ANALYSIS OF A CLINICAL TRIAL IN THE TREATMENT OF CARCINOMA OF THE OROPHARYNX

A DATA ANALYSIS PROJECT SUBMITTED FOR STAT356, METU

ELİF DOĞAN DAR
22 MAY 2017

1.INTRODUCTION

In this study, we are given data of 195 patients with squamous carcinoma of 3 sites in the mouth and throat reported by six institutions. Patients in the study are randomly assigned to two different treatment groups. This study included measurements of many covariates which would be expected to relate to survival experience. Six such variables are given in the data (sex, T staging, N staging, age, general condition, and grade).

We would like to know whether one of these two treatments is more preferable than the other. Also, we would like to determine the extent to which the covariates relate to survival time.

In this project, we will be performing a linear regression to predict survival time of the patients which is given as number of days. Most of the covariates are factors, some of the are integer valued. Among them, we will not be using CASE variable, which doesn't provide any information. We will make a variable transformation for ENTRY_DT variable and try to compare the model with and without that variable. To compare these models, we will perform 5-fold cross-validation technique and compare according to the mean of the sum of squared errors(sse) across 5-folds.

2.DATA

Data of this project, named "PHARYNX", consists of 13 variables and 195 observations. Source of the data is "The Statistical Analysis of Failure Time Data, by JD Kalbfleisch & RL Prentice, (1980), Published by John Wiley & Sons". The dependent variable is TIME which is the survival time in days from the day of diagnosis. We will not be using variable CASE, which is the order of the data, it doesn't give any information about the data. There are 6 factor variables in the data, namely; INST(participating institution), SEX(1 if the patient is male, 2 if the patient is female), TX(1 if it is standard treatment, 2 if it is new treatment), SITE(site of the primary tumor, 1=faucial arch, 2=tonsillar fossa, 3=posterior pillar, 4=pharyngeal tongue, 5=posterior wall), STATUS(0=censored, 1=dead) and N_STAGE(N=0 refers to there being no clinical evidence of a lymph node metastasis and N=1, N=2, N=3 indicate, in increasing magnitude, the extent of existing lymph node involvement). There is a continuous covariate AGE which is in years at the time of diagnosis. There are some ordinal variables, GRADE (the degree to which the tumor cell resembles the host cell; 1=well differentiated, 2=moderately differentiated, 3=poorly differentiated, 9=missing), COND (Condition: 1=no disability, 2=restricted work, 3=requires assistance with self-care, 4=bed confined, 9=missing) and T_STAGE (1=primary tumor measuring 2 cm or less in largest diameter, 2=primary tumor measuring 2 cm to 4 cm in largest diameter with minimal infiltration in depth, 3=primary tumor measuring more than 4 cm, 4=massive invasive tumor). We will treat these variables as factors because although these variables have a natural order, the meaning of one unit change is not fixed. Finally, there is variable ENTRY_DT (the date of study entry: Day of year and year, dddyy). To be able to use that variable, we will create a new variable which gives day of this date from a reference point, namely the first day of 1968.

3. METHODOLOGY

I will use multiple linear regression linear model since dependent variable is continuous. While doing the variable selection, I will do stepwise, backward and forward variable selection. The stepwise selection algorithm starts with null model and compares Akaike Information Criteria(AIC) of all the models where one variable is added and the null model. It picks the model with the smallest AIC value. Then it repeats the procedure but this time all models where one variable is added, all models where one variable is deleted and current model are compared. Again, it picks the model which decreases AIC value the most. This algorithm stops when none of the alternative models decrease the AIC. In forward selection, we start with the null model but in every step either we add a variable or stop. In backward selection, we start with the full model and go backward.

I will use k-fold cross validation with k=5 to compare and validate these models. In cross validation technique, we divide data into k equal chunks. We train the model with k-1 of them and test with the remaining chunk, and calculate the sum of squared errors for test data. We repeat this

procedure k times until all of the chunks are being used as test data once. In the end, we take the mean of this k different sum of squared errors. This way we are using all observations as test once, also we are getting rid of the possible bias because of the selection of the training set and outliers.

