Detecting Breast Cancer With Logistic Regression

Project Summary

In this study, the goal was to identify numerical factors based on measurements taken from needle biopsies of
masses from 569 women that would distinguish cancerous from non-cancerous tumors, in order to assist physicians in
accurate diagnosis and appropriate treatment. The data contains the target variable “diagnosis” which is the final, presumed
correct, diagnosis of whether the tumor is malignant or benign.

Data Processing

#1

```
> mydata <- read.csv("Desktop/wbcd0.csv", head=TRUE, sep="\t")

> summary(mydata)
 diagnosis       radius          texture         perimeter
 B:357       Min.   : 6.981   Min.   : 9.71   Min.   : 43.79
 M:212       1st Qu.:11.700   1st Qu.:16.17   1st Qu.: 75.17
             Median :13.370   Median :18.84   Median : 86.24
             Mean   :14.127   Mean   :19.29   Mean   : 91.97
             3rd Qu.:15.780   3rd Qu.:21.80   3rd Qu.:104.10
             Max.   :28.110   Max.   :39.28   Max.   :188.50
      area          smoothness       compactness
 Min.   : 143.5   Min.   :0.05263   Min.   :0.01938
 1st Qu.: 420.3   1st Qu.:0.08637   1st Qu.:0.06492
 Median : 551.1   Median :0.09587   Median :0.09263
 Mean   : 654.9   Mean   :0.09636   Mean   :0.10434
 3rd Qu.: 782.7   3rd Qu.:0.10530   3rd Qu.:0.13040
 Max.   :2501.0   Max.   :0.16340   Max.   :0.34540
    concavity          points          symmetry
 Min.   :0.00000   Min.   :0.00000   Min.   :0.1060
 1st Qu.:0.02956   1st Qu.:0.02031   1st Qu.:0.1619
 Median :0.06154   Median :0.03350   Median :0.1792
 Mean   :0.08880   Mean   :0.04892   Mean   :0.1812
 3rd Qu.:0.13070   3rd Qu.:0.07400   3rd Qu.:0.1957
 Max.   :0.42680   Max.   :0.20120   Max.   :0.3040
   dimension
 Min.   :0.04996
 1st Qu.:0.05770
 Median :0.06154
 Mean   :0.06280
 3rd Qu.:0.06612
```

Max. :0.09744

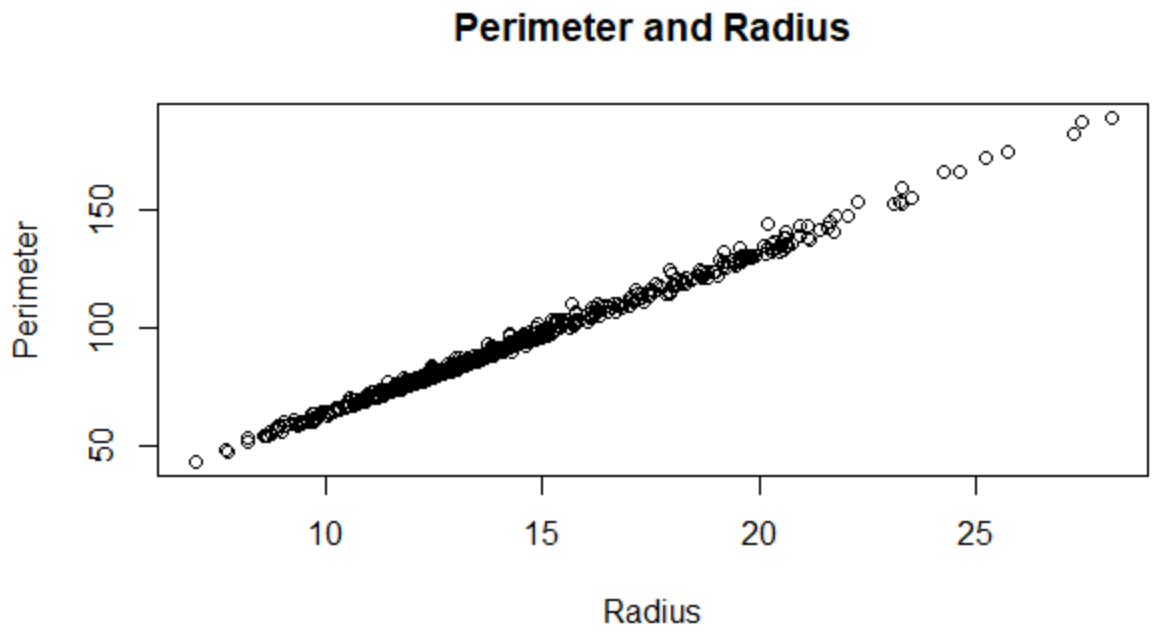
#2

Simplifying names of this database, we have:

