Tuan Dang
04/29/2021

Project 3: Finding the Best Logistic Regression Model to Detect Cancerous Tumors

Part 1.

Summary statistics for each variable:

```
table(data$diagnosis)
  B   M
357 212
```

```
summary(data$radius)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
  6.981  11.700  13.370  14.127  15.780  28.110
```

```
summary(data$texture)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
   9.71   16.17   18.84    19.29   21.80    39.28
```

```
summary(data$perimeter)
   Min. 1st Qu.  Median    Mean    3rd Qu.    Max.
  43.79   75.17   86.24     91.97    104.10  188.50
```

```
summary(data$area)
   Min. 1st Qu.  Median    Mean    3rd Qu.    Max.
  143.5   420.3   551.1     654.9      782.7  2501.0
```

```
summary(data$smoothness)
   Min.        1st Qu.  Median    Mean     3rd Qu.    Max.
0.05263    0.08637    0.09587   0.09636   0.10530 0.16340
```

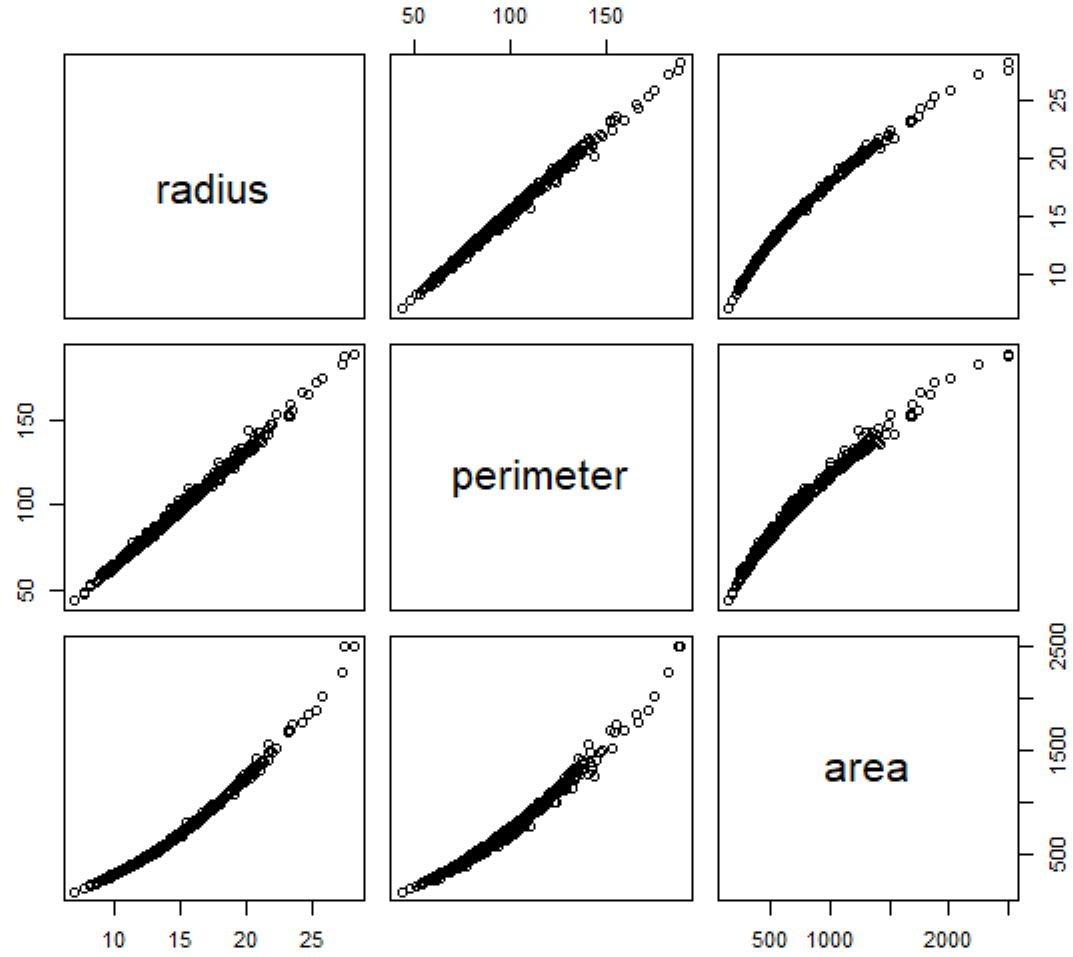
```
summary(data$compactness)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
```

```
0.01938 0.06492 0.09263 0.10434 0.13040 0.34540

summary(data$concavity)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
0.00000 0.02956 0.06154 0.08880 0.13070 0.42680

summary(data$points)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
0.00000 0.02031 0.03350 0.04892 0.07400 0.20120

summary(data$symmetry)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
 0.1060  0.1619  0.1792  0.1812  0.1957  0.3040

summary(data$dimension)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
0.04996 0.05770 0.06154 0.06280 0.06612 0.09744
```