4. RESULTS

Treatment of missing values: In cond and grade variables missing values are coded as 9. Therefore, we start with replacing them with "NA", so that R can understand them as missing values. We could set 9 as NA value while loading the data. But it could result in changing some values which are also 9 in other variables to "NA" values wrongly. Therefore, I chose to change afterward. After that, I deleted variables case and entry_dt from data because they don't give information about the dependent variable. However, later on, I realized that entry_dt is also explaining dependent variable to some extent. Therefore, first I will proceed with entry_dt deleted, later on, some transformation of entry_dt will be added and I will compare these models.

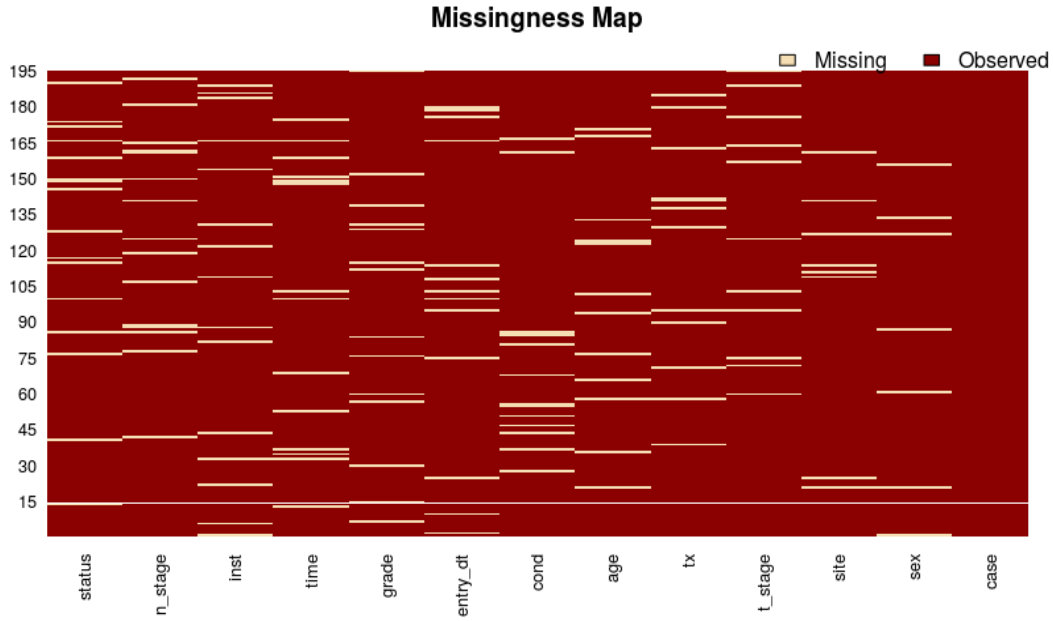


Figure 4.1

I predicted the missing values by using the package "Amelia". First I looked at the missingness map (Figure 4.1) with the command missmap(). This map of the missing values shows that there doesn't seem to be a certain pattern in missing values and we can assume that they are random. This assumption is crucial for the usage of Amelia. Then I used function amelia() which gave some unexpected results, to make it better I added information of boundaries of the variables. Also, there is a randomness factor in Amelia, so to make results better I repeated Amelia procedure 5 times, this way I got 5 different imputations of the data. These imputations agree on non-missing values but can be different on missing values. Then to be able to get the final imputations I took the mode of factor variables and median of the non-factor variables across all 5 imputations. I used median instead of mean because all continuous variables I have are integer-valued, so that kind of imputation will be against the nature of the data. Also median is more robust than mean, so it won't be much affected by one or even two outliers out of 5 imputations.

To check whether our imputations are consistent with the observed data or not we can use density plots. In Figure 4.2, you can see that observed and imputed values seem to have similar distributions. Discrepancies in the densities are mostly because the number of imputations is way smaller than the number of observed values.