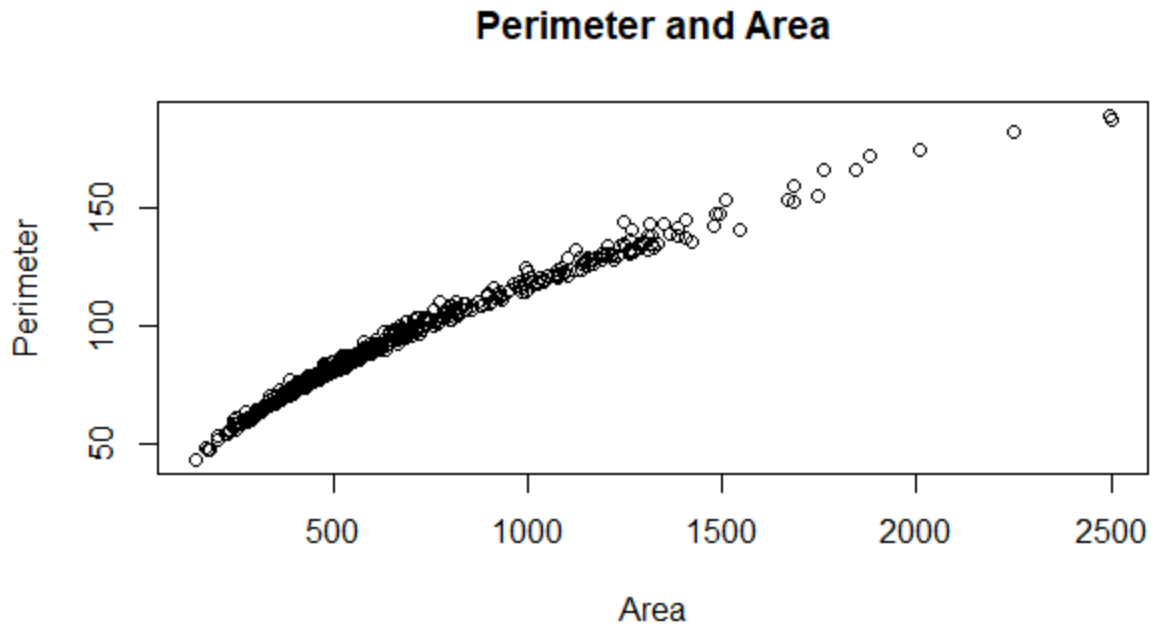
```
> radius <- mydata$radius
> texture <- mydata$texture
> perimeter <- mydata$perimeter
> area <- mydata$area
> smoothness <- mydata$smoothness
> compactness <- mydata$compactness
> concavity <- mydata$concavity
> points <- mydata$points
> symmetry <- mydata$symmetry
> dimension <- mydata$dimension
```

Pair plots of radius, perimeter, and area:

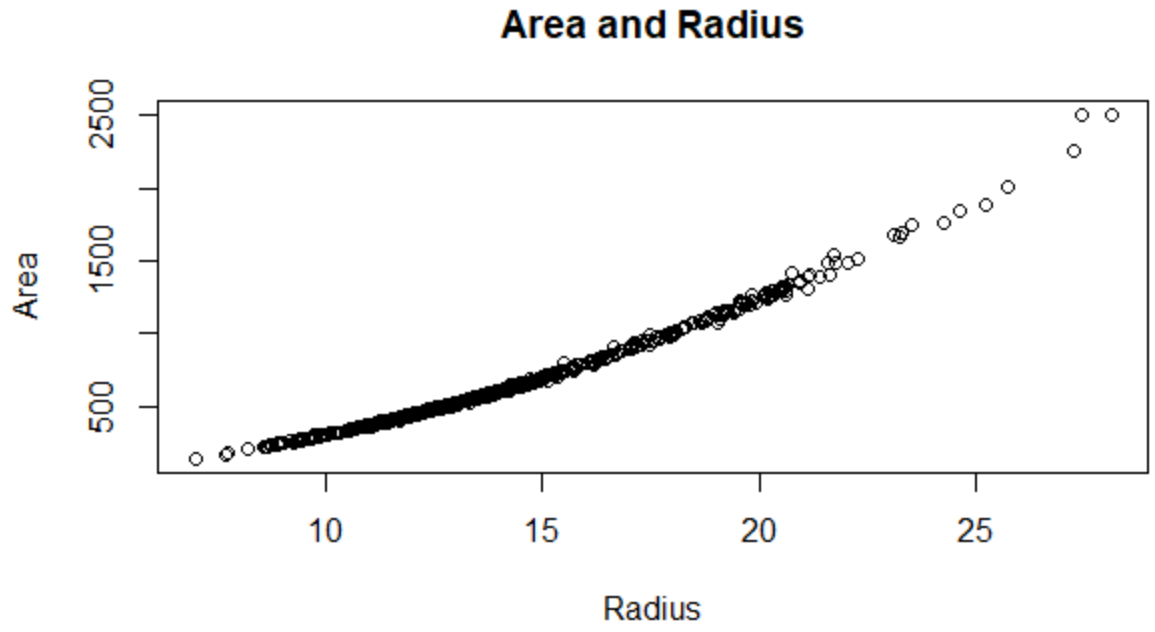
```
> plot(radius, perimeter, xlab = "Radius", ylab = "Perimeter", main ="Perimeter and Radius")
```




```
> plot(area, perimeter, xlab = "Area", ylab = "Perimeter", main ="Perimeter and Area")
```




```
> plot(radius, area, xlab = "Radius", ylab = "Area", main ="Area and Radius")
```



