Covariate Analysis of Cervical Cancer

Report Done By: Benjamin Mak, Junying Zhu, Junyi Lin

|  |  |  |
| --- | --- | --- |
| Name | Student Number | Responsibilities |
| Ben Mak | 1007225085 | Modelling, Graphing |
| Junying Zhu | 1005127168 | Data Analysis, Model Validation |
| Junyi Lin | 1005800625 | Introduction, Data Analysis and Building Model |

Covariate Analysis of Cervical Cancer

2023-03-20

## Abstract

Cervical cancer is a devastating disease that affects millions of people around the world, and finding ways to prevent its relapse is crucial in improving patients’ outcomes and quality of life. Our goal is to Indicate the factors that predict the event of no-relapse occurring and also classify Patients according to individual risk of relapse.

We got 14 variables in this study,and the data contains 905 Patients.By analyzing these variables, we can determine which ones are most predictive of the event of no relapse. Furthermore, we needed to classify patients based on their likelihood of relapse from the disease.

## Variables

MRNO: Patient number.

SURGDAT: Surgery date (diagnosis date).

ADJ\_RAD: Categorical variable: 0 – if patient did NOT receive radiation therapy; 1 – if patient received radiation therapy. (radiation therapy given only when parameters are deemed severe enough by doctor)

AGE\_1 Age of patient at time of diagnosis.

CLS\_1 Categorical variable: (Prognostics) (Capillary Lymphatic Spaces) 0 – negative; 1, 2 – positive.

DIS\_STA Categorical variable: 0 – no evidence of disease; 1 – alive with disease; 2 – dead of disease; 3 – dead of complications (disease present); 4 – dead of complications (disease absent); 5 – dead of unrelated causes.

GRAD\_1 Categorical variable: cell differentiation 1 – better; 2 – moderate; 3 – worst; 0 – indicating a missing value.

HISTOLOG Categorical variable: ranging from 0 to 6.

MARGINS Categorical variable: disease left after surgery 0 – clear; 1 – para-vaginal area; 2 – vaginal area; 3 – both.

MAXDEPT Continuous variable: depth of tumor (mm); a record of 0 indicates that device could not measure depth due to small measurement.

PELLYMPH 0 – negative; 1 – positive.

RECURRN1 Date of reoccurrence of disease (if no reoccurrence, there is no entry recorded).

SIZE\_1 Size of tumor (mm) upon diagnosis.

##Libraries Used - dplyr - readxl - corrplot - ResourceSelection - pRoc

## Research and Hypothesis

The first research question is to ask which of the factors best predict the chance of no relapse of cervical cancer occuring. In order to find the factors, it is first important to categorize the different levels of relapsing of the disease. Consider the event of no relapse occuring when the date of the relapse does not exist (Cervical Cancer - Case Study FAQ). We used this fact to construct our new dependent variable called “relapse”, where we set the value as 0 if the patient did experience relapse, and 1 if the patient did not experience relapse.

The idea of the research is to find any particular factors that could be closely related to the relapse of cervical cancer in patients. Given the sample provided, we examined the common patterns found in each patient’s information in each predictor. We wanted to conduct 2 selected models, one for each research question. To predict the event of no relapse, we chose to conduct research using grouped data, particularly only with patients whose relapse result was 0. To generalize the classification of patients based on relapse, we decided to determine the predictors based on relation with the relapse variable overall. The difference to consider for each estimation is that one model estimation will be based on grouped data while the other is based on ungrouped data respectively. The difference in both case studies is the model validation.

## Collinearity Between Dependent Category and Checking for Multicollinearity

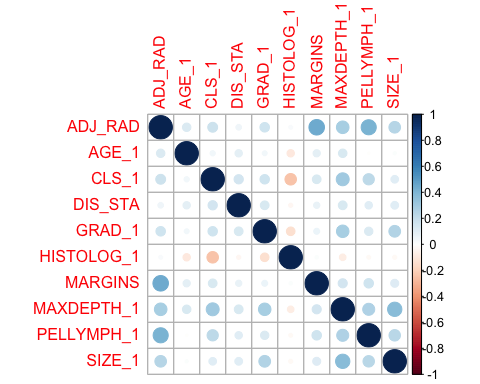
To begin with, we needed to consider any multi collinearity between any of the significant variables. To examine this, we constructed a correlation matrix that examined all significant variables that were provided in the data.