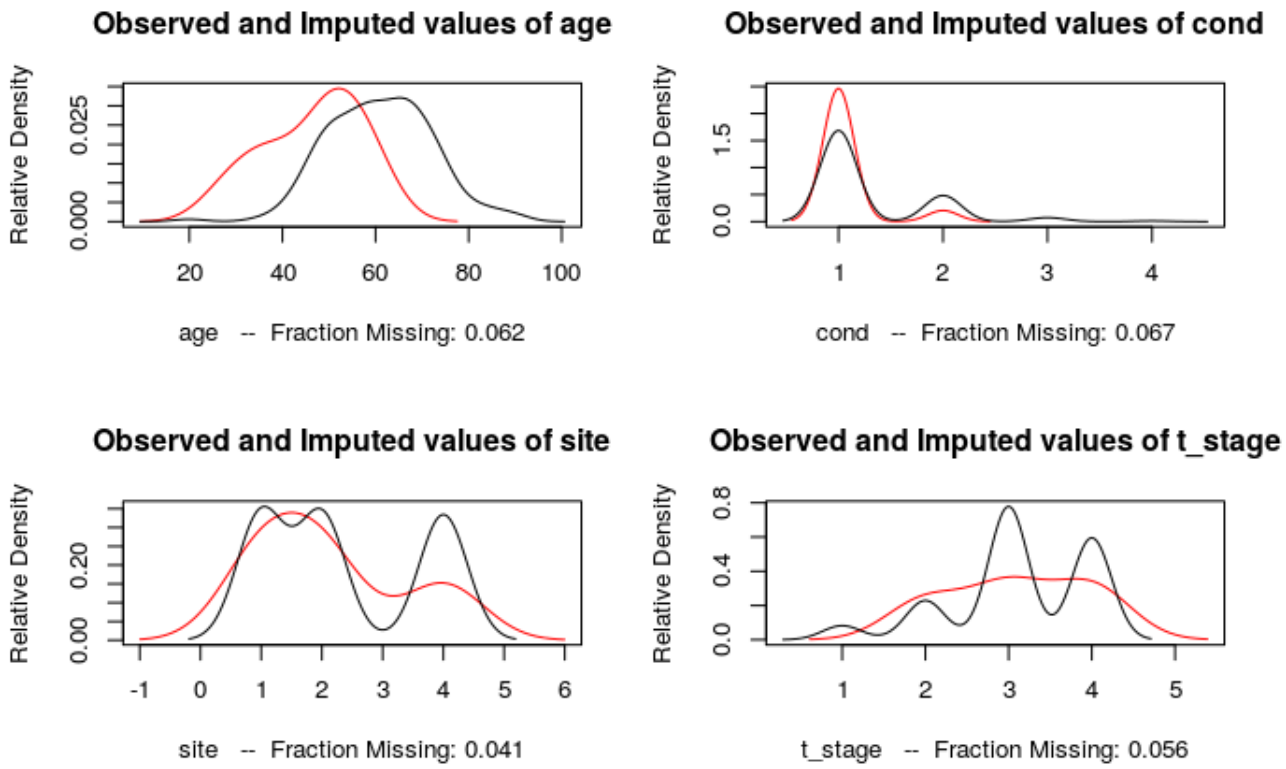


Figure 4. 2

To be able to understand whether we made good imputations or not, we should be able to compare imputed values with observed values of the same data point, which is not possible. To overcome this we can use overimpute() function in that package. This function treats all observed values as missing values one by one and makes imputations for them, then it compares these predictions with the observed values and gives the result as a plot. Points near the line in the graph shows that they are good predictions and faraway points indicate that they are bad predictions. You can see the overimpute graph (Figure 4.3) of the variable age for example if we treat observed values of age as missing and try to impute them only 5 percent of them would be predicted poorly. We have a similar result for the variable time.

Exploratory Data Analysis: First we start with the correlation matrix. Most of the variables have either no or very little linear relation. The only relation which is high but not high enough to cause multicollinearity is between time and status with a correlation coefficient of – 0.65. It makes sense, time for alive people is higher than dead. At this point, before proceeding with data visualizations I changed inst, tx, sex, site and status variables to factor. Because unit change in those variables is not fixed. I will use “ggplot2” package for visualizations.

We would like to see whether data is homogeneously distributed among variables or not. As you can see in Figure 4.4, sex seems to be equally distributed among treatment groups. Also, you can see in Figure 4.5, sex seems to be equally distributed among different institutions too. From t_stage-status graph (Figure 4.6) we can see that percentage of alive is decreasing with t_stage increasing. Because of the linear relationship between status and time, most probably it has the same relation with time. As seen in the bar plot of the variable cond shows that there are fewer observations with grade=4, actually only 1. It might cause a problem in the future. We will get back to that.