Correlation analysis:

```
> cor.test(radius,area)
```

Pearson's product-moment correlation

```
data:  radius and area
t = 148.32, df = 567, p-value < 2.2e-16
alternative hypothesis: true correlation is not equal to 0
95 percent confidence interval:
 0.9851095 0.9892674
sample estimates:
      cor
0.9873572
```

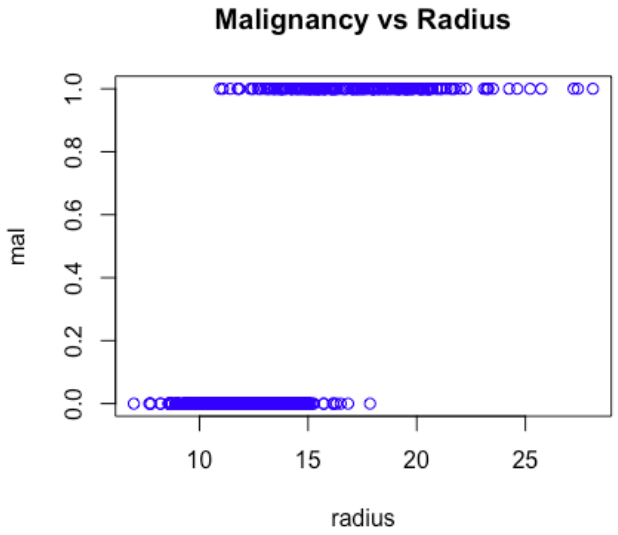
```
> cor.test(perimeter,area)

        Pearson's product-moment correlation

data:  perimeter and area
t = 143.48, df = 567, p-value < 2.2e-16
alternative hypothesis: true correlation is not equal to 0
95 percent confidence interval:
 0.9841091 0.9885448
sample estimates:
      cor
0.9865068
```

```
> cor.test(radius, perimeter)

        Pearson's product-moment correlation

data:  radius and perimeter
t = 362.99, df = 567, p-value < 2.2e-16
alternative hypothesis: true correlation is not equal to 0
95 percent confidence interval:
 0.9974716 0.9981808
sample estimates:
      cor
0.9978553
```

Based on our graphs and correlation test, we can conclude that the three variables: radius, perimeter, and area are actually very correlated with each other, with the correlation parameters >0.98; which are very close to 1. Therefore, we can get away with just one variable without affecting the regression model. From here, we continue our project, choosing radius to represent the three.

