Group 21

STAC51 CATEGORICAL DATA ANALYSIS CASE STUDY

Topic: What Attributes Presented in our Dataset Would Help Predict the Event of No Relapse of Cervical Cancer?

Author(s): Ankit Jhurani, David Do, John Roy Jarlos, Sampson Liang

University of Toronto Scarborough | Winter 2021

Assigned jobs and descriptions:

- Ankit Jhurani – 1003697065

Introduction, Hypothesis and Initial Data Analysis, Presentation wrote the introduction, concluded the hypothesis, performed the initial data analysis, helped in finding the final model and made the presentation.

- Sampson Liang - 1005348474

Data Cleaning - performed basic drop.na functions and categorizing all the categorical variables with their respective defined descriptions that has some discrepancies in the listed values of some of the parameters

- David Do - 1004440009

Process in finding the final model - performed the full model reduced model approach in finding the suitable model for our data set including Step AIC function to find the final model

- John Roy Jarlos – 1004916781

Model Validation - performed model validation including any remedial measures that would make the final model suitable.

UNIVERSITY OF TORONTO SCARBOROUGH, STAC51 WINTER 2021

Case Study: On Predicting the Survival of Patients Following Surgeries Related to Cervical Cancer

Table of Contents

1.Introduction

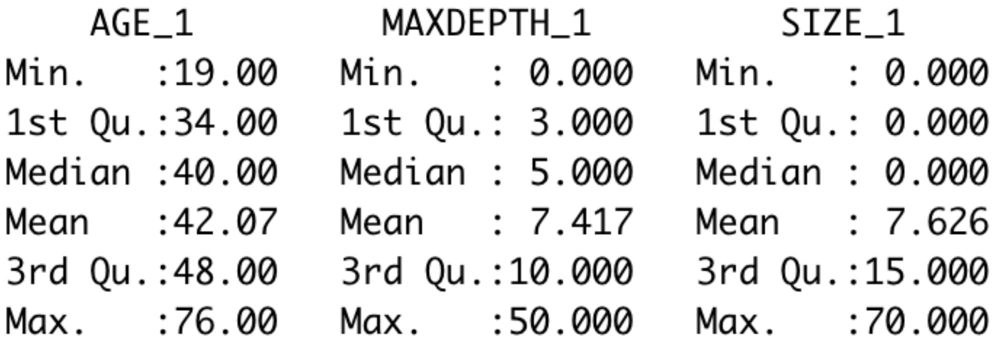
Cancer as we know it is a disease caused when cells divide uncontrollably and spread into surrounding tissues. Now when it starts in the Cervix, it is known as cervical cancer. The cervix connects the vagina to the upper part of the uterus. This is the portion of the women where a baby grows when she is pregnant. It is a known fact that all women are at risk and it occurs most often in women over the age of 30. ("Basic information about cervical cancer," 2021) It is the second most common type of cancer in women worldwide. Most of the cases occur in less developed countries where no effective screening systems are available. In high income countries, cervical cancer incidents have more than halved over the past 30 years since the introduction of formal screening programmes. Women in their early stages of tumor can be cured, although long term morbidity from treatment is common. (Waggoner, 2003) If complications do occur then radical hysterectomy or chemoradiation, or a combination of both are used to treat these complications from the early stages. ("Cervical cancer," 2019) However, even with these options available, the study on how to reduce the continuing effects of cervical cancer beyond the first stage is still far from being solved. Furthermore, treatment of recurrent cervical cancer seems to remain ineffective. Quality of life should be taken into account in treatment of women with primary and recurrent cervical cancer. (Waggoner, 2003)

2.Hypothesis

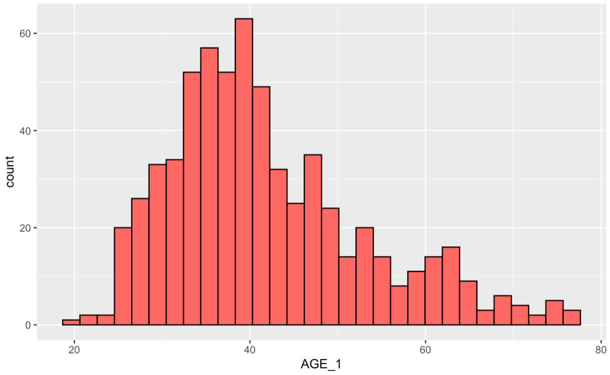
Every human being is different and predicting reasons for relapse varies and is inconsistent from patient to patient. From the data collected, 905 cervical cancer patients are considered. There are different attributes that could contribute to relapse. However, they are hard to determine and then create a reasonable solution. This paper's purpose is to filter out attributes from our data that contribute to no relapse of cervical cancer after surgeries needed to treat this disease such as Radical hysterectomy and radiation therapy. This purpose will be accomplished through various categorical and regression tests like: Generalized Linear Model, drop_na functions, Step AIC to finalize our final model and Model Validation techniques to check if the final model is specified. In addition to this, the paper will classify patients according to their individual risk of relapse.