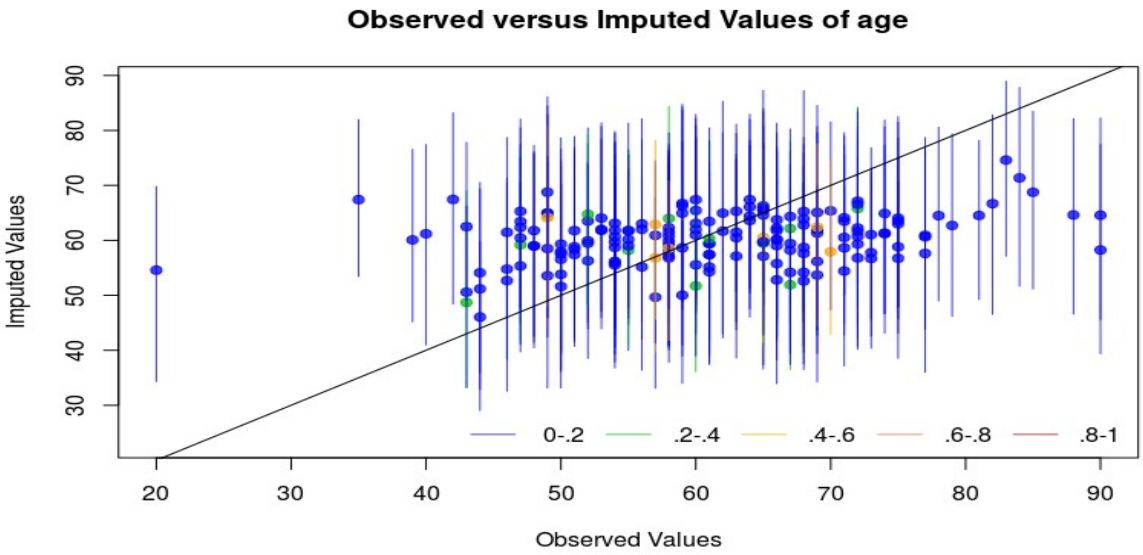


Figure 4. 3

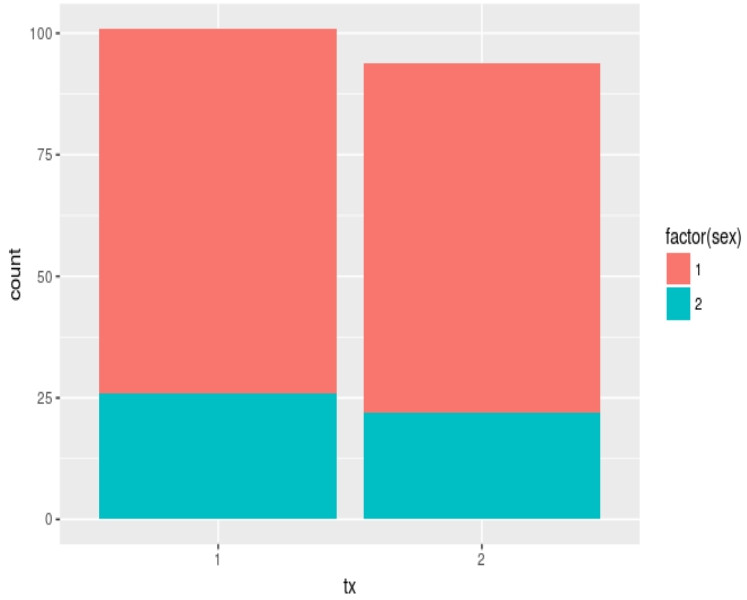


Figure 4. 4

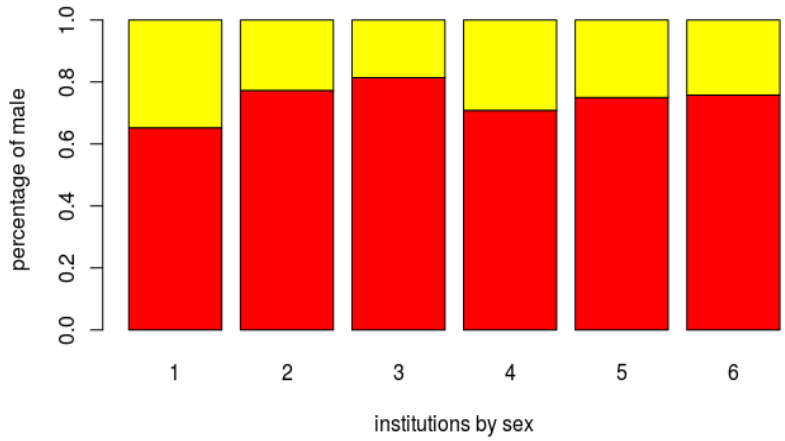


Figure 4. 5

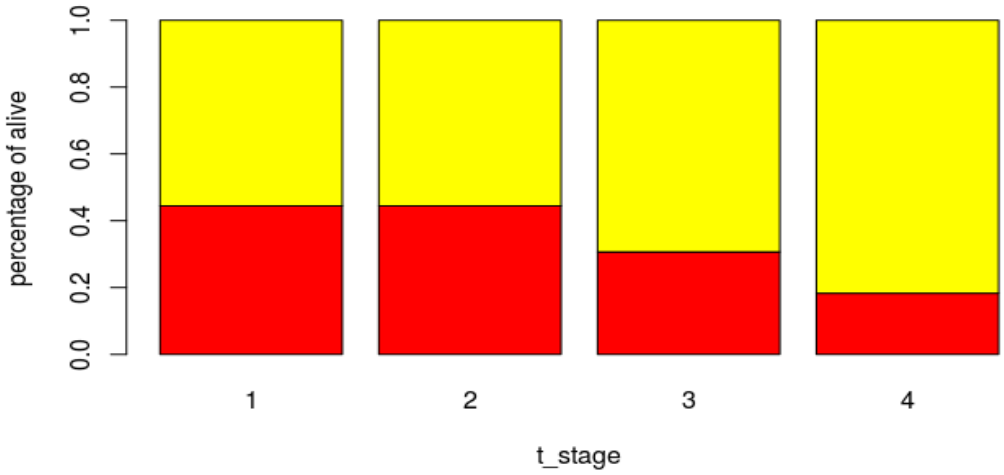


Figure 4. 6

As you can see in the boxplot of time with respect to different institutions (Figure 4.7), mean of the time variable is approximately same for all institutions except inst=3, and variance is similar except inst=3 and 4. It also didn't have any linear relationship with other variables. We can suspect that inst variable will not be in the final model. From stacked bar chart of time variable by sex which can be seen in Figure 4.8, it looks like percentage of the women in time variable increases over time. As we can see in the boxplot of time by tx(Figure 4.9), mean of second treatment is slightly smaller but the variation is very small compared to the first treatment, because of that there are many outliers in time with second treatment. In the graph of time and age by status graph(Figure 4.10), lines are almost parallel except the beginning where the line is being curved because of some outlier. Therefore, most probably there is an interaction between age and status. Same applies to the graph of scatter cond and time by status, Figure 4.11. There doesn't seem to be an interaction between condition and status.