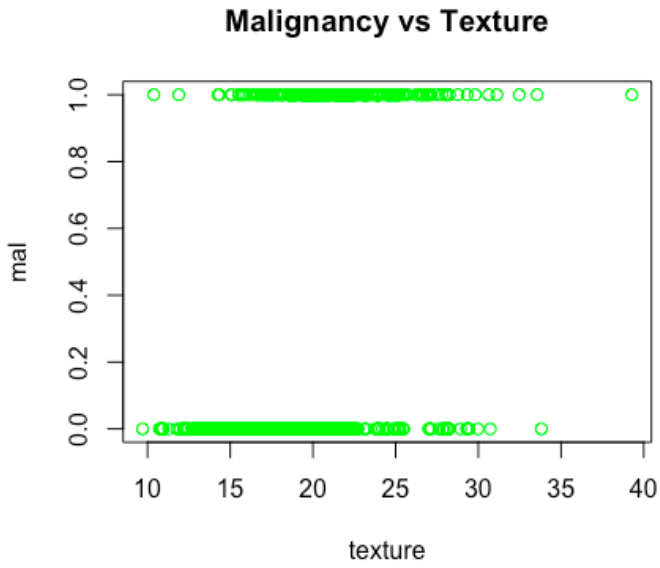
#3

```
> mal <- ifelse(mydata$diagnosis=="M", 1, 0)

> plot(radius, mal, main="Malignancy vs Radius", col="blue")
> modelrad <- glm(mal~radius,family=binomial)
> summary(modelrad)
```




```
Call:
glm(formula = mal ~ radius, family = binomial)
Deviance Residuals:
    Min       1Q   Median       3Q      Max
-2.5470  -0.4694  -0.1746   0.1513   2.8098
Coefficients:
             Estimate   Std. Error   z value   Pr(>|z|)
(Intercept) -15.24587   1.32463   -11.51   <2e-16 ***
radius        1.03359    0.09311    11.10   <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: 330.01  on 567  degrees of
freedom
AIC: 334.01
Number of Fisher Scoring iterations: 6
```

```
> plot(texture, mal, main="Malignancy vs Texture", col="green")
> modeltext <- glm(mal~texture,family=binomial)
> summary(modeltext)
```




```
Call:
glm(formula = mal ~ texture, family = binomial)
Deviance Residuals:
    Min       1Q   Median       3Q      Max
-2.3942  -0.8451  -0.5881   1.1022   2.3477
Coefficients:
             Estimate   Std. Error  z value   Pr(>|z|)
(Intercept) -5.12577   0.52638   -9.738    <2e-16 ***
texture      0.23464    0.02614    8.975    <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: 646.52 on 567 degrees of freedom
AIC: 650.52
Number of Fisher Scoring iterations: 4

```
> plot(smoothness, mal, main="Malignancy vs Smoothness", col="red")
> modelsmooth <- glm(mal~smoothness, family=binomial)
> summary(modelsmooth)
```

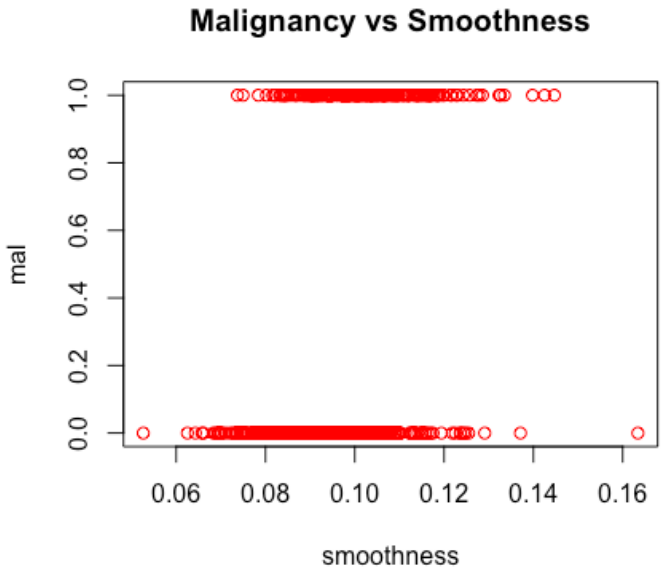
Call:
glm(formula = mal ~ smoothness, family = binomial)
Deviance Residuals:

Min	1Q	Median	3Q	Max
-2.6352	-0.9148	-0.6357	1.1428	2.0404

Coefficients:

	Estimate	Std. Error	z value	Pr(>|z|)	
(Intercept)	-6.3773	0.7474	-8.532	< 2e-16	***
smoothness	60.0857	7.5497	7.959	1.74e-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: 673.95 on 567 degrees of freedom
AIC: 677.95
Number of Fisher Scoring iterations: 3



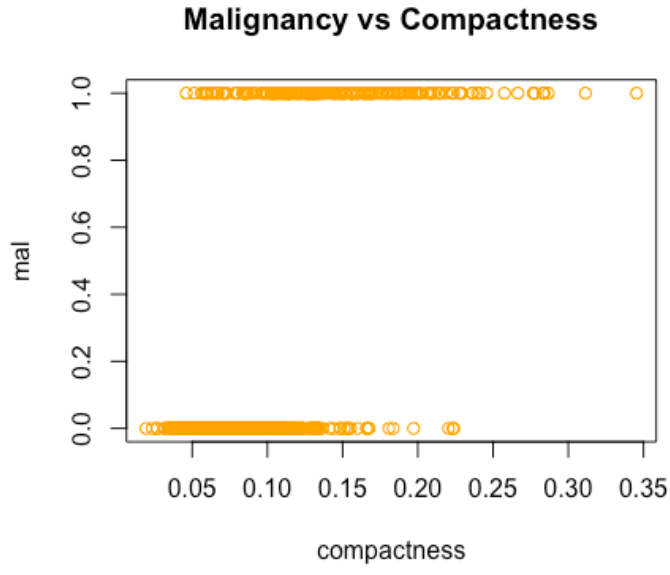