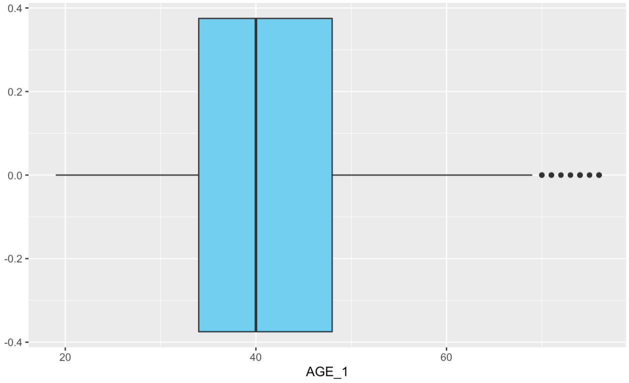
3.Initial Data Analysis

A data set consisting of 905 cervical cancer patients will be considered. In this data set, each patient is categorized by the following categorical variables: whether they have received radiation therapy or not, whether they test positive for capillary lymphatic spaces or not, status of the patient since their last follow up, level of cell differentiation, histology of the patient's cancer cells, amount of disease left after surgery, and pelvis lymph node involvement. Additionally, each patient has had their surgery date, age, tumor depth, size of tumor, date of recurrence, and last follow up date recorded. A summary statistic for the continuous variables along with graphs displaying the distributions of each of the categorical variables can be found below.

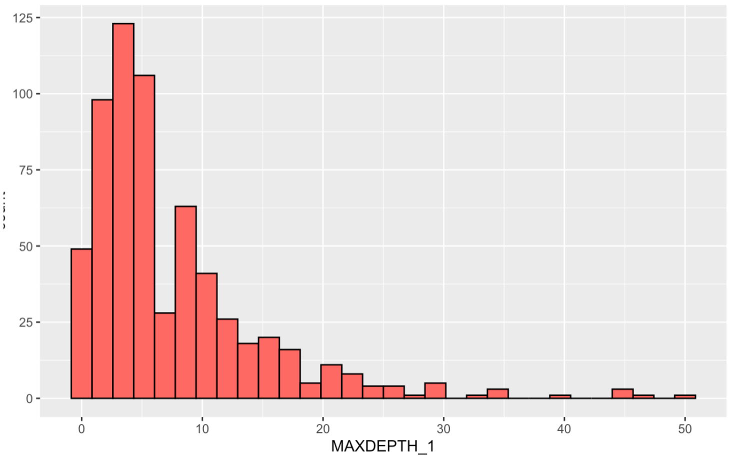


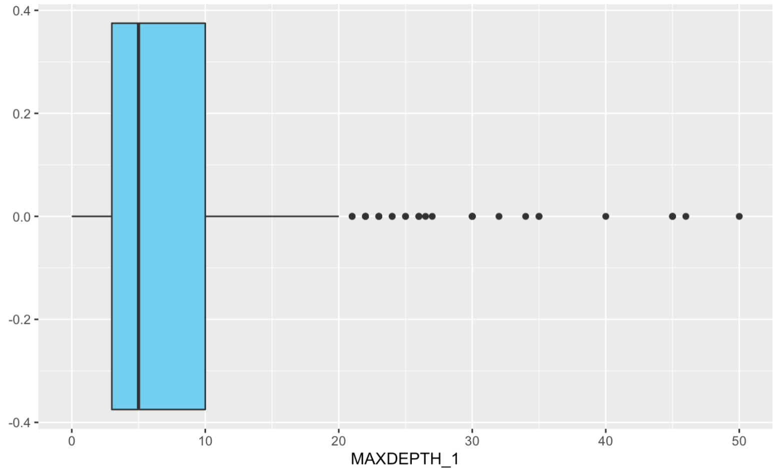
```
     AGE_1          MAXDEPTH_1         SIZE_1     
 Min.   :19.00   Min.   : 0.000   Min.   : 0.000  
 1st Qu.:34.00   1st Qu.: 3.000   1st Qu.: 0.000  
 Median :40.00   Median : 5.000   Median : 0.000  
 Mean   :42.07   Mean   : 7.417   Mean   : 7.626  
 3rd Qu.:48.00   3rd Qu.:10.000   3rd Qu.:15.000  
 Max.   :76.00   Max.   :50.000   Max.   :70.000  
```