##   
## Attaching package: 'dplyr'

## The following objects are masked from 'package:stats':  
##   
## filter, lag

## The following objects are masked from 'package:base':  
##   
## intersect, setdiff, setequal, union

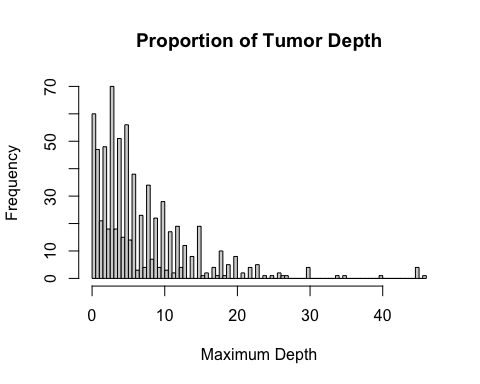
## corrplot 0.92 loaded

 In this plot, we recognized that there were not any significant multicorrelations between the variables, implying that .If we found that any of these variables had some significance with our relapse variable, we would also need to consider the interactions between each of these terms in our final selected model.

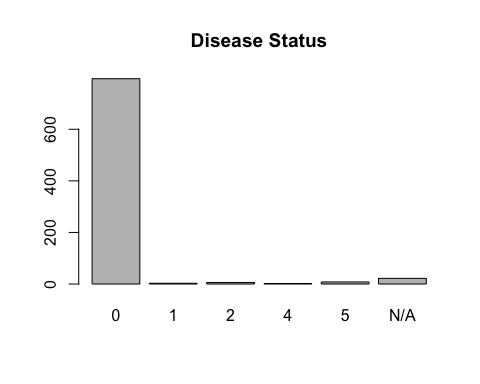
## Analysis of Grouped Variables

To begin with, we need to consider the singular independence by constructing simple histograms and bar graphs that display the proportions of each variable.

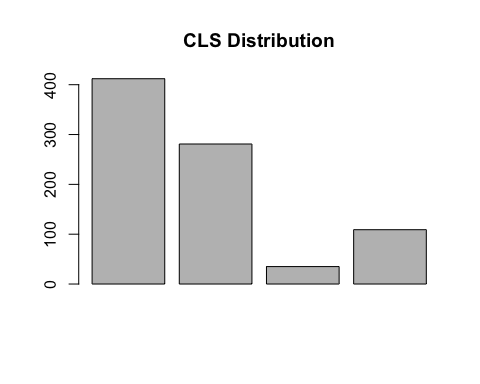
hist(test$MAXDEPTH\_1,main = "Proportion of Tumor Depth", breaks = 100, xlab = "Maximum Depth")



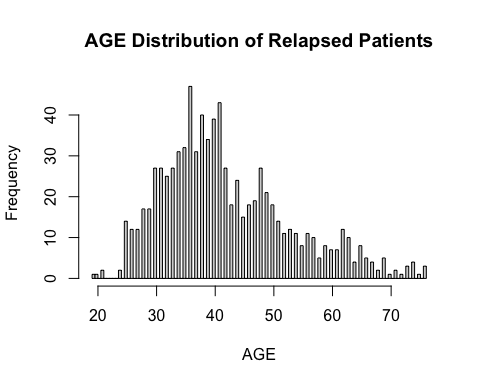
dis\_sta\_count <- test %>% count(DIS\_STA)  
barplot(dis\_sta\_count$n, main = "Disease Status", name = c(0,1,2,4,5, "N/A"))



plot\_count <- test %>% count(CLS\_1)  
barplot(plot\_count$n, main = "CLS Distribution")