```
> plot(compactness, mal, main="Malignancy vs Compactness", col="orange")
> modelcomp <-
glm(mal~compactness,family=binomial)
> summary(modelcomp)
```

Call:
glm(formula = mal ~ compactness, family = binomial)
Deviance Residuals:

Min	1Q	Median	3Q	Max
-2.7454	-0.6513	-0.3985	0.6546	2.3615

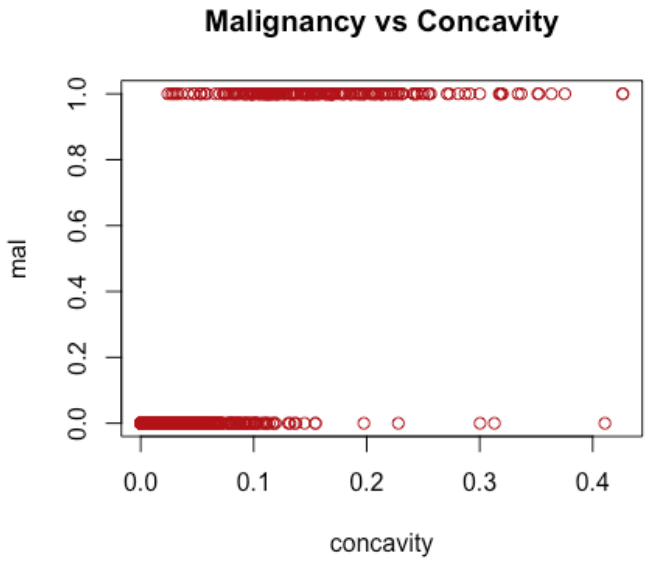
Coefficients:

	Estimate	Std. Error	z value	Pr(>|z|)	
(Intercept)	-4.4001	0.3563	-12.35	<2e-16	***
compactness	36.3798	3.1868	11.42	<2e-16	***

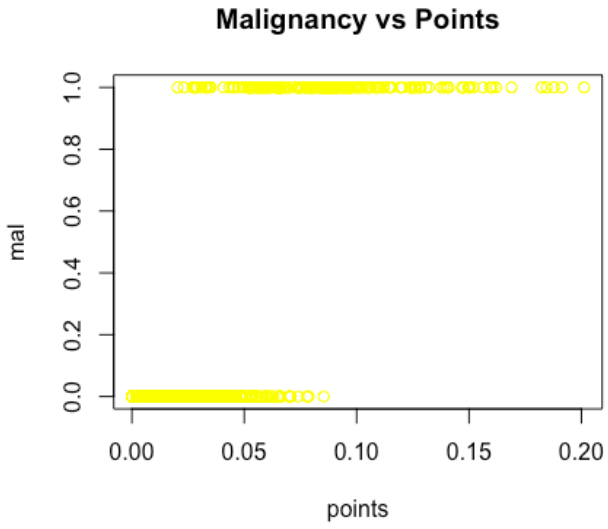


```
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
(Dispersion parameter for binomial family taken to be 1)
Null deviance: 751.44  on 568  degrees of freedom
Residual deviance: 508.79  on 567  degrees of freedom
AIC: 512.79
Number of Fisher Scoring iterations: 5
```

```
> plot(concavity, mal, main="Malignancy vs Concavity", col="brown")
> modelcon <- glm(mal~concavity,family=binomial)
> summary(modelcon)
```



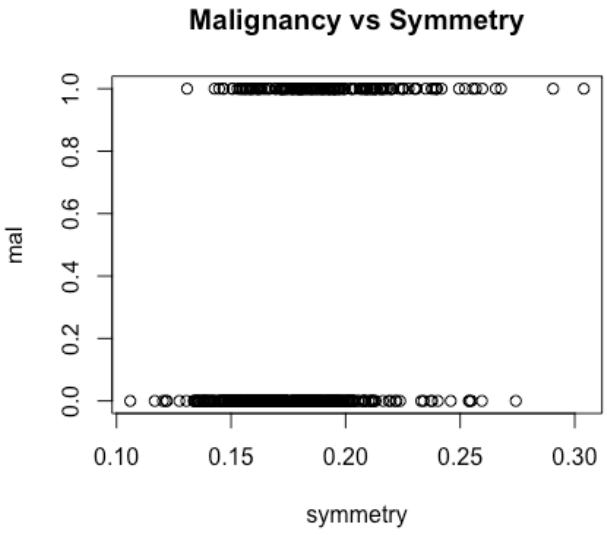