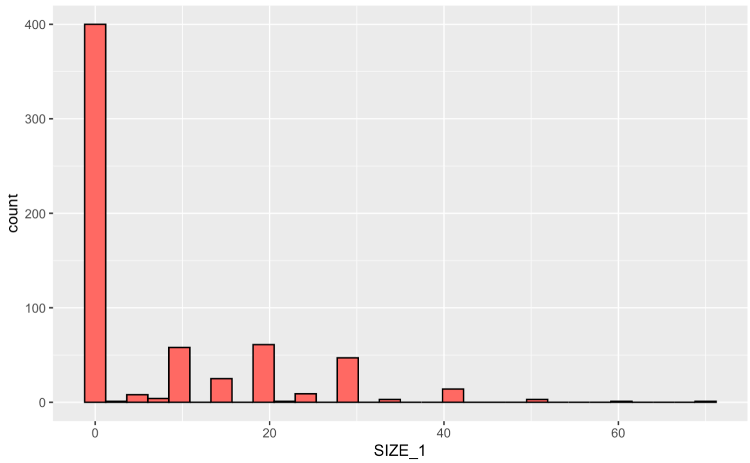


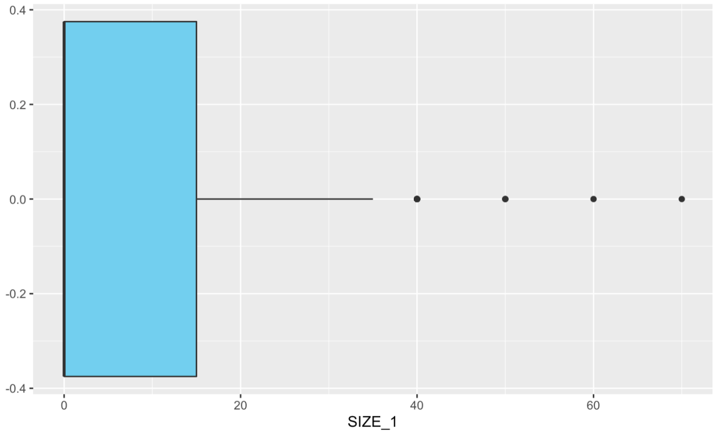
The summary statistics of each numerical variable are computed using the summary() function. AGE_1 gives us the age of each person in our dataset. The mean is at 42.07 years old, with the youngest at 19 years old and the oldest at 76 years old. The median for AGE_1 is at 40 years old.





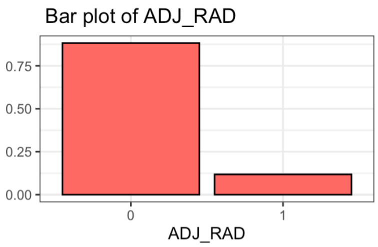
MAXDEPT_1 variable gives us the depth of the tumor in millimeters (mm). The dataset gives us a mean of 7.417, with zero as its minimum and 50 as its maximum. Its median is at 5.

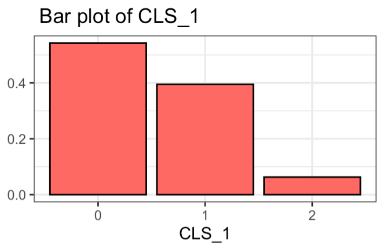


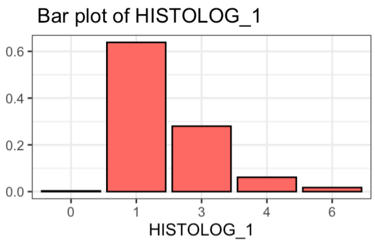


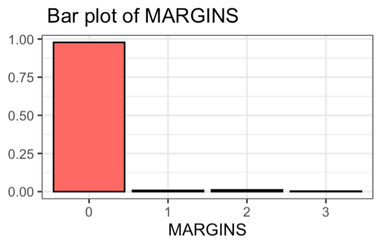
SIZE_1 variable gives us the size of each tumor in millimeters (mm) upon its diagnosis. The dataset gives us a mean of 7.626, minimum of zero meaning that the tumor could not be measured as it is diminutive, maximum at 70.