Part 2.

```
pairs(data[c("radius","perimeter","area")])
```

Scatterplot Matrix of Radius, Perimeter, and Area



From these scatter plots, we can see that there is a clear relationship between radius, perimeter, and area, which makes sense since radius can be used to calculate both of the latter variables (assuming that the tumors are spherical). This leads to a strong linear relationship between radius and perimeter, and a strong quadratic relationship between radius and area. Similar to the age and age squared variables from Project 2, I predict that if we include all of these variables together in a model, it would show that area has a much larger significance than radius and perimeter. I'm also predicting that we can exclude perimeter, and possibly radius, from the model while keeping area, because the perimeter is essentially the same as radius multiplied by a constant coefficient, so it would be redundant in a model.

Should use cor.test() here.

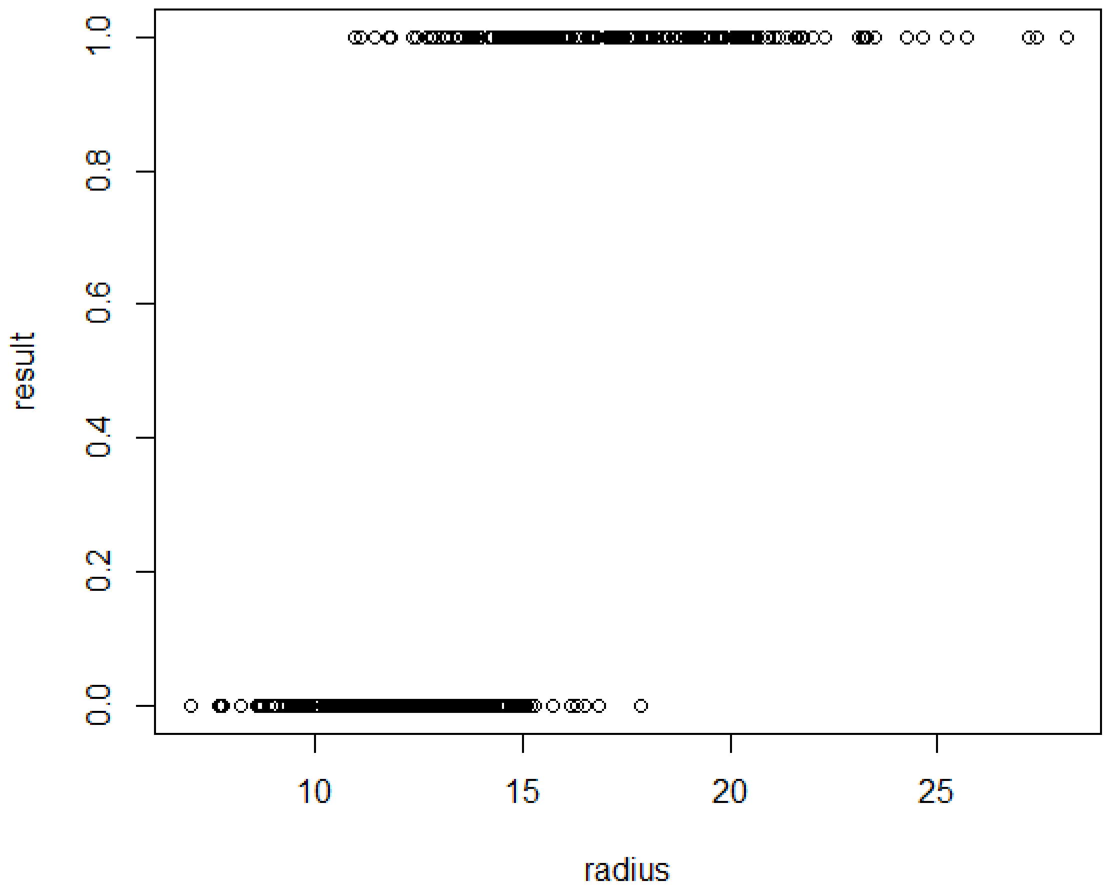
Part 3.