```
Call:
glm(formula = mal ~ concavity, family = binomial)
Deviance Residuals:
    Min       1Q   Median       3Q      Max
-4.7647  -0.4612  -0.2912   0.3293   2.4310
Coefficients:
             Estimate  Std. Error  z value  Pr(>|z|)
(Intercept)  -3.7850    0.2914   -12.99   <2e-16 ***
concavity    36.8457    3.0314    12.15   <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: 383.23  on 567  degrees of freedom
AIC: 387.23
Number of Fisher Scoring iterations: 6
```



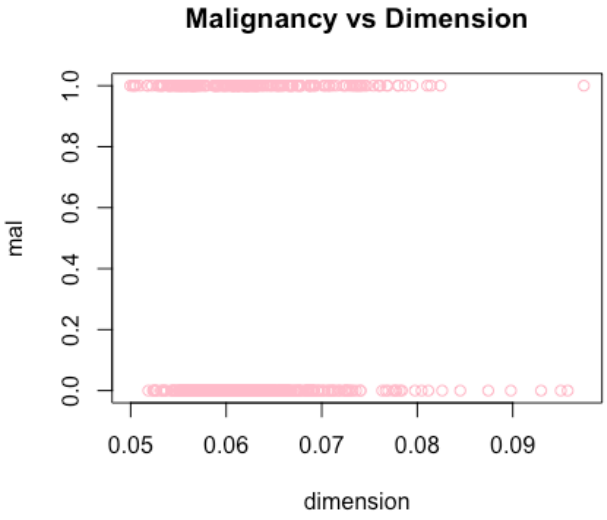

```
> plot(points, mal, main="Malignancy vs Points",
col="yellow")
> modelpoint <- glm(mal~points,family=binomial)
> summary(modelpoint)
Call:
glm(formula = mal ~ points, family = binomial)
Deviance Residuals:
    Min       1Q   Median       3Q      Max
-2.5783  -0.3251  -0.1532   0.1573   2.7187
Coefficients:
             Estimate  Std. Error z value Pr(>|z|)
```

```
(Intercept)  -5.8437    0.4803  -12.17   <2e-16 ***
points      106.9937    9.1190   11.73   <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: 258.92  on 567  degrees of freedom
AIC: 262.92
Number of Fisher Scoring iterations: 7
```

```
> plot(symmetry, mal, main="Malignancy vs Symmetry", col="black")
> modelsym <- glm(mal~symmetry,family=binomial)
> summary(modelsym)
```




```
Call:
glm(formula = mal ~ symmetry, family = binomial)
Deviance Residuals:
    Min       1Q   Median       3Q      Max
-2.0637  -0.9187  -0.6879   1.1674   2.0446
Coefficients:
             Estimate  Std. Error   z value   Pr(>|z|)
(Intercept)   -5.5689    0.6981   -7.977    1.50e-15 ***
symmetry     27.6042  3.7655    7.331     2.29e-13 ***
---
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: 686.80  on 567  degrees of freedom
AIC: 690.8
Number of Fisher Scoring iterations: 4
```




```
> plot(dimension, mal, main="Malignancy vs
Dimension", col="pink")
> modeldim <- glm(mal~dimension,family=binomial)
> summary(modeldim)
Call:
glm(formula = mal ~ dimension, family = binomial)
Deviance Residuals:
    Min       1Q   Median       3Q      Max
-0.9816  -0.9692  -0.9598   1.3985   1.4639
```

Coefficients:

	Estimate	Std. Error	z value	Pr(>|z|)
(Intercept)	-0.2837	0.7799	-0.364	0.716
dimension	-3.7819	12.3519	-0.306	0.759

(Dispersion parameter for binomial family taken to be 1)
Null deviance: 751.44 on 568 degrees of freedom
Residual deviance: 751.35 on 567 degrees of freedom
AIC: 755.35
Number of Fisher Scoring iterations: 4

If we were to use a single predictors variable, we can end up with the variable that doesn't predict malignancy best. For example, based on the graphs produced above, while predictor dimension produces almost parallel lines of data and its P-value isn't significant at the .05 level, that of the other predictors have much more significant P-values and their graphs almost look like the S shape that we want for our logistic regression model. Also, the more predictor variables we have, when putting them together in a regression model, the more predictors that will appear to be more significant than others and therefore, improving our diagnosis's accuracy.

#4

```
> install.packages("gmodels")

> train_case <- mydata[1:469, ]
> test_case <- mydata[470:569, ]

> tumordata <- data.frame (mal, radius, texture, smoothness, concavity, compactness, points, symmetry, dimension)