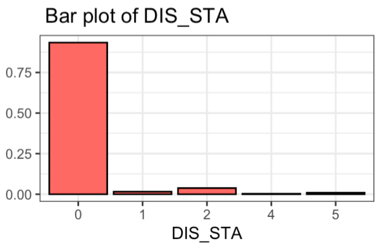
Bar Graph of Categorical Variables

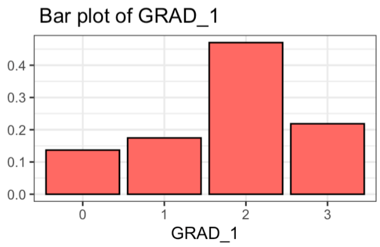

Bar plot of ADJ_RAD

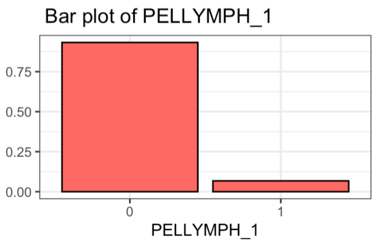

Bar plot of CLS_1

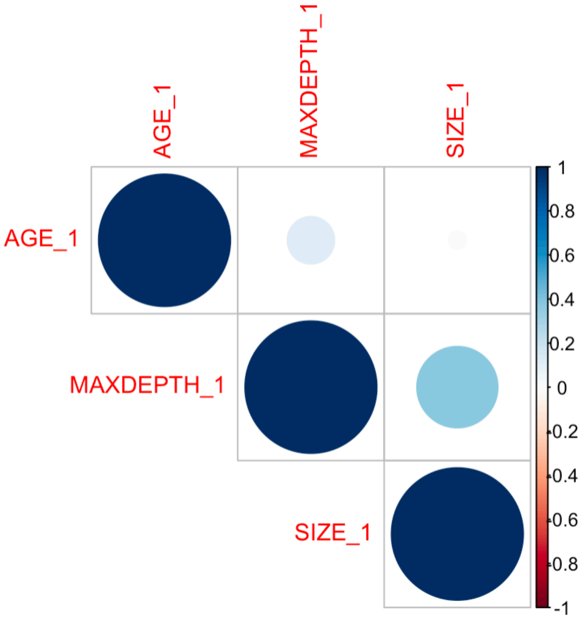

Bar plot of HISTOLOG_1


Bar plot of MARGINS


Bar plot of DIS_STA


Bar plot of GRAD_1

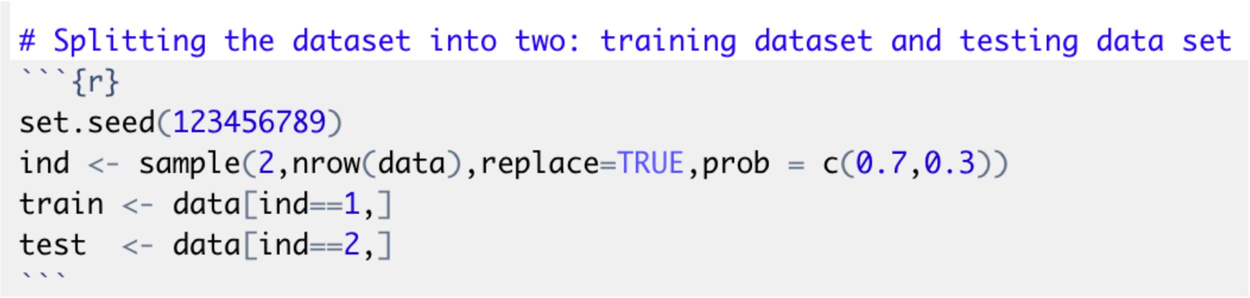

Bar plot of PELLYMPH_1



We used library "corplot" to use a graph that would give us the correlation of each numerical variables. Out of all the variables present in our data after the data cleaning, these are the variables that remain numerical and therefore, we were able to get the correlation of each variable to one another. For example, the correlation of MAXDEPTH_1 and SIZE_1 has a light blue color, and it indicates that these variables contain positive correlation between each other.

Data Cleaning

The data set has been modified such that patients with missing values are not considered and have been dropped from the final data set. A variable 'diff' has been added that records the number of years between each patient's surgery date and their last follow up date. Furthermore, columns representing categorical variables in the data set have been converted from a numeric variable into a factor in order to better operate within the tests we will be performing. Finally, the data has been sampled and split into both a training data set, and a testing data set for the purpose of our analysis.



~~~
# Splitting the dataset into two: training dataset and testing data set
```{r}
set.seed(123456789)
ind <- sample(2,nrow(data),replace=TRUE,prob = c(0.7,0.3))
train <- data[ind==1,]
test  <- data[ind==2,]
```
~~~

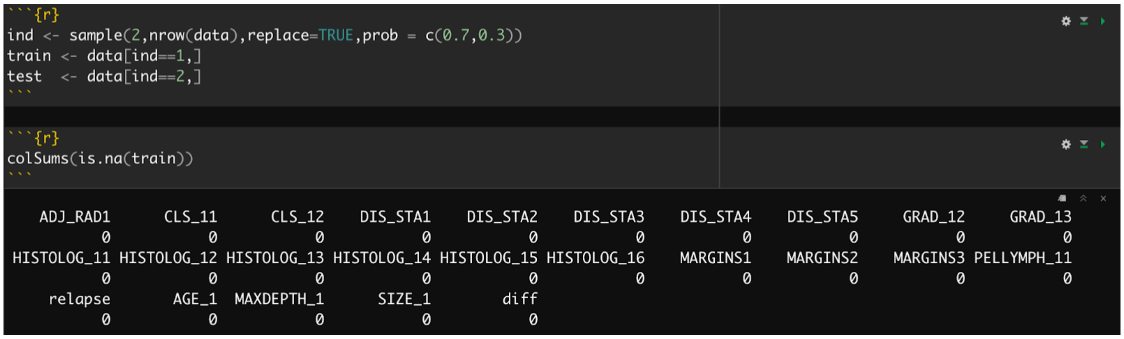
4.Process in Finding the Final Model

The process in finding our final model includes the clean dataset. This is because we want to reduce inaccuracies and discrepancies. Accuracy is needed especially in matters that pertain to human anatomy.

As explained earlier on how we obtained our clean data, we proceed to understand relapse of patients by calculating the difference between Surgery Date (SURGDATE) and Last Follow Up date (FU_DATE). The purpose of this is to determine whether or not patients have actually had cases of relapse. From the output it is clear that not all patients have had a recurrence date. This could be because the patient was actually cured, but we could also consider that the patient might have passed away maybe because of other issues not mentioned in the description of data. In addition to this, we conduct a one-hot encoding to convert an unordered categorical vector to multiple binarized vectors where each binary vector of 1's and 0's. Furthermore, for the numeric variables we first create a min-max function.

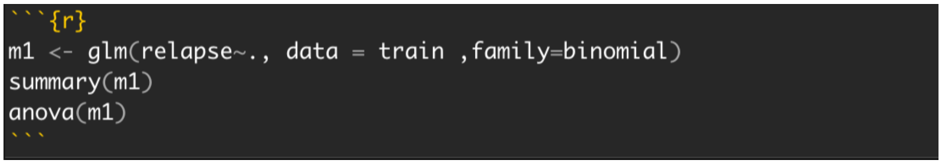
~~~
```{r}
minmax <- function(x) {
    return ((x - min(x)) / (max(x) - min(x)))
}
```
~~~

We then proceed to do the same one-hot encode for the numeric variables. This is then followed by dividing the cleaned dataset into training and testing data. Training data contains 70% of the data whereas the testing dataset contains 30%.