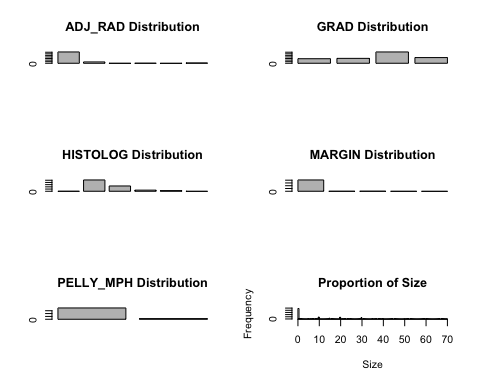


hist(test$AGE\_1, main = "AGE Distribution of Relapsed Patients", breaks= 100, xlab = "AGE")

 We found that these four variables must have a strong correlation with relapse occurring. There was a common pattern in each of the patients that did not suffer relapse from the disease: Age of patents were mostly around 30-40 years old, the depth of tumors in each patient were 0-10 mm, size of tumors were 0-1 mm and patients reported to be alive and free of disease. However, one other thing that we found was that the proportion of patients who had no relapse occur and also were alive and had no sign of disease were significantly high. In fact, we found that approximately 2/3 of the overall sample obtained contained patients that met both of these requirements. This would be a problem for classifying patients based on relapse, as this factor would indicate a bias in patients who classified with no relapse.

We also decided to check if the non-correlated predictors had a different type of distribution. Similarly, we checked by using a simple bar graph distributing each of the remaining variables.

par(mfrow = c(3,2))  
plot\_count <- test %>% count(ADJ\_RAD)  
barplot(plot\_count$n, main = "ADJ\_RAD Distribution")  
  
plot\_count <- test %>% count(GRAD\_1)  
barplot(plot\_count$n, main = "GRAD Distribution")  
  
plot\_count <- test %>% count(HISTOLOG\_1)  
barplot(plot\_count$n, main = "HISTOLOG Distribution")  
  
plot\_count <- test %>% count(MARGINS)  
barplot(plot\_count$n, main = "MARGIN Distribution")  
  
plot\_count <- test %>% count(PELLYMPH\_1)  
barplot(plot\_count$n, main = "PELLY\_MPH Distribution")  
  
hist(test$SIZE\_1, main = "Proportion of Size", breaks = 100, xlab = "Size")

 Unlike the first four indicators analyzed, we found that the other variables did not follow a similar pattern, there was not any significant skew right. Instead, we found that the other factors seemed to be “normally distributed”. For some other factors that were skewed right, it did not seem reasonable based on the meanings of the variables. For instance, most patients had HISTOLOG = 1, which is Squamous cell carcinoma, which is a cancer cell that is often to relapse in patients. All in all, this indicated that there was not any significant correlations between relapse and these other predictors.

## Constructing the Fitted Model

Based on our findings from the previous analysis, we started to construct a selected model that would best predict the outcomes of the relapse variable. To verify our findings for independent correlations, we constructed a combined model to examine the dependence between relapse and all of the possible predictors. For each variable, we tested the null hypothesis as the predictor coefficient is equal to 0, and our alternate hypothesis as the coefficient not equal to 0. In the case we fail to reject the null hypothesis, the coefficient could be dropped from the model. We decided to be willing to make a 5% error, which meant that if any p-value fell below the 5% threshold, that we reject the null hypothesis, and reject the null hypothesis if it is greater.

combined\_fit <- model <- glm(relapse ~ DIS\_STA + ADJ\_RAD + AGE\_1 + CLS\_1 + GRAD\_1 + HISTOLOG\_1 + MARGINS+ MAXDEPTH\_1+ PELLYMPH\_1+ SIZE\_1, family = binomial, data = data)  
summary(combined\_fit)

