -1.870 0.061461 .   
## texture.c 9.471e+00 2.600e+00 3.642 0.000270 \*\*\*  
## perimeter.c -5.063e+02 1.340e+02 -3.779 0.000157 \*\*\*  
## area.c 2.622e+02 6.835e+01 3.837 0.000125 \*\*\*  
## smoothness.c 1.066e+02 6.066e+01 1.757 0.078889 .   
## compactness.c 1.769e+02 5.916e+01 2.990 0.002790 \*\*   
## concave.points.c 5.669e+01 1.728e+01 3.280 0.001037 \*\*   
## texture.c:perimeter.c -7.184e+02 6.272e+02 -1.145 0.252068   
## texture.c:area.c 3.689e+02 3.156e+02 1.169 0.242437   
## texture.c:smoothness.c 4.994e+02 2.604e+02 1.918 0.055067 .   
## texture.c:compactness.c 1.457e+02 2.759e+02 0.528 0.597420   
## texture.c:concave.points.c 6.147e+01 8.589e+01 0.716 0.474163   
## perimeter.c:area.c -2.286e+01 1.172e+01 -1.951 0.051005 .   
## perimeter.c:smoothness.c -2.548e+04 1.088e+04 -2.343 0.019139 \*   
## perimeter.c:compactness.c -5.681e+02 6.790e+02 -0.837 0.402797   
## perimeter.c:concave.points.c 4.922e+03 1.650e+03 2.983 0.002853 \*\*   
## area.c:smoothness.c 1.251e+04 5.416e+03 2.310 0.020861 \*   
## area.c:compactness.c 2.927e+02 3.542e+02 0.826 0.408708   
## area.c:concave.points.c -2.430e+03 8.391e+02 -2.896 0.003783 \*\*   
## smoothness.c:compactness.c 1.184e+04 4.699e+03 2.520 0.011737 \*   
## smoothness.c:concave.points.c 2.729e+02 8.986e+02 0.304 0.761406   
## compactness.c:concave.points.c -1.795e+03 8.502e+02 -2.112 0.034698 \*   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## (Dispersion parameter for binomial family taken to be 1)  
##   
## Null deviance: 751.440 on 568 degrees of freedom  
## Residual deviance: 91.767 on 547 degrees of freedom  
## AIC: 135.77  
##   
## Number of Fisher Scoring iterations: 10

From the above output, the centered values with the interaction values of texture mean,perimeter and area mean are most significant. There are no interaction effects that that are significant at this moment since all their p values are greater that 0.05. We decided to run another stepwise regression on this model.

#Interactions on the centered and interacted model

binary.logit.sw2 = step (binary.logit3, directon='both')

## Start: AIC=135.77  
## diagnosis ~ (texture.c + perimeter.c + area.c + smoothness.c +   
## compactness.c + concave.points.c)^2

## Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred  
  
## Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred  
  
## Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred  
  
## Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred  
  
## Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred  
  
## Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred  
  
## Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred  
  
## Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred  
  
## Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred  
  
## Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred  
  
## Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred  
  
## Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred

## Df Deviance AIC  
## - smoothness.c:concave.points.c 1 91.860 133.86  
## - texture.c:compactness.c 1 92.052 134.05  
## - texture.c:concave.points.c 1 92.283 134.28  
## - area.c:compactness.c 1 92.474 134.47  
## - perimeter.c:compactness.c 1 92.482 134.48  
## - texture.c:perimeter.c 1 93.216 135.22  
## - texture.c:area.c 1 93.282 135.28  
## <none> 91.767 135.77  
## - perimeter.c:area.c 1 95.347 137.35  
## - texture.c:smoothness.c 1 95.946 137.95  
## - compactness.c:concave.points.c 1 96.377 138.38  
## - area.c:smoothness.c 1 99.181 141.18  
## - area.c:concave.points.c 1 99.264 141.26  
## - perimeter.c:smoothness.c 1 99.363 141.36  
## - perimeter.c:concave.points.c 1 99.636 141.64  
## - smoothness.c:compactness.c 1 99.954 141.95

## Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred

##   
## Step: AIC=133.86  
## diagnosis ~ texture.c + perimeter.c + area.c + smoothness.c +   
## compactness.c + concave.points.c + texture.c:perimeter.c +   
## texture.c:area.c + texture.c:smoothness.c + texture.c:compactness.c +   
## texture.c:concave.points.c + perimeter.c:area.c + perimeter.c:smoothness.c +   
## perimeter.c:compactness.c + perimeter.c:concave.points.c +   
## area.c:smoothness.c + area.c:compactness.c + area.c:concave.points.c +   
## smoothness.c:compactness.c + compactness.c:concave.points.c

## Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred  
  
## Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred  
  
## Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred  
  
## Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred  
  
## Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred  
  
## Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred  
  
## Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred  
  
## Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred  
  
## Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred  
  
## Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred  
  
## Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred

## Df Deviance AIC  
## - texture.c:compactness.c 1 92.218 132.22  
## - texture.c:concave.points.c 1 92.325 132.32  
## - perimeter.c:compactness.c 1 92.833 132.83  
## - area.c:compactness.c 1 92.885 132.88  
## - texture.c:perimeter.c 1 93.387 133.39  
## - texture.c:area.c 1 93.454 133.45  
## <none> 91.860 133.86  
## - perimeter.c:area.c 1 95.347 135.35  
## - texture.c:smoothness.c 1 95.967 135.97  
## - compactness.c:concave.points.c 1 96.556 136.56  
## - area.c:smoothness.c 1 99.271 139.27  
## - area.c:concave.points.c 1 99.317 139.32  
## - perimeter.c:smoothness.c 1 99.445 139.44  
## - perimeter.c:concave.points.c 1 99.656 139.66  
## - smoothness.c:compactness.c 1 100.525 140.53

## Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred

##   
## Step: AIC=132.22  
## diagnosis ~ texture.c + perimeter.c + area.c + smoothness.c +   
## compactness.c + concave.points.c + texture.c:perimeter.c +   
## texture.c:area.c + texture.c:smoothness.c + texture.c:concave.points.c +   
## perimeter.c:area.c + perimeter.c:smoothness.c + perimeter.c:compactness.c +   
## perimeter.c:concave.points.c + area.c:smoothness.c + area.c:compactness.c +   
## area.c:concave.points.c + smoothness.c:compactness.c + compactness.c:concave.points.c

## Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred  
  
## Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred  
  
## Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred  
  
## Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred  
  
## Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred  
  
## Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred  
  
## Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred  
  
## Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred  
  
## Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred  
  
## Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred

## Df Deviance AIC  
## - texture.c:concave.points.c 1 92.872 130.87  
## - perimeter.c:compactness.c 1 93.005 131.00  
## - area.c:compactness.c 1 93.075 131.07  
## <none> 92.218 132.22  
## - perimeter.c:area.c 1 95.886 133.89  
## - texture.c:smoothness.c 1 96.226 134.23  
## - texture.c:perimeter.c 1 97.427 135.43  
## - texture.c:area.c 1 98.123 136.12  
## - compactness.c:concave.points.c 1 98.363 136.36  
## - area.c:smoothness.c 1 99.516 137.52  
## - perimeter.c:smoothness.c 1 99.700 137.70  
## - smoothness.c:compactness.c 1 100.835 138.84  
## - area.c:concave.points.c 1 100.952 138.95  
## - perimeter.c:concave.points.c 1 101.395 139.40

## Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred

##   
## Step: AIC=130.87  
## diagnosis ~ texture.c + perimeter.c + area.c + smoothness.c +   
## compactness.c + concave.points.c + texture.c:perimeter.c +   
## texture.c:area.c + texture.c:smoothness.c + perimeter.c:area.c +   
## perimeter.c:smoothness.c + perimeter.c:compactness.c + perimeter.c:concave.points.c +   
## area.c:smoothness.c + area.c:compactness.c + area.c:concave.points.c +   
## smoothness.c:compactness.c + compactness.c:concave.points.c

## Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred  
  
## Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred  
  
## Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred  
  
## Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred  
  
## Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred  
  
## Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred  
  
## Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred  
  
## Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred

## Df Deviance AIC  
## - perimeter.c:compactness.c 1 93.365 129.37  
## - area.c:compactness.c 1 93.434 129.43  
## <none> 92.872 130.87  
## - perimeter.c:area.c 1 97.225 133.22  
## - texture.c:perimeter.c 1 97.656 133.66  
## - texture.c:area.c 1 98.226 134.23  
## - compactness.c:concave.points.c 1 99.978 135.98  
## - area.c:smoothness.c 1 99.982 135.98  
## - perimeter.c:smoothness.c 1 100.230 136.23  
## - smoothness.c:compactness.c 1 101.447 137.45  
## - area.c:concave.points.c 1 102.083 138.08  
## - perimeter.c:concave.points.c 1 102.638 138.64  
## - texture.c:smoothness.c 1 104.303 140.30

## Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred

##   
## Step: AIC=129.36  
## diagnosis ~ texture.c + perimeter.c + area.c + smoothness.c +   
## compactness.c + concave.points.c + texture.c:perimeter.c +   
## texture.c:area.c + texture.c:smoothness.c + perimeter.c:area.c +   
## perimeter.c:smoothness.c + perimeter.c:concave.points.c +   
## area.c:smoothness.c + area.c:compactness.c + area.c:concave.points.c +   
## smoothness.c:compactness.c + compactness.c:concave.points.c

## Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred  
  
## Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred  
  
## Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred  
  
## Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred  
  
## Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred  
  
## Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred  
  
## Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred

## Df Deviance AIC  
## - area.c:compactness.c 1 93.484 127.48  
## <none> 93.365 129.37  
## - perimeter.c:area.c 1 97.449 131.45  
## - texture.c:perimeter.c 1 98.068 132.07  
## - texture.c:area.c 1 98.655 132.66  
## - area.c:smoothness.c 1 100.938 134.94  
## - compactness.c:concave.points.c 1 101.017 135.02  
## - perimeter.c:smoothness.c 1 101.107 135.11  
## - area.c:concave.points.c 1 102.091 136.09  
## - smoothness.c:compactness.c 1 102.318 136.32  
## - perimeter.c:concave.points.c 1 102.638 136.64  
## - texture.c:smoothness.c 1 104.633 138.63

## Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred

##   
## Step: AIC=127.48  
## diagnosis ~ texture.c + perimeter.c + area.c + smoothness.c +   
## compactness.c + concave.points.c + texture.c:perimeter.c +   
## texture.c:area.c + texture.c:smoothness.c + perimeter.c:area.c +   
## perimeter.c:smoothness.c + perimeter.c:concave.points.c +   
## area.c:smoothness.c + area.c:concave.points.c + smoothness.c:compactness.c +   
## compactness.c:concave.points.c

## Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred  
  
## Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred  
  
## Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred  
  
## Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred  
  
## Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred  
  
## Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred

## Df Deviance AIC  
## <none> 93.484 127.48  
## - texture.c:perimeter.c 1 98.381 130.38  
## - perimeter.c:area.c 1 98.399 130.40  
## - texture.c:area.c 1 98.975 130.97  
## - area.c:smoothness.c 1 100.944 132.94  
## - compactness.c:concave.points.c 1 101.034 133.03  
## - perimeter.c:smoothness.c 1 101.115 133.12  
## - area.c:concave.points.c 1 102.191 134.19  
## - smoothness.c:compactness.c 1 102.436 134.44  
## - perimeter.c:concave.points.c 1 102.782 134.78  
## - texture.c:smoothness.c 1 104.788 136.79

We got the smallest value of the AIC value in the model with the interaction effects which has the value 127.47, which is less that the one in the retained first model after step wise regression. Therefore, this qualifies as our final model.