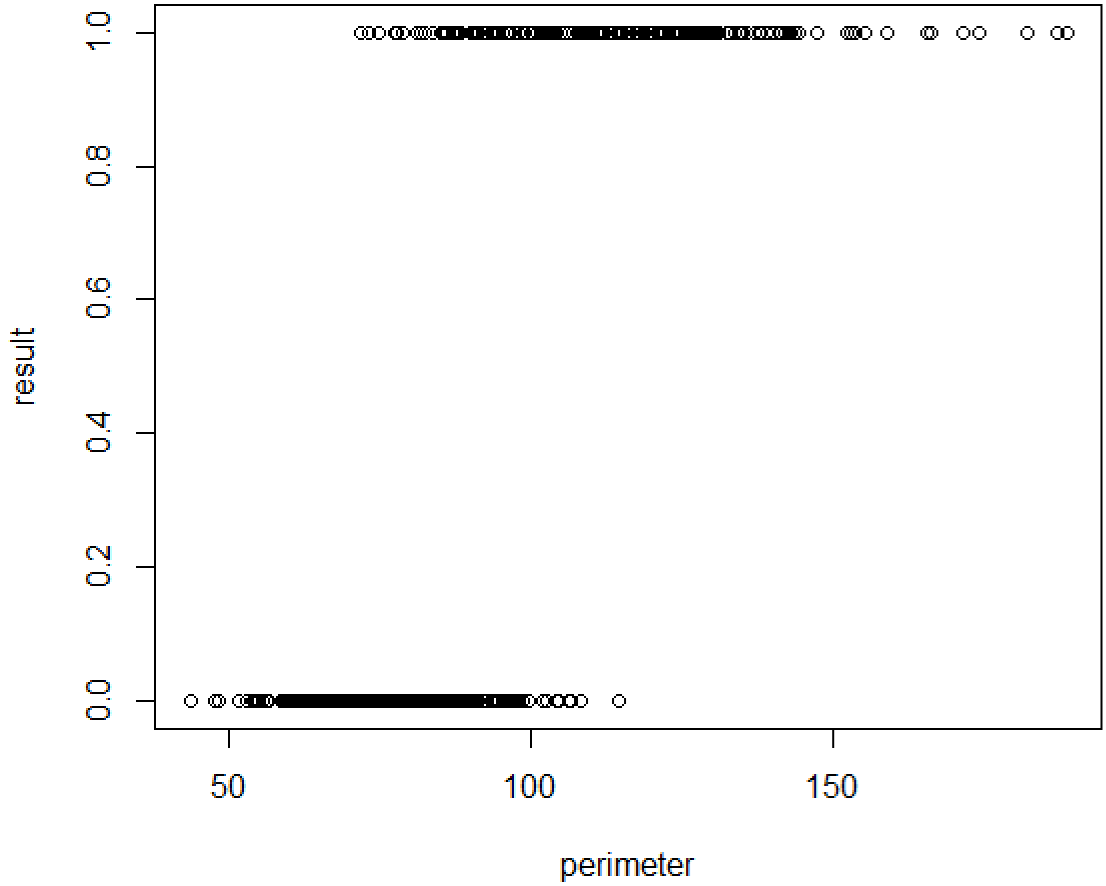
```
data$result <- ifelse(data$diagnosis == "M", 1, 0)
```

```
table(data$result)
  0   1
357 212
```