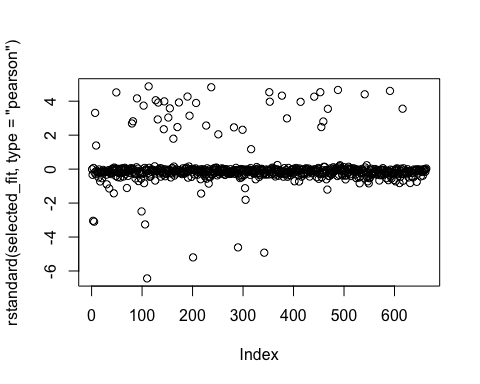
##   
## Call:  
## glm(formula = relapse ~ DIS\_STA + ADJ\_RAD + AGE\_1 + CLS\_1 + GRAD\_1 +   
## HISTOLOG\_1 + MARGINS + MAXDEPTH\_1 + PELLYMPH\_1 + SIZE\_1,   
## family = binomial, data = data)  
##   
## Deviance Residuals:   
## Min 1Q Median 3Q Max   
## -2.6981 -0.2995 -0.2053 -0.1524 3.0957   
##   
## Coefficients:  
## Estimate Std. Error z value Pr(>|z|)   
## (Intercept) -3.43022 1.04742 -3.275 0.00106 \*\*   
## DIS\_STA 1.02615 0.17181 5.973 2.33e-09 \*\*\*  
## ADJ\_RAD 0.17848 0.41176 0.433 0.66469   
## AGE\_1 -0.04286 0.01922 -2.229 0.02579 \*   
## CLS\_1 0.79725 0.31204 2.555 0.01062 \*   
## GRAD\_1 0.29848 0.25772 1.158 0.24680   
## HISTOLOG\_1 0.08859 0.15940 0.556 0.57839   
## MARGINS -0.56748 0.68657 -0.827 0.40850   
## MAXDEPTH\_1 0.05318 0.02071 2.568 0.01024 \*   
## PELLYMPH\_1 0.40249 0.56696 0.710 0.47776   
## SIZE\_1 0.01378 0.01396 0.987 0.32357   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## (Dispersion parameter for binomial family taken to be 1)  
##   
## Null deviance: 305.30 on 644 degrees of freedom  
## Residual deviance: 220.09 on 634 degrees of freedom  
## (260 observations deleted due to missingness)  
## AIC: 242.09  
##   
## Number of Fisher Scoring iterations: 6

To test any interaction between any of the correlated variables, we decided to conduct a backwards elimination using StepAIC. We ran into complications with forward and both direction eliminations because they had relatively higher AICs, implying that their p-values would be less accurate. For the backwards elimination, although we obtained many interaction terms, we decided to only keep the interaction terms that had lower p-values, which are the ones that are more significant.

saturated.fit <- glm(relapse ~ DIS\_STA\*AGE\_1\*CLS\_1\*MAXDEPTH\_1,data = data)  
fit <- glm(relapse~1, data = data)  
  
backwards\_fit <- step(saturated.fit, direction= "backward", trace = 0)  
summary(backwards\_fit)

##   
## Call:  
## glm(formula = relapse ~ DIS\_STA \* AGE\_1 \* CLS\_1 \* MAXDEPTH\_1,   
## data = data)  
##   
## Deviance Residuals:   
## Min 1Q Median 3Q Max   
## -1.02039 -0.04524 -0.01628 -0.00593 0.98924   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) -0.0372782 0.0555549 -0.671 0.502450   
## DIS\_STA 0.2053145 0.1071912 1.915 0.055881 .   
## AGE\_1 0.0008897 0.0013075 0.680 0.496472   
## CLS\_1 -0.0900949 0.0844575 -1.067 0.286483   
## MAXDEPTH\_1 0.0124471 0.0065465 1.901 0.057701 .   
## DIS\_STA:AGE\_1 -0.0036342 0.0020309 -1.789 0.074007 .   
## DIS\_STA:CLS\_1 0.4484517 0.1299847 3.450 0.000597 \*\*\*  
## AGE\_1:CLS\_1 0.0022567 0.0020067 1.125 0.261192   
## DIS\_STA:MAXDEPTH\_1 0.0945111 0.0108428 8.716 < 2e-16 \*\*\*  
## AGE\_1:MAXDEPTH\_1 -0.0002201 0.0001522 -1.446 0.148710   
## CLS\_1:MAXDEPTH\_1 0.0086626 0.0076222 1.136 0.256170   
## DIS\_STA:AGE\_1:CLS\_1 -0.0058794 0.0024593 -2.391 0.017100 \*   
## DIS\_STA:AGE\_1:MAXDEPTH\_1 -0.0018476 0.0002170 -8.512 < 2e-16 \*\*\*  
## DIS\_STA:CLS\_1:MAXDEPTH\_1 -0.1059616 0.0128847 -8.224 1.08e-15 \*\*\*  
## AGE\_1:CLS\_1:MAXDEPTH\_1 -0.0001586 0.0001792 -0.885 0.376527   
## DIS\_STA:AGE\_1:CLS\_1:MAXDEPTH\_1 0.0020207 0.0002473 8.172 1.60e-15 \*\*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## (Dispersion parameter for gaussian family taken to be 0.0350376)  
##   
## Null deviance: 39.339 on 662 degrees of freedom  
## Residual deviance: 22.669 on 647 degrees of freedom  
## (242 observations deleted due to missingness)  
## AIC: -322.62  
##   
## Number of Fisher Scoring iterations: 2