#Final Model Summary And plot

binary.logit4 = glm (data = Breastcancer, diagnosis ~ texture.c + perimeter.c + area.c + smoothness.c   
 + compactness.c + concave.points.c + texture.c:perimeter.c + texture.c:area.c +  
 texture.c:smoothness.c + perimeter.c:area.c + perimeter.c:smoothness.c +  
 perimeter.c:concave.points.c +area.c:smoothness.c + area.c:concave.points.c +  
 smoothness.c:compactness.c + compactness.c:concave.points.c  
 , family = 'binomial')

## Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred

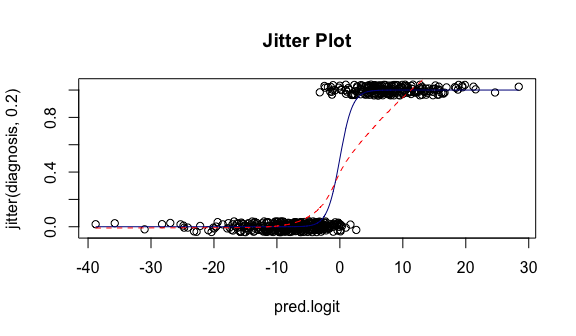
summary(binary.logit4)

##   
## Call:  
## glm(formula = diagnosis ~ texture.c + perimeter.c + area.c +   
## smoothness.c + compactness.c + concave.points.c + texture.c:perimeter.c +   
## texture.c:area.c + texture.c:smoothness.c + perimeter.c:area.c +   
## perimeter.c:smoothness.c + perimeter.c:concave.points.c +   
## area.c:smoothness.c + area.c:concave.points.c + smoothness.c:compactness.c +   
## compactness.c:concave.points.c, family = "binomial", data = Breastcancer)  
##   
## Deviance Residuals:   
## Min 1Q Median 3Q Max   
## -2.31315 -0.05493 -0.00325 0.00996 2.54235   
##   
## Coefficients:  
## Estimate Std. Error z value Pr(>|z|)   
## (Intercept) -1.110e+00 4.863e-01 -2.282 0.022486 \*   
## texture.c 8.510e+00 1.880e+00 4.526 6.00e-06 \*\*\*  
## perimeter.c -4.751e+02 1.142e+02 -4.161 3.17e-05 \*\*\*  
## area.c 2.462e+02 5.819e+01 4.232 2.32e-05 \*\*\*  
## smoothness.c 9.215e+01 5.058e+01 1.822 0.068449 .   
## compactness.c 1.616e+02 4.889e+01 3.306 0.000948 \*\*\*  
## concave.points.c 5.709e+01 1.635e+01 3.492 0.000479 \*\*\*  
## texture.c:perimeter.c -3.084e+02 1.412e+02 -2.184 0.028939 \*   
## texture.c:area.c 1.689e+02 7.340e+01 2.301 0.021371 \*   
## texture.c:smoothness.c 5.681e+02 1.884e+02 3.015 0.002566 \*\*   
## perimeter.c:area.c -2.321e+01 9.759e+00 -2.378 0.017414 \*   
## perimeter.c:smoothness.c -2.261e+04 8.652e+03 -2.614 0.008952 \*\*   
## perimeter.c:concave.points.c 5.160e+03 1.534e+03 3.364 0.000769 \*\*\*  
## area.c:smoothness.c 1.110e+04 4.288e+03 2.588 0.009655 \*\*   
## area.c:concave.points.c -2.544e+03 7.856e+02 -3.238 0.001205 \*\*   
## smoothness.c:compactness.c 1.057e+04 3.740e+03 2.827 0.004698 \*\*   
## compactness.c:concave.points.c -2.057e+03 6.973e+02 -2.951 0.003172 \*\*   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## (Dispersion parameter for binomial family taken to be 1)  
##   
## Null deviance: 751.440 on 568 degrees of freedom  
## Residual deviance: 93.484 on 552 degrees of freedom  
## AIC: 127.48  
##   
## Number of Fisher Scoring iterations: 9