> model1 <- glm(tumordata$mal~tumordata$radius+tumordata$texture+tumordata$smoothness+tumordata$points+tumordata$concavity+tumordata$compactness+tumordata$symmetry+tumordata$dimension,family=binomial, data=train_case)
> summary(model1)

Call:
glm(formula = tumordata$mal ~ tumordata$radius + tumordata$texture +
    tumordata$smoothness + tumordata$points + tumordata$concavity +
    tumordata$compactness + tumordata$symmetry + tumordata$dimension,
    family = binomial, data = train_case)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-2.10517  -0.13757  -0.03442   0.01821   3.02582
```

Coefficients:

	Estimate	Std. Error	z value	Pr(>|z|)	
(Intercept)	-33.44994	7.95034	-4.207	2.58e-05	***
tumordata$radius	0.98884	0.23882	4.141	3.46e-05	***
tumordata$texture	0.38180	0.06284	6.076	1.24e-09	***
tumordata$smoothness	75.75837	33.12887	2.287	0.0222	*
tumordata$points	58.57517	28.27231	2.072	0.0383	*
tumordata$concavity	14.93512	8.24096	1.812	0.0699	.
tumordata$compactness	-16.58939	12.96410	-1.280	0.2007	
tumordata$symmetry	18.28180	10.90985	1.676	0.0938	.
tumordata$dimension	-26.44212	84.03225	-0.315	0.7530	

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: 151.69 on 560 degrees of freedom
AIC: 169.69

Number of Fisher Scoring iterations: 8

```
> anova(model1, test = "Chisq")
```

Analysis of Deviance Table
Model: binomial, link: logit
Response: tumordata$mal
Terms added sequentially (first to last)

	Df	Deviance	Resid. Df	Resid. Dev	Pr(>Chi)	
NULL			568	751.44		
tumordata$radius	1	421.43	567	330.01	< 2.2e-16	***
tumordata$texture	1	38.89	566	291.12	4.489e-10	***
tumordata$smoothness	1	103.83	565	187.29	< 2.2e-16	***
tumordata$points	1	26.97	564	160.32	2.066e-07	***
tumordata$concavity	1	1.98	563	158.34	0.15907	
tumordata$compactness	1	3.74	562	154.60	0.05325	.
tumordata$symmetry	1	2.81	561	151.79	0.09376	.
tumordata$dimension	1	0.10	560	151.69	0.75237	

Signif. codes: 0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

As shown in model 1, tumordata$radius, tumordata$texture, tumordata$smoothness, and tumordata$points are significant variables with really low p values. Therefore, we create model2 with only these 4 variables.

```
> model2 <-
glm(tumordata$mal~tumordata$radius+tumordata$texture+tumordata$smoothness+tumordata$
points,family=binomial, data=train_case)
> summary(model2)
Call:
glm(formula = tumordata$mal ~ tumordata$radius + tumordata$texture +
    tumordata$smoothness + tumordata$points, family = binomial,
    data = train_case)

Deviance Residuals:
     Min        1Q    Median        3Q       Max
-2.42132  -0.15010  -0.04247   0.02603   2.86598

Coefficients:
                          Estimate     Std. Error    z value   Pr(>|z|)
(Intercept)               -28.57552    4.81406    -5.936     2.92e-09 ***
tumordata$radius          0.85081    0.17112     4.972      6.63e-07 ***
tumordata$texture         0.35845    0.05985    5.990       2.10e-09 ***
tumordata$smoothness  52.26403  26.08496   2.004   0.0451 *
tumordata$points          78.73692   16.59332   4.745     2.08e-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: 160.32  on 564  degrees of freedom
AIC: 170.32

Number of Fisher Scoring iterations: 8

> anova(model2, test = "Chisq")
Analysis of Deviance Table
Model: binomial, link: logit
Response: tumordata$mal
Terms added sequentially (first to last)

                         Df   Deviance Resid. Df Resid. Dev   Pr(>Chi)
NULL                                       568     751.44
```

```
tumordata$radius         1   421.43    567     330.01    < 2.2e-16 ***
tumordata$texture        1    38.89    566     291.1     4.489e-10 ***
tumordata$smoothness     1   103.83    565     187.29     < 2.2e-16 ***
tumordata$points         1    26.97     564     160.32    2.066e-07 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