selected\_fit <-glm(relapse ~ DIS\_STA+AGE\_1+CLS\_1+MAXDEPTH\_1+DIS\_STA:MAXDEPTH\_1 + DIS\_STA:AGE\_1:CLS\_1 + DIS\_STA:CLS\_1:MAXDEPTH\_1 + DIS\_STA:AGE\_1:CLS\_1:MAXDEPTH\_1 , data = data)  
  
plot(rstandard(selected\_fit, type = "pearson"))

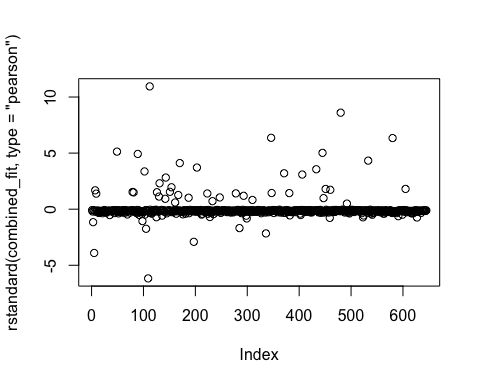
 In the residual plot for our new selected model, we found that the residuals are relative at 0, meaning that the selected model is potentially a good fit. furthermore, there were not many outliers, meaning that over dispersion was not significant enough to consider in our study.

## Model Validation

Now that we have selected a model, we needed to validate it to ensure that it is a sufficient model that can represent the event of relapse of cervical cancer. To start, we noted that for our first case, we needed to use the holmer-lemeshow test to test if our selected model is adequate enough compared to the saturated model, since the data in the case of no relapse is grouped. We also constructed a ROC curve and examined the area below the curve.

plot(rstandard(combined\_fit, type = "pearson"))  
library(ResourceSelection)

## ResourceSelection 0.3-5 2019-07-22



library(pROC)

## Type 'citation("pROC")' for a citation.

##   
## Attaching package: 'pROC'

## The following objects are masked from 'package:stats':  
##   
## cov, smooth, var

hoslem.test(selected\_fit$y, fitted(selected\_fit), g = 4)

##   
## Hosmer and Lemeshow goodness of fit (GOF) test  
##   
## data: selected\_fit$y, fitted(selected\_fit)  
## X-squared = 2.1585, df = 2, p-value = 0.3398

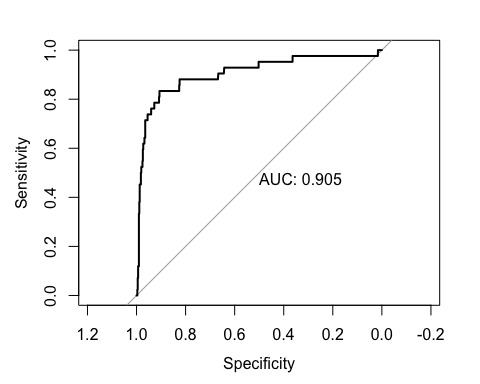
anova(selected\_fit, saturated.fit, test = "Chisq")