The most significant predictors from the final model therefore are radius.c, texture.c, area.c, concave.points.c and radius.c:area.c. We could not take out some of the non significant predictors since it will increase the AIC value

#Final model Diagnosis

pred.logit = predict (binary.logit4)  
plot(jitter(diagnosis,0.2) ~ pred.logit, data = Breastcancer, main = 'Jitter Plot')  
  
  
pred.r = predict(binary.logit4, type='response')  
pred.order =order(pred.logit)  
  
lines(pred.logit[pred.order],pred.r[pred.order],col='darkblue')  
  
lines(lowess(pred.logit[pred.order], pred.r[pred.order]),col='red',lty=2)



summary(binary.logit4)

##   
## Call:  
## glm(formula = diagnosis ~ texture.c + perimeter.c + area.c +   
## smoothness.c + compactness.c + concave.points.c + texture.c:perimeter.c +   
## texture.c:area.c + texture.c:smoothness.c + perimeter.c:area.c +   
## perimeter.c:smoothness.c + perimeter.c:concave.points.c +   
## area.c:smoothness.c + area.c:concave.points.c + smoothness.c:compactness.c +   
## compactness.c:concave.points.c, family = "binomial", data = Breastcancer)  
##   
## Deviance Residuals:   
## Min 1Q Median 3Q Max   
## -2.31315 -0.05493 -0.00325 0.00996 2.54235   
##   
## Coefficients:  
## Estimate Std. Error z value Pr(>|z|)   
## (Intercept) -1.110e+00 4.863e-01 -2.282 0.022486 \*   
## texture.c 8.510e+00 1.880e+00 4.526 6.00e-06 \*\*\*  
## perimeter.c -4.751e+02 1.142e+02 -4.161 3.17e-05 \*\*\*  
## area.c 2.462e+02 5.819e+01 4.232 2.32e-05 \*\*\*  
## smoothness.c 9.215e+01 5.058e+01 1.822 0.068449 .   
## compactness.c 1.616e+02 4.889e+01 3.306 0.000948 \*\*\*  
## concave.points.c 5.709e+01 1.635e+01 3.492 0.000479 \*\*\*  
## texture.c:perimeter.c -3.084e+02 1.412e+02 -2.184 0.028939 \*   
## texture.c:area.c 1.689e+02 7.340e+01 2.301 0.021371 \*   
## texture.c:smoothness.c 5.681e+02 1.884e+02 3.015 0.002566 \*\*   
## perimeter.c:area.c -2.321e+01 9.759e+00 -2.378 0.017414 \*   
## perimeter.c:smoothness.c -2.261e+04 8.652e+03 -2.614 0.008952 \*\*   
## perimeter.c:concave.points.c 5.160e+03 1.534e+03 3.364 0.000769 \*\*\*  
## area.c:smoothness.c 1.110e+04 4.288e+03 2.588 0.009655 \*\*   
## area.c:concave.points.c -2.544e+03 7.856e+02 -3.238 0.001205 \*\*   
## smoothness.c:compactness.c 1.057e+04 3.740e+03 2.827 0.004698 \*\*   
## compactness.c:concave.points.c -2.057e+03 6.973e+02 -2.951 0.003172 \*\*   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## (Dispersion parameter for binomial family taken to be 1)  
##   
## Null deviance: 751.440 on 568 degrees of freedom  
## Residual deviance: 93.484 on 552 degrees of freedom  
## AIC: 127.48  
##   
## Number of Fisher Scoring iterations: 9