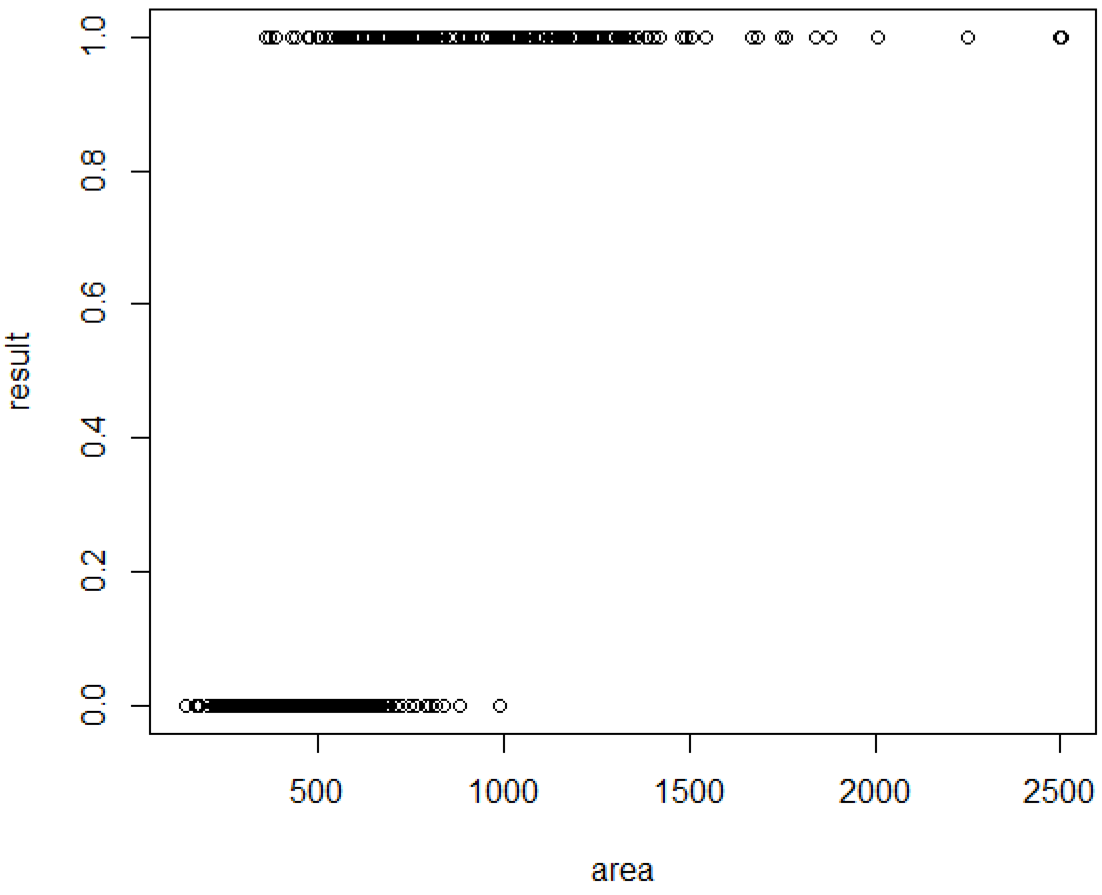
I decided to produce graphs for all three variables from question 2 to see what they would look like.

```
plot(result ~ radius, data=data)
plot(result ~ perimeter, data=data)
plot(result ~ area, data=data)
```


result
0.0
0.2
0.4
0.6
0.8
1.0
10
15
20
25
radius



1.0
0.8
0.6
0.4
0.2
0.0
result
50
100
150
perimeter



As expected from part 2, radius and perimeter have an almost identical effect on the response, while the area looks a bit different. It is not very clear which variable is the best for the model. In fact, all of them would be horrible at predicting the response if they were the only predictors in the model, seeing that there is so much overlap between the 0 and 1 dots. For example, in the scatter plot between area and result, we can see that when the area is between 500 and 1000, the dots can appear in either column, which would make it hard to find a line that would help predict the tumors.

Part 4.

R Command:

```
data_train <- data[1:469, ]
data_test <- data[470:569, ]
```

For the first model, I wanted to use all the available variables to see what that would look like and to have a baseline to improve the model.

***Model 1**

```
model1 <- glm(result ~ radius + texture + perimeter + area + smoothness + compactness + concavity + points + symmetry + dimension, data = data_train, family = binomial)

summary(model1)
```

```
Call:
glm(formula = result ~ radius + texture + perimeter + area +
    smoothness + compactness + concavity + points + symmetry +
    dimension, family = binomial, data = data_train)

Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept)  -6.57840   14.48977  -0.454   0.6498
radius       -4.75549    3.93868  -1.207   0.2273
texture       0.36996    0.06954   5.320 1.04e-07 ***
perimeter     0.39859    0.54856   0.727   0.4675
area          0.03418    0.01880   1.819   0.0690 .
smoothness   80.85450   36.11867   2.239   0.0252 *
compactness -19.66762   22.22589  -0.885   0.3762
concavity     7.33147    8.90268   0.824   0.4102
points       73.67390   31.18390   2.363   0.0181 *
symmetry     11.38851   11.65115   0.977   0.3283
```

```
dimension   -66.53464   94.81830  -0.702   0.4829
---

    Null deviance: 617.53  on 468  degrees of freedom
Residual deviance: 120.74  on 458  degrees of freedom
AIC: 142.74
```

From the summary, we can see that some variables stand out as being more significant than others (i.e. texture, area, smoothness, points), which gives us an idea of what variables we can include or exclude to improve the model. We will also see the ANOVA to investigate further.