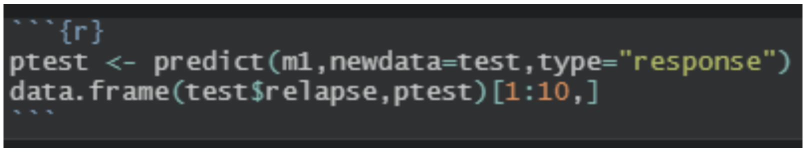
~~~
```{r}
ind <- sample(2,nrow(data),replace=TRUE,prob = c(0.7,0.3))
train <- data[ind==1,]
test  <- data[ind==2,]
~~~

```{r}
colSums(is.na(train))
```

   ADJ_RAD1   CLS_11   CLS_12   DIS_STA1   DIS_STA2   DIS_STA3   DIS_STA4   DIS_STA5   GRAD_12   GRAD_13
          0   0   0   0   0   0   0   0   0   0
HISTOLOG_11 HISTOLOG_12 HISTOLOG_13 HISTOLOG_14 HISTOLOG_15 HISTOLOG_16   MARGINS1   MARGINS2   MARGINS3 PELLYMPH_11
          0   0   0   0   0   0   0   0   0   0
    relapse   AGE_1 MAXDEPTH_1   SIZE_1   diff
          0   0   0   0   0
```

Now that we have our necessary and cleaned up inputs, we go ahead and run a Generalized Linear Model under the Binomial family using the training data.



```
```{r}
m1 <- glm(relapse~., data = train ,family=binomial)
summary(m1)
anova(m1)
```
```
```

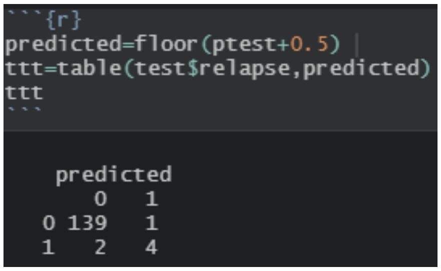
The output to this model provides us with an initial understanding of all the categorical variables that were used in this model. From the output we get that DIS_STA1, DIS_STA2 and MAXDEPTH_1 are important variables with significant information for predicting whether an individual will relapse or not.

We can now use this model to predict the response variable using the below code.

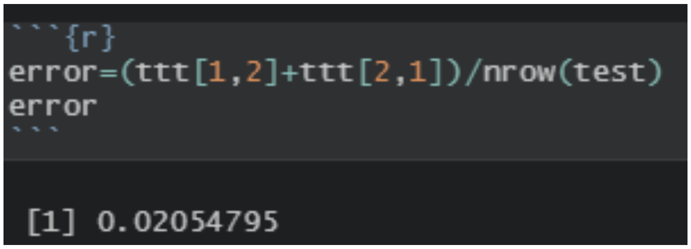


~~~
```{r}
ptest <- predict(m1,newdata=test,type="response")
data.frame(test$relapse,ptest)[1:10,]
```
~~~

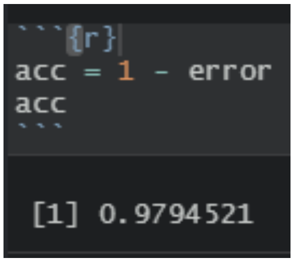
Afterwards, we can compare the predicted results to the actual results in a table and use the info to find the error and accuracy of the model, which turns out to be almost 98% accurate with 2% error.



~~~
```{r}
predicted=floor(ptest+0.5)
ttt=table(test$relapse,predicted)
ttt
```

   predicted
      0   1
  0 139   1
  1   2   4
~~~

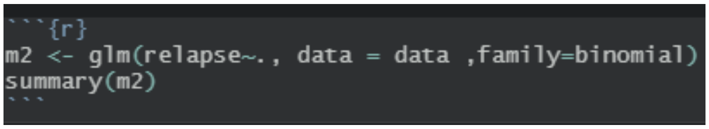


~~~
```{r}
error=(ttt[1,2]+ttt[2,1])/nrow(test)
error
```

[1] 0.02054795
~~~



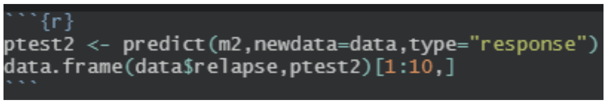
~~~
```{r}
acc = 1 - error
acc
```

[1] 0.9794521
~~~

Now, we make another Generalized Linear Model, but this time instead of the training data, we use the actual data. Looking at the results from the code below, we can see that the significant variables are DIS_STA1, DIS_STA2, and MAXDEPTH_1, which are the same variables as last time. The major difference between this model and the previous model is that the MAXDEPTH_1 variable is much more significant in this model than the previous one.

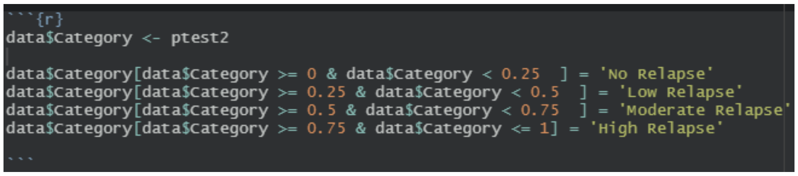