The final model jitter plot has shows that the final model is fitting the our data better than the first model which suggests an improvement in our model.

cbind.data.frame (exp.beta = exp (binary.logit4$coefficients),   
 exp (as.data.frame (confint(binary.logit4))))

## Waiting for profiling to be done...

## Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred  
  
## Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred  
  
## Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred  
  
## Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred  
  
## Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred  
  
## Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred  
  
## Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred  
  
## Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred  
  
## Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred  
  
## Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred  
  
## Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred  
  
## Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred  
  
## Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred  
  
## Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred  
  
## Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred  
  
## Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred  
  
## Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred  
  
## Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred  
  
## Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred  
  
## Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred  
  
## Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred  
  
## Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred  
  
## Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred  
  
## Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred  
  
## Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred  
  
## Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred  
  
## Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred  
  
## Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred  
  
## Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred  
  
## Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred  
  
## Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred  
  
## Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred  
  
## Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred  
  
## Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred  
  
## Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred  
  
## Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred  
  
## Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred  
  
## Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred  
  
## Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred  
  
## Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred  
  
## Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred  
  
## Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred  
  
## Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred  
  
## Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred  
  
## Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred  
  
## Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred  
  
## Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred  
  
## Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred  
  
## Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred  
  
## Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred  
  
## Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred  
  
## Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred  
  
## Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred  
  
## Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred  
  
## Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred  
  
## Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred  
  
## Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred  
  
## Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred  
  
## Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred  
  
## Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred  
  
## Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred  
  
## Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred  
  
## Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred  
  
## Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred  
  
## Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred  
  
## Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred  
  
## Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred  
  
## Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred  
  
## Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred  
  
## Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred  
  
## Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred  
  
## Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred  
  
## Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred  
  
## Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred  
  
## Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred  
  
## Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred  
  
## Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred  
  
## Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred  
  
## Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred  
  
## Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred  
  
## Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred  
  
## Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred  
  
## Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred  
  
## Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred  
  
## Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred  
  
## Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred  
  
## Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred  
  
## Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred  
  
## Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred  
  
## Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred  
  
## Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred  
  
## Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred  
  
## Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred  
  
## Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred  
  
## Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred  
  
## Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred  
  
## Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred  
  
## Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred  
  
## Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred  
  
## Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred  
  
## Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred  
  
## Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred  
  
## Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred  
  
## Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred  
  
## Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred  
  
## Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred  
  
## Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred  
  
## Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred  
  
## Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred  
  
## Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred  
  
## Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred  
  
## Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred  
  
## Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred  
  
## Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred  
  
## Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred  
  
## Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred  
  
## Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred  
  
## Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred  
  
## Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred  
  
## Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred  
  
## Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred  
  
## Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred  
  
## Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred  
  
## Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred  
  
## Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred  
  
## Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred  
  
## Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred  
  
## Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred  
  
## Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred  
  
## Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred  
  
## Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred  
  
## Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred  
  
## Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred  
  
## Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred  
  
## Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred  
  
## Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred  
  
## Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred  
  
## Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred  
  
## Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred  
  
## Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred  
  
## Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred  
  
## Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred  
  
## Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred  
  
## Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred  
  
## Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred  
  
## Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred  
  
## Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred  
  
## Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred  
  
## Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred  
  
## Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred  
  
## Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred  
  
## Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred  
  
## Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred  
  
## Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred  
  
## Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred  
  
## Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred  
  
## Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred  
  
## Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred  
  
## Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred  
  
## Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred  
  
## Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred  
  
## Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred  
  
## Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred  
  
## Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred  
  
## Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred  
  
## Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred  
  
## Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred  
  
## Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred  
  
## Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred  
  
## Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred  
  
## Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred  
  
## Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred  
  
## Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred  
  
## Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred  
  
## Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred  
  
## Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred  
  
## Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred  
  
## Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred  
  
## Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred  
  
## Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred

## exp.beta 2.5 % 97.5 %  
## (Intercept) 3.296318e-01 1.195057e-01 8.409776e-01  
## texture.c 4.963591e+03 1.696761e+02 3.000981e+05  
## perimeter.c 4.861923e-207 1.565152e-312 2.113289e-114  
## area.c 8.708560e+106 8.422961e+59 8.192034e+160  
## smoothness.c 1.050152e+40 2.129084e-01 6.560783e+85  
## compactness.c 1.553823e+70 5.765612e+29 5.760579e+113  
## concave.points.c 6.228281e+24 8.679608e+11 9.521013e+39  
## texture.c:perimeter.c 1.203222e-134 3.636976e-260 4.613340e-16  
## texture.c:area.c 2.320964e+73 9.794424e+11 8.495162e+138  
## texture.c:smoothness.c 5.195387e+246 9.065950e+98 Inf  
## perimeter.c:area.c 8.351982e-11 1.487717e-19 6.104222e-02  
## perimeter.c:smoothness.c 0.000000e+00 0.000000e+00 0.000000e+00  
## perimeter.c:concave.points.c Inf Inf Inf  
## area.c:smoothness.c Inf Inf Inf  
## area.c:concave.points.c 0.000000e+00 0.000000e+00 0.000000e+00  
## smoothness.c:compactness.c Inf Inf Inf  
## compactness.c:concave.points.c 0.000000e+00 0.000000e+00 4.426768e-268

#Max Likelihood test

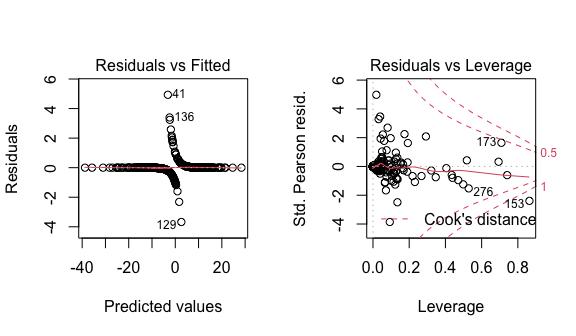
1 - pchisq(binary.logit4$null.deviance - binary.logit4$deviance, binary.logit4$df.null -  
 binary.logit4$df.residual)

## [1] 0

The likelihood test from the above has a value of 0. We were worried about multicollinearirty of the data which was shown by the matrix. However due to our low VIF value we do not have be worried about the collinearity.

#Residual Plot

par(mfrow = c(1,2))   
plot(binary.logit4, which=c(1,5))



The residual vs fitted plot is normal with a deviation of a few points namely points 380, 153 and 276. Point 389 is a high leverage point with pount 153 and 276 having hugh cook`s distance value.

#------------------------Deviance test of lack of fit--------------------  
  
# First model  
pchisq(deviance(binary.logit), df.residual(binary.logit), lower=F)

## [1] 1

# Final model  
pchisq(deviance(binary.logit4), df.residual(binary.logit4), lower=F)

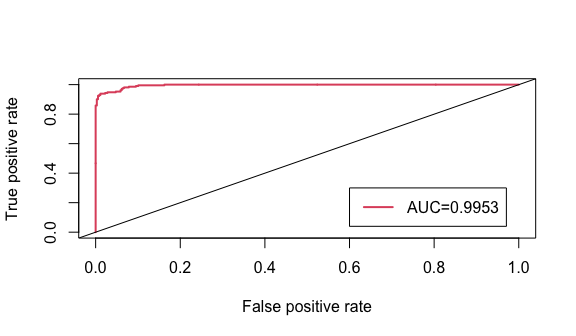
## [1] 1

There is no significant lack of fit in either the first model or the final model (p > 0.05), with the deviance from both models equal to 1.