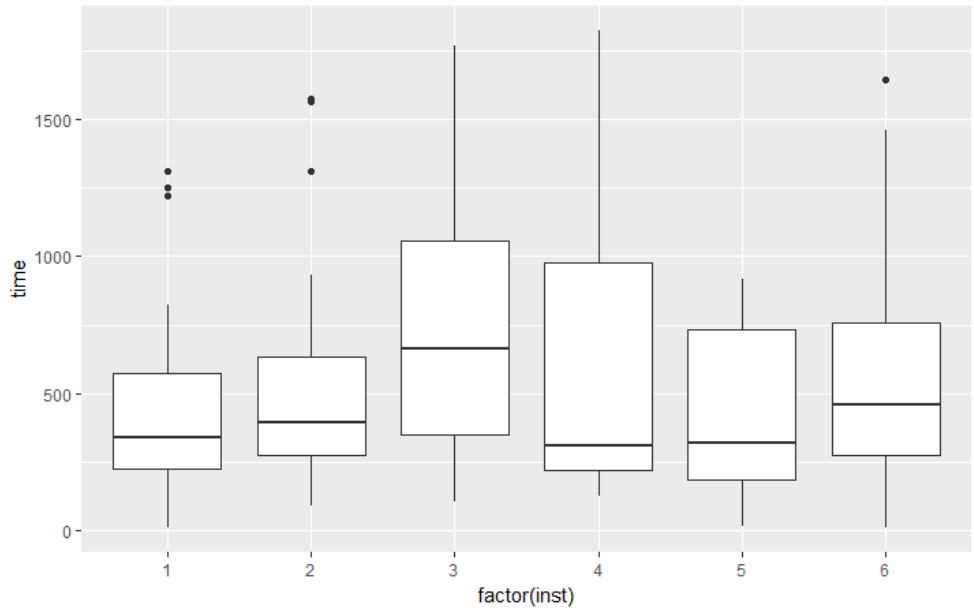


Figure 4. 7

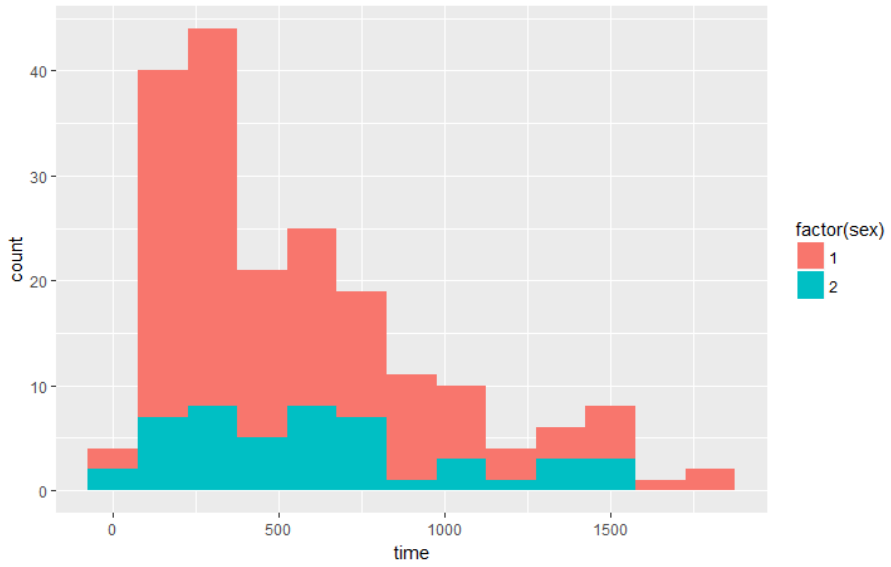


Figure 4. 8

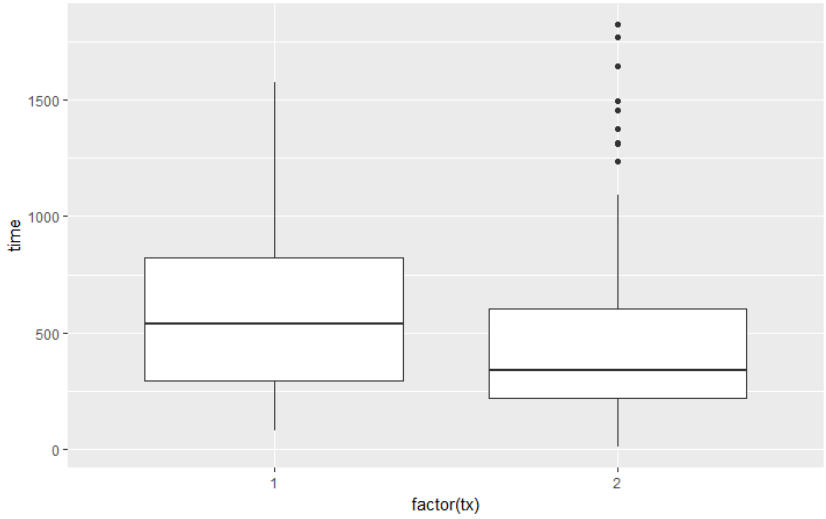