~~~
```{r}
m2 <- glm(relapse~., data = data ,family=binomial)
summary(m2)
```
~~~

Using this model, we can use the probability values the predict function gives in order to classify patients into the four following categories: “No relapse”, “Low Relapse”, “Moderate Relapse”, and “High Relapse”.

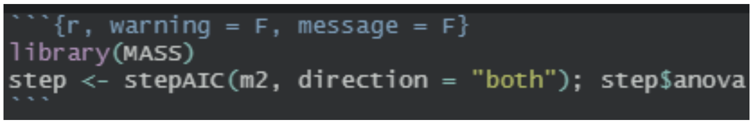


~~~~
```{r}
ptest2 <- predict(m2,newdata=data,type="response")
data.frame(data$relapse,ptest2)[1:10,]
```
~~~~



~~~~
```{r}
data$Category <- ptest2

data$Category[data$Category >= 0 & data$Category < 0.25  ] = 'No Relapse'
data$Category[data$Category >= 0.25 & data$Category < 0.5  ] = 'Low Relapse'
data$Category[data$Category >= 0.5 & data$Category < 0.75  ] = 'Moderate Relapse'
data$Category[data$Category >= 0.75 & data$Category <= 1] = 'High Relapse'

```
~~~~

Lastly, we can use the step function from the MASS library to remove unnecessary variables from a model.



~~~~
```{r, warning = F, message = F}
library(MASS)
step <- stepAIC(m2, direction = "both"); step$anova
```
~~~~

5.Model Validation

Before we conclude that the final model that we computed can predict the outcome of relapse, we need to check first if the final model is correctly specified. For model validation, doing the goodness of fit to check if the model fits the data would be the first step. However, the dataset was an ungrouped dataset, therefore using the goodness of fit and deviance would not work.

For ungrouped data just like the cervical cancer dataset, using the Hosmer Lemeshow test would give us the goodness-of-fit for the ungrouped data. The Hosmer Lemeshow test computes a chi-square statistic from observed and expected frequencies in each of the group quantiles. The first group of the quantiles consist of the observations with the lowest 10% predicted probabilities.

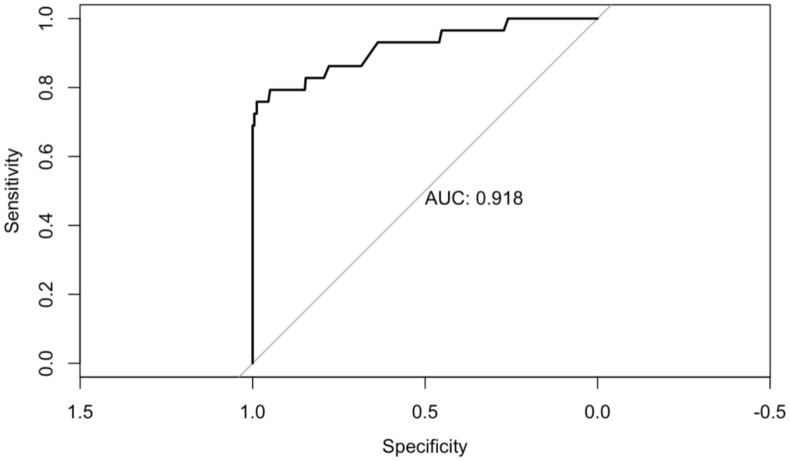
Bartlett states that "large p-value does not mean the model fits well, since lack of evidence against a null hypothesis is not equivalent to evidence in favour of the alternative hypothesis. In particular, if our sample size is small, a high p-value from the test may simply be a consequence of the test having lower power to detect misspecification, rather than being indicative of good fit" (Bartlett, 2014)

In our model, our Hosmer Lemeshow Test *p value = 0.07578*. The p value is greater than our alpha = 0.05, therefore we fail to reject our null hypothesis and conclude that our current model fits the data well.

```
        Hosmer and Lemeshow goodness of fit (GOF) test

data:  final.model$y, fitted(final.model)
X-squared = 14.237, df = 8, p-value = 0.07578
```

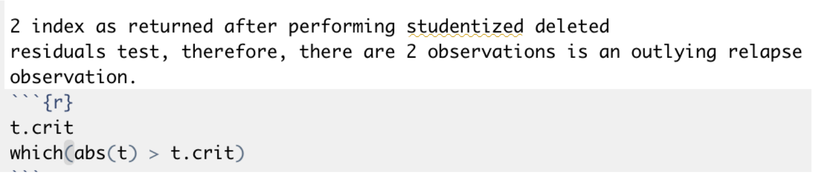
The ROC curve shows the plot of sensitivity as a function of (1 - specificity) for the possible cut-offs for a test or a combination of tests. Mandrekar tells us that the area under an ROC curve provides a measure of discrimination and allows investigators to compete the performance for two or more diagnostic tests (Mandrekar, 2015)

AUC score of 0.5 gives us no discrimination. 0.7-0.8 suggests an acceptable score. 0.9 is considered excellent and above 0.9 would be outstanding. Based on our data, our *AUC is 0.918.* This falls under the outstanding group for AUC. We can conclude that there is 91.8% chance that the model will be able to distinguish or separate the positive class from the negative class.



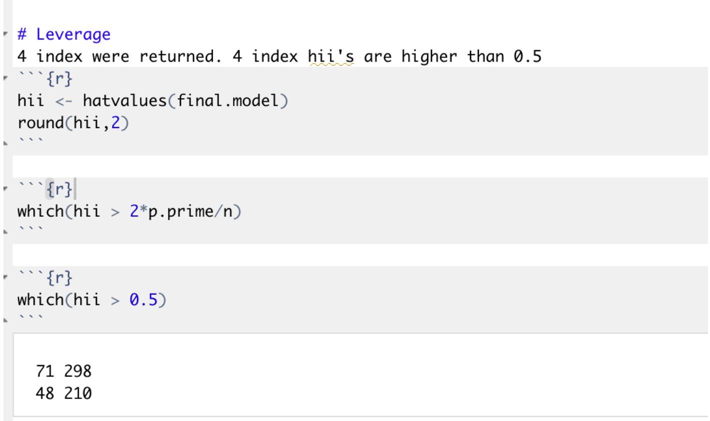
To identify outlying points in our data, using the studentized deleted residuals formula will help us identify which points are outlying if there are any. The studentized deleted residuals will delete the observations one by one and refit the model each time.

The outcome of the studentized deleted residuals that we computed concluded that there are 2 outlying relapse observations in our dataset.