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

Analysis of Deviance Table

            Df Deviance Resid. Df Resid. Dev  Pr(>Chi)
NULL                          468     617.53
radius       1   348.75       467     268.79 < 2.2e-16 ***
texture      1    33.55       466     235.23 6.935e-09 ***
perimeter    1    55.05       465     180.18 1.174e-13 ***
area         1     4.12       464     176.07  0.042464 *
smoothness   1    35.74       463     140.33 2.258e-09 ***
compactness  1     1.75       462     138.58  0.186275
concavity    1     9.47       461     129.11  0.002092 **
points       1     6.80       460     122.32  0.009138 **
symmetry     1     1.08       459     121.24  0.298000
dimension    1     0.50       458     120.74  0.479796
```

From the ANOVA of model 1, we can see that the four variables mentioned earlier are still significant, although it is interesting to see that other variables also popped up as well. Suddenly, radius and perimeter seem more significant than area, which is conflicting with the Wald tests from the model summary. Concavity is a new variable that appeared as significant. For model 2, I will include all of these variables together to see what happens.

*Model 2

```
model2 <- glm(result ~ radius + perimeter + texture + area + smoothness +
concavity + points, data = data_train, family = binomial)

Call:

glm(formula = result ~ radius + perimeter + texture + area +
    smoothness + concavity + points, family = binomial, data = data_train)

Coefficients:
             Estimate Std. Error z value Pr(>|z|)
(Intercept) -10.05756   10.97195  -0.917  0.35932
radius       -0.66157    2.70619  -0.244  0.80687
perimeter    -0.21255    0.30378  -0.700  0.48412
texture       0.36691    0.06826   5.375 7.65e-08 ***
area          0.03346    0.01635   2.047  0.04062 *
smoothness   61.21069   30.19820   2.027  0.04267 *
concavity     2.54766    7.69032   0.331  0.74043
points       82.58062   29.89174   2.763  0.00573 **
---
    Null deviance: 617.53  on 468  degrees of freedom
Residual deviance: 124.36  on 461  degrees of freedom
AIC: 140.36
```

Excluding the other variables did help lower the AIC, which is good for the model's ability to make predictions. The summary is similar to model 1; we still see that texture, area, smoothness, and points stand out from the rest.

```
anova(model2, test = "Chisq")
Analysis of Deviance Table

           Df Deviance Resid. Df Resid. Dev  Pr(>Chi)
NULL                         468     617.53
radius      1   348.75       467     268.79 < 2.2e-16 ***
perimeter   1    63.35       466     205.44 1.732e-15 ***
texture     1    25.26       465     180.18 5.019e-07 ***
area        1     4.12       464     176.07  0.042464 *
smoothness  1    35.74       463     140.33 2.258e-09 ***
concavity   1     7.70       462     132.63  0.005537 **
points      1     8.28       461     124.36  0.004019 **
```

The ANOVA for model 2 is also very similar to model 1, which does not give us any new information. For model 3, I decided to stick to the four "main" variables that we have seen so far, and this turned out to produce the best model by far. Before I reached this conclusion, I did try many other models by modifying which variables to include (e.g. changing between radius, perimeter, and area to see which one works best).

***Model 3**

```
model3 <- glm(result ~ texture + area + smoothness + points, data = data_train,
family = binomial)
```

```
Call:
glm(formula = result ~ texture + area + smoothness + points,
    family = binomial, data = data_train)
```

```
Coefficients:
             Estimate Std. Error z value Pr(>|z|)
(Intercept) -23.255151   4.146513  -5.608 2.04e-08 ***
texture       0.355079   0.066553   5.335 9.54e-08 ***
area          0.009870   0.002159   4.572 4.83e-06 ***
smoothness   59.858381  27.361664   2.188   0.0287 *
points       74.869573  17.288830   4.331 1.49e-05 ***
---

    Null deviance: 617.53  on 468  degrees of freedom
Residual deviance: 128.05  on 464  degrees of freedom
AIC: 138.05
```

Model 3 is the best one that I could find where all the variables are significant and we are keeping the number of predictors to a minimum. This also gave the lowest AIC value (138.05), which ideally means this is better at making predictions compared to the previous models.