Figure 4. 9

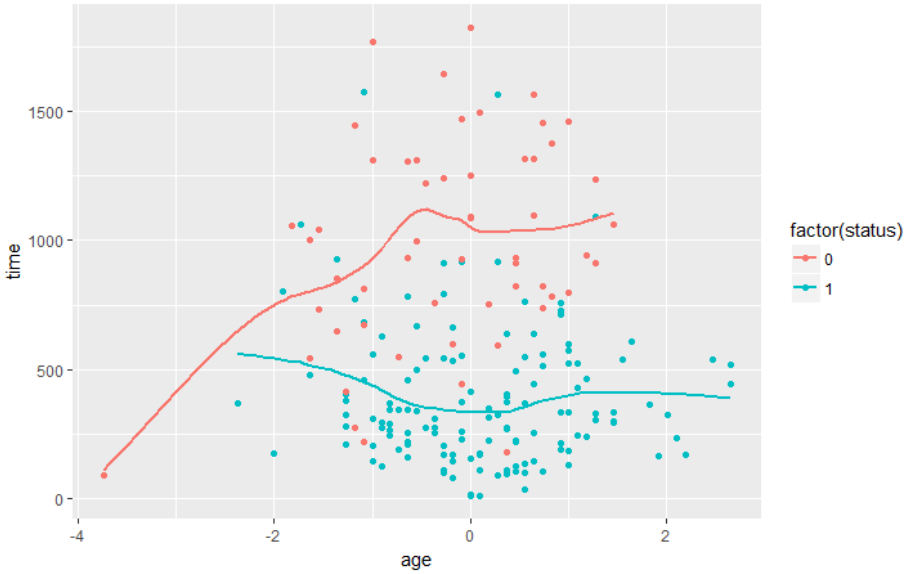


Figure 4. 10

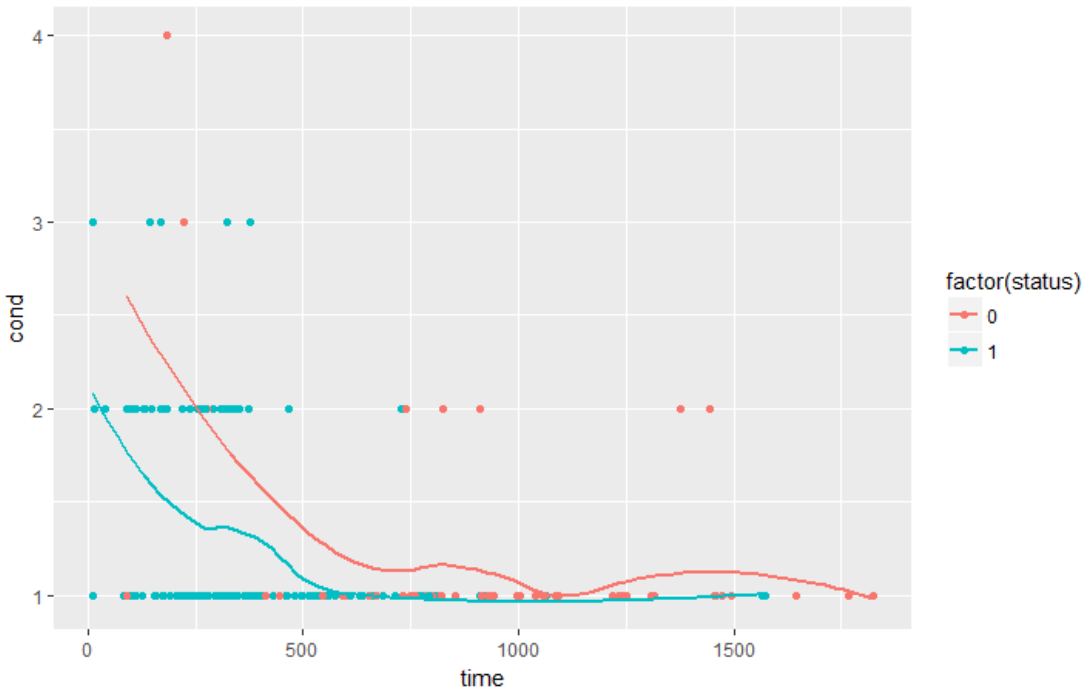