~~~~
2 index as returned after performing studentized deleted
residuals test, therefore, there are 2 observations is an outlying relapse
observation.
```{r}
t.crit
which(abs(t) > t.crit)
~~~~

After getting the studentized deleted residuals, we then compute the leverage which is the measure of the distance of the x value for the ith data point and the mean of the x values for all n data points (Using Leverages to Help Identify Extreme “X” Values, PennState)



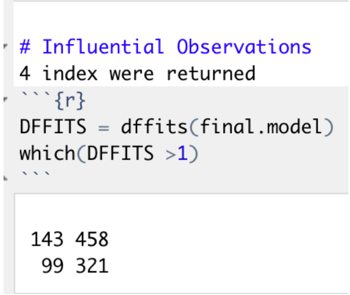
~~~~
# Leverage
4 index were returned. 4 index hii's are higher than 0.5
```{r}
hii <- hatvalues(final.model)
round(hii,2)
```

```{r}
which(hii > 2*p.prime/n)
```

```{r}
which(hii > 0.5)
```

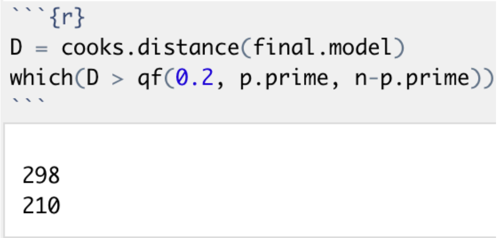
  71 298
  48 210
~~~~

We then computed the difference in fitted values or DFFits. DFFits considers the influence of the ith observation on the fitted value Yi. 4 indexes were returned after computing the DFFits.



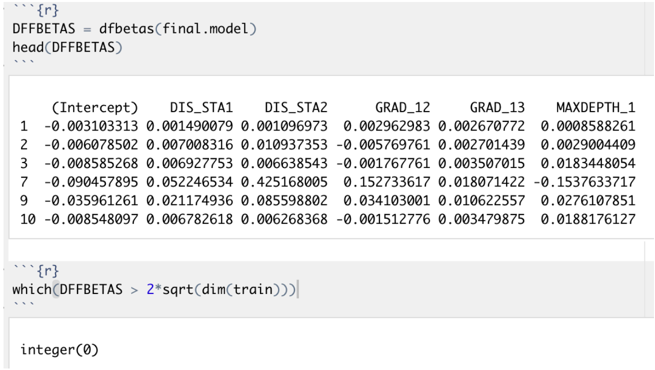
~~~
# Influential Observations
4 index were returned
```{r}
DFFITS = dffits(final.model)
which(DFFITS >1)
```

 143 458
  99 321
~~~

After computing the difference in fitted values, we then computed Cook's distance. Cook’s distance considered the influence of the ith observation on all n fitted values. 2 indexes were returned after performing the Cook’s distance.



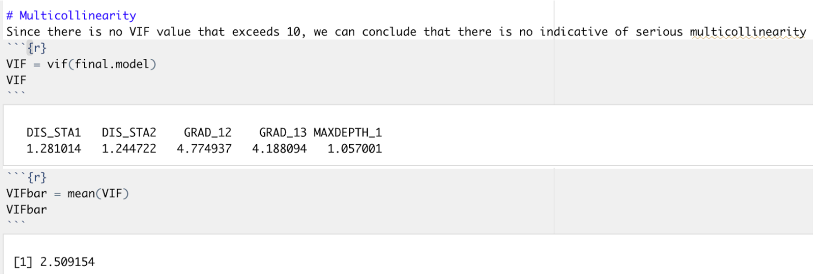
~~~
```{r}
D = cooks.distance(final.model)
which(D > qf(0.2, p.prime, n-p.prime))
```

 298
 210
~~~