However, AIC increases in model2 (from 169.69 to 170.32). Therefore, we create model3 with all variables except for tumordata$dimention because of its strongly insignificant p value (0.75).

```
> model3 <-
glm(tumordata$mal~tumordata$radius+tumordata$texture+tumordata$smoothness+tumordata$
points+tumordata$concavity+tumordata$compactness+tumordata$symmetry,family=binomial,
data=train_case)
> summary(model3)

Call:
glm(formula = tumordata$mal ~ tumordata$radius + tumordata$texture +
    tumordata$smoothness + tumordata$points + tumordata$concavity +
    tumordata$compactness + tumordata$symmetry, family = binomial,
    data = train_case)

Deviance Residuals:
     Min        1Q    Median        3Q       Max
-2.05762  -0.13809  -0.03481   0.01803   3.00137

Coefficients:
                           Estimate     Std. Error    z value    Pr(>|z|)
(Intercept)              -35.14941     5.91489   -5.943      2.81e-09 ***
tumordata$radius          1.02607     0.20893     4.911      9.06e-07 ***
tumordata$texture         0.38373     0.06265     6.125      9.06e-10 ***
tumordata$smoothness     73.08878    32.05063   2.280      0.0226 *
tumordata$points         61.04882     27.24291   2.241      0.0250 *
tumordata$concavity      13.80627     7.45887     1.851      0.0642 .
tumordata$compactness -19.41566    9.40834   -2.064     0.0390 *
tumordata$symmetry       18.37705    10.92109     1.683      0.0924 .
---
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: 151.79  on 561  degrees of freedom
```

AIC: 167.79

Number of Fisher Scoring iterations: 8

```
> anova(model3, test = "Chisq")
Analysis of Deviance Table
Model: binomial, link: logit
Response: tumordata$mal
Terms added sequentially (first to last)

                         Df Deviance Resid. Df Resid. Dev  Pr(>Chi)
NULL                                       568     751.44
tumordata$radius          1   421.43       567     330.01    < 2.2e-16***
tumordata$texture         1    38.89       566     291.12    4.489e-10***
tumordata$smoothness      1   103.83       565     187.29    < 2.2e-16***
tumordata$points          1    26.97       564     160.32    2.066e-07***
tumordata$concavity       1     1.98       563     158.34    0.15907
tumordata$compactness     1     3.74       562     154.60    0.05325 .
tumordata$symmetry        1     2.81       561     151.79    0.09376 .
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

Model3 shows the best AIC value of 167.79 (lowest compared to model1 and model2). We choose this logistic regression model to produce the confusion matrix below.

```
> library(gmodels)
> pred_case <- predict(model3, newdata = test_case, type="response")
> pre <- ifelse(pred_case > .5, 1, 0)
> CrossTable(tumordata$mal[470:569],pre[1:100])

   Cell Contents
|-------------------------------|
|                             N |
| Chi-square contribution |
|           N / Row Total |
|           N / Col Total |
|         N / Table Total |
|-------------------------------|


Total Observations in Table:  100
```

	pre[1:100]		
tumordata$mal[470:569]	0	1	Row Total
0	44	17	61
	0.087	0.193	
	0.721	0.279	0.610
	0.638	0.548	
	0.440	0.170	
1	25	14	39
	0.136	0.302	
	0.641	0.359	0.390
	0.362	0.452	
	0.250	0.140	
Column Total	69	31	100
	0.690	0.310	

We classify 44 + 14 = 58 of the 100 test cases correctly, for an accuracy rate of 58%, and consequently an error rate of 42%. To predict malignancy more easily, we change the lower cutoff points to 0.4; 0.3; and 0,2.

The 0.4 cutoff point lowers the accuracy rate to 57% with 25 fail predictions of malignancy unchanged. This indicates that the s curve is skewed more to no malignancy, so lower cutoff point 0.3 might increase the predictability for malignant patients as shown below.

```
> pre <- ifelse(pred_case > .3, 1, 0)
> CrossTable(tumordata$mal[470:569],pre[1:100])
```

```
   Cell Contents
|-------------------------|
|                       N |
| Chi-square contribution |
|           N / Row Total |
|           N / Col Total |
|         N / Table Total |
|-------------------------|
```

Total Observations in Table: 100

	pre[1:100]		
tumordata$mal[470:569]	0	1	Row Total
0	43	18	61
	0.402	0.714	
	0.705	0.295	0.610
	0.672	0.500	
	0.430	0.180	
1	21	18	39
	0.628	1.117	
	0.538	0.462	0.390
	0.328	0.500	
	0.210	0.180	
Column Total	64	36	100
	0.640	0.360	

The confusion matrix of 0.3 cutoff point shows higher accuracy rate (61%) with lower fail prediction for malignancy (21) and small changes in the correct prediction of benignity (44 to 43). However, much lower cutoff point 0.2 has a lower accuracy rate (59%) due to 3 less correct benignity predictions compared to the original cut off point.

In conclusion, a cutoff point of 0.3 produces the best result.