```
anova(model3, test = "Chisq")
Analysis of Deviance Table

           Df Deviance Resid. Df Resid. Dev  Pr(>Chi)
NULL                         468     617.53
texture     1   89.018       467     528.52 < 2.2e-16 ***
area        1  294.179       466     234.34 < 2.2e-16 ***
smoothness  1   84.426       465     149.91 < 2.2e-16 ***
points      1   21.864       464     128.05 2.927e-06 ***
```

Lastly, I wanted to include an LR test to make sure that we eliminated the right variables. This ANOVA command includes model 3 and model 1 since model 1 includes every variable. The p-value that we get is very large, which means we can not reject the null hypothesis, meaning that the excluded variables were probably insignificant.

```
anova(model3, model1, test = "Chisq")
Analysis of Deviance Table

Model 1: result ~ texture + area + smoothness + points
Model 2: result ~ radius + texture + perimeter + area + smoothness + compactness
+ concavity + points + symmetry + dimension

  Resid. Df Resid. Dev Df Deviance Pr(>Chi)
1       464     128.05
2       458     120.74  6     7.31   0.2931
```

***Creating a confusion matrix for the best model that we found:**

```
install.packages("gmodels")
library(gmodels)
data_predictions <- predict(model3, newdata = data_test, type = "response")
pre <- ifelse(data_predictions > 0.5, 1, 0)
CrossTable(data$result[470:569], pre[1:100])
```

```
Total Observations in Table:  100

                     | pre[1:100]
data$result[470:569] |         0 |         1 | Row Total |
---------------------|-----------|-----------|-----------|
                   0 |        60 |         1 |        61 |
                     |    10.444 |    19.397 |           |
                     |     0.984 |     0.016 |     0.610 |
                     |     0.923 |     0.029 |           |
                     |    0.600 |    0.010 |          |
---------------------   |----------|----------|----------|
                      1 |        5 |       34 |       39 |
                     |   16.336 |   30.339 |          |
                     |    0.128 |    0.872 |    0.390 |
                     |    0.077 |    0.971 |          |
                     |    0.050 |    0.340 |          |
---------------------   |----------|----------|----------|
           Column Total |       65 |       35 |      100 |
                     |    0.650 |    0.350 |          |
---------------------   |----------|----------|----------|
```

When we pick the standard cutoff probability of 0.5, we get a very good success rate of 94% ((60+34)/100). I experimented with other cutoff probabilities, but couldn't find anything better than 94%, although I found something else interesting. When we pick 0.35 as our cutoff probability, the success rate is also 94% ((58+36)/100) as shown in the cross table below. However, the probability that our model would make the mistake of predicting a cancerous tumor as being benign is lower (3% compared to 5%), which is what we want because we wouldn't want to tell someone that they don't have cancer when in actuality they do. So, although in either case, the success rate is 94%, we might want to choose 0.35 as the cutoff probability instead.

Note: I also tried 0.1, which lowered the success rate down to 89%, but the chances of making this error become 0%. So, I suppose there is an ethical argument to be made that this model would help save more patients.

```
pre <- ifelse(data_predictions > 0.35, 1, 0)
CrossTable(data$result[470:569], pre[1:100])
```

```
Total Observations in Table:  100
```

```
                    | pre[1:100]
data$result[470:569] |          0 |          1 |Row Total|
--------------------- |-----------|-----------|-----------|
                   0 |         58 |          3 |        61 |
                     |    11.616 |    18.168|           |
                     |     0.951 |     0.049 |     0.610|
                     |     0.951 |     0.077 |           |
                     |     0.580 |     0.030 |           |
--------------------- |-----------|-----------|-----------|
                   1 |          3 |         36 |        39 |
                     |    18.168 |    28.417|           |
                     |     0.077 |     0.923 |     0.390|
                     |     0.049 |     0.923 |           |
                     |     0.030 |     0.360 |           |
--------------------- |-----------|-----------|-----------|
        Column Total  |         61 |         39 |       100 |
                     |     0.610 |     0.390 |           |
--------------------- |-----------|-----------|-----------|
```