After computing the Cook’s distance, we then computed the difference in betas or DFBetas. DFBetas measure the influence of an observation on each regression coefficient. There are zero observations that are influential in our regression coefficients.



~~~
```{r}
DFFBETAS = dfbetas(final.model)
head(DFFBETAS)
```

    (Intercept)    DIS_STA1    DIS_STA2      GRAD_12     GRAD_13    MAXDEPTH_1
1  -0.003103313 0.001490079 0.001096973  0.002962983 0.002670772  0.0008588261
2  -0.006078502 0.007008316 0.010937353 -0.005769761 0.002701439  0.0029004409
3  -0.008585268 0.006927753 0.006638543 -0.001767761 0.003507015  0.0183448054
7  -0.090457895 0.052246534 0.425168005  0.152733617 0.018071422 -0.1537633717
9  -0.035961261 0.021174936 0.085598802  0.034103001 0.010622557  0.0276107851
10 -0.008548097 0.006782618 0.006268368 -0.001512776 0.003479875  0.0188176127

```{r}
which(DFFBETAS > 2*sqrt(dim(train)))
```

 integer(0)
~~~

After checking for any influential observations using DFFits, Cook's distance and DFBetas, we then compute if there are any multicollinearity using the variance inflation factor or the VIF. This is used to detect any presence of multicollinearity. Since there is no VIF value that exceeds 10, we can conclude that there is no multicollinearity present in our dataset.



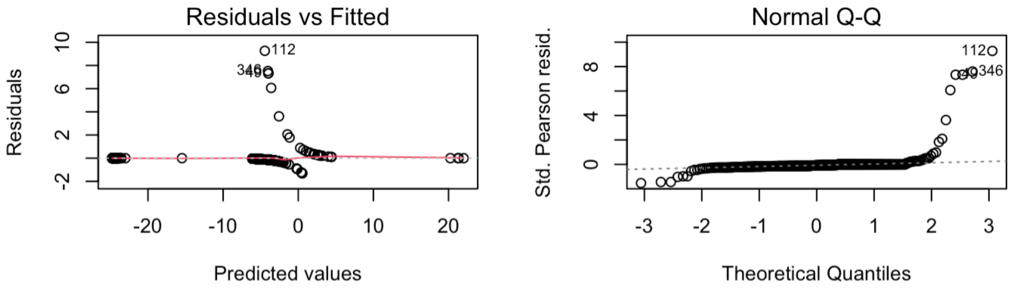
~~~~
# Multicollinearity
Since there is no VIF value that exceeds 10, we can conclude that there is no indicative of serious multicollinearity
```{r}
VIF = vif(final.model)
VIF
```

   DIS_STA1   DIS_STA2    GRAD_12    GRAD_13 MAXDEPTH_1
   1.281014   1.244722   4.774937   4.188094   1.057001

```{r}
VIFbar = mean(VIF)
VIFbar
```

 [1] 2.509154
~~~~

6.Limitation



For the final model we selected, we can see from the residual plot that the residuals do not look like a random scatter, they seem to form a reciprocal graph.

The Normal Q-Q plot is not close to the line near the top right most of the graph. So, after conducting the Shapiro-Wilk test, where the null hypothesis is that the residuals are normally distributed in some population, the p-value < 0.05. Thus, we reject the null-hypothesis and conclude that the residuals are not normally distributed in some population.

After rejecting null hypothesis and concluding that the residuals are not normally distributed from the Shapiro-Wilk test, we conducted a box-cox transformation, but we ran an error that says that the response variable must be positive.

Our team tried to shift the variables as a form of transformation, but we ran into another problem, the binomial no longer works. We exhausted all the remedial measures that we know to solve the errors in our residual plot. We will leave this for future work as we believe that the concepts to solve this problem are not within the range of this course.

7.Bibliography

Bartlett, J. (2014, February 16). The Hosmer-Lemeshow goodness of fit test for logistic regression. The Stats Geek. https://thestatsgeek.com/2014/02/16/the-hosmer-lemeshow-goodness-of-fit-test-for-logistic-regression/

Mandrekar. (2015, November 20). Receiver Operating Characteristic Curve in Diagnostic Test Assessment. ScienceDirect. https://www.sciencedirect.com/science/article/pii/S1556086415306043

Waggoner, S. E. (2003, June 28). Cervical cancer. ScienceDirect.com | Science, health and medical journals, full text articles and books. https://www.sciencedirect.com/science/article/pii/S0140673603137786?casa_token=cYS0kaXrcUoAAAAA:u1Jg8_uUrLMDCmJMgEVBXrYn90fZ0_etvhBcq2Gid5y0uB4gwlByl1xZTD5QHMUn84KfhhE6tG4

Basic information about cervical cancer. (2021, February 16). Centers for Disease Control and Prevention. https://www.cdc.gov/cancer/cervical/basic_info/index.htm#:~:text=When%20cancer%20starts%20in%20the,in%20women%20over%20age%2030

Cervical cancer. (2019, January 18). ScienceDirect.com | Science, health and medical journals, full text articles and books. https://www.sciencedirect.com/science/article/pii/S014067361832470X?casa_token=oLrEos72hvkAAAAA:_oK7n5Jg94nHSryDGWi8RlE8NiWlUXzLZdrd9Kaigw6-DY8HSGovctvk4w_nBrX4dz6QURIzni8

11.2 - Using Leverages to Help Identify Extreme x Values | STAT 501. (n.d.). PennState: Statistics Online Courses. Retrieved April 12, 2021, from https://online.stat.psu.edu/stat501/lesson/11/11.2