library(ROCR)  
pred1 <- prediction(binary.logit4$fitted.values, binary.logit4$y)  
perf1 <- performance(pred1,"tpr","fpr")  
auc1 <- performance(pred1,"auc")@y.values[[1]]   
auc1

## [1] 0.9953227

plot(perf1, lwd=2, col=2)  
abline(0,1)  
legend(0.6, 0.3, c(paste ("AUC=", round (auc1, 4), sep="")), lwd=2, col=2)



roc.x = slot (perf1, "x.values")[[1]]   
roc.y = slot (perf1, "y.values") [[1]]  
cutoffs = slot (perf1, "alpha.values") [[1]]  
auc.table = cbind.data.frame(cutoff=pred1@cutoffs,tp=pred1@tp, fp=pred1@fp, tn=pred1@tn,fn=pred1@fn)  
names (auc.table) = c("Cutoff", "TP", "FP", "TN", "FN")   
auc.table$sensitivity = auc.table$TP / (auc.table$TP +auc.table$FN)  
auc.table$specificity = auc.table$TN / (auc.table$TN + auc.table$FP)   
auc.table$FalsePosRate = 1 - auc.table$specificity   
auc.table$sens\_spec = auc.table$sensitivity + auc.table$specificity  
auc.best = auc.table [auc.table$sens\_spec == max (auc.table$sens\_spec),]   
auc.best

## Cutoff TP FP TN FN sensitivity specificity FalsePosRate sens\_spec  
## 264 0.5548863 199 4 353 13 0.9386792 0.9887955 0.01120448 1.927475

Based on our ROC value our model managed to fit 99.53 % of our data. We got high specifity and sensitivity values. Our cut off point was 0.555. We had 199 true positives, 4 false positives, 353 true negetives and 13 false positive. This shows how well our model managed to fit the data set.

preds = predict (binary.logit4, se.fit = T)  
pred.df = cbind.data.frame (Breastcancer, as.data.frame (preds))  
pred.df$lwr = pred.df$fit - 1.96 \* pred.df$se.fit  
pred.df$upr = pred.df$fit + 1.96 \* pred.df$se.fit  
pred.df$fit.pr = round (exp (pred.df$fit) / (1 + exp (pred.df$fit)), 3)   
pred.df$lwr.pr = round (exp (pred.df$lwr) / (1 + exp (pred.df$lwr)), 3)  
pred.df$upr.pr = round (exp (pred.df$upr) / (1 + exp (pred.df$upr)), 3)

pred.df [c(30,47,78,21,16,11), c(2,5:7,14:16)]

## diagnosis perimeter\_mean area\_mean smoothness\_mean se.fit residual.scale  
## 30 1 4.744932 6.861816 0.09847 0.8617935 1  
## 47 0 3.945651 5.307773 0.08600 5.4611682 1  
## 78 1 4.789157 6.913737 0.10650 4.1923501 1  
## 21 0 4.450036 6.253829 0.10750 1.7851696 1  
## 16 1 4.571924 6.490420 0.11390 2.5426633 1  
## 11 1 4.631812 6.681858 0.08206 1.0948803 1  
## lwr  
## 30 0.6158831  
## 47 -38.9487736  
## 78 -1.4364696  
## 21 -13.1854639  
## 16 7.0871456  
## 11 1.1555543

pred.df$pred.dis = ifelse (pred.df$fit.pr >= auc.best$Cutoff[1], "Pred.Yes", "Pred.No")  
table (pred.df$diagnosis, pred.df$pred.dis)

##   
## Pred.No Pred.Yes  
## 0 353 4  
## 1 13 199

# Conclusions

The most significant predictors are radius mean, texture mean, area mean and concave points mean. Our model successfully predicted 199 as patients diagnosed with breast cancer, and had false positive of 4 patients. Therefore, that is why are ROC curve had a high AUC value of 0.9953.