## Analysis of Deviance Table  
##   
## Model 1: relapse ~ DIS\_STA + AGE\_1 + CLS\_1 + MAXDEPTH\_1 + DIS\_STA:MAXDEPTH\_1 +   
## DIS\_STA:AGE\_1:CLS\_1 + DIS\_STA:CLS\_1:MAXDEPTH\_1 + DIS\_STA:AGE\_1:CLS\_1:MAXDEPTH\_1  
## Model 2: relapse ~ DIS\_STA \* AGE\_1 \* CLS\_1 \* MAXDEPTH\_1  
## Resid. Df Resid. Dev Df Deviance Pr(>Chi)   
## 1 654 29.179   
## 2 647 22.669 7 6.51 < 2.2e-16 \*\*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1

roc(selected\_fit$y ~ fitted(selected\_fit), plot = TRUE, print.auc = TRUE)

## Setting levels: control = 0, case = 1

## Setting direction: controls < cases



##   
## Call:  
## roc.formula(formula = selected\_fit$y ~ fitted(selected\_fit), plot = TRUE, print.auc = TRUE)  
##   
## Data: fitted(selected\_fit) in 621 controls (selected\_fit$y 0) < 42 cases (selected\_fit$y 1).  
## Area under the curve: 0.905

In the end, we found that our selected model is sufficient compared to our selected model. For the homer lemeshow test, we found that the p-value was significantly small, so fail to reject the null hypothesis and conclude based on this test,that the simpler model is sufficient. We ended up getting a similar result from our ROC curve construction. The area obtained from the ROC was relatively high. So we make the same conclusion that the simpler model is sufficient.

## Part 2: Classifying Patients based in Relapse

Instead of finding the correlation between predictors and the relapse at 0, we now are trying to find predictors that predict relapse in general. We decided to keep the predictors that would determine the event of patients that did not relapse. This was because This was based on our finding that most patients in the sample were for certain to not face any relapse of the disease. We changed the dependent variable to the whole relapse variable instead of considering the patients who had relapse = 0. Because the dependent variable is now ungrouped, we verify the model by constructing a Cochran–Mantel–Haenszel test.Similarly, we considered the p-value compared to a 5% level of significance

drop1(selected\_fit, test = "Chisq")

## Single term deletions  
##   
## Model:  
## relapse ~ DIS\_STA + AGE\_1 + CLS\_1 + MAXDEPTH\_1 + DIS\_STA:MAXDEPTH\_1 +   
## DIS\_STA:AGE\_1:CLS\_1 + DIS\_STA:CLS\_1:MAXDEPTH\_1 + DIS\_STA:AGE\_1:CLS\_1:MAXDEPTH\_1  
## Df Deviance AIC scaled dev. Pr(>Chi)   
## <none> 29.179 -169.25   
## DIS\_STA:AGE\_1:CLS\_1:MAXDEPTH\_1 1 30.607 -139.58 31.666 1.831e-08 \*\*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1

The p-value obtained from this test was also significantly small. So we concluded that the estimated model for the relapse variable is adequate.

## Conclusions

In the end, we found that the following predictors would determine the likelihood of patients suffering from relapse of the disease: DIS\_STA, MAX\_DEPTH, AGE and SIZE. We were able to find these by using the Wald Test on a combined estimated model and backwards elimination step AIC to determine any interaction terms to include in the model.

However, we do not believe that the sample obtained does not accurately reflect the possibility of relapsation and ran into a few limitations in our conducted research.

After conducting the Data Analysis, it was found that there was some bias in the data, as the sample obtained contained a proportion of more patients without the disease. This implies that in the data, that the predictors obtained for patients with the disease cannot be accurately generalized for the group. If the sample contained more patients that did and currently are suffering relapse, that our findings based on this sample would have some significant changes.

Secondly, we found that not all patients information did not all have a follow-up. Furthermore, we were not able to classify what date is reasonable for a follow-up and which date is not. The follow up dates found in all of the patients had a wide range in time and so we were not able to find a common time where patients potentially were involved in a follow-up.

## Resources

Cervical Cancer - Case Study FAQ, web.archive.org/web/20030502072908/<http:/www.math.yorku.ca/Who/Faculty/Ng/ssc2002/CervicalFAQ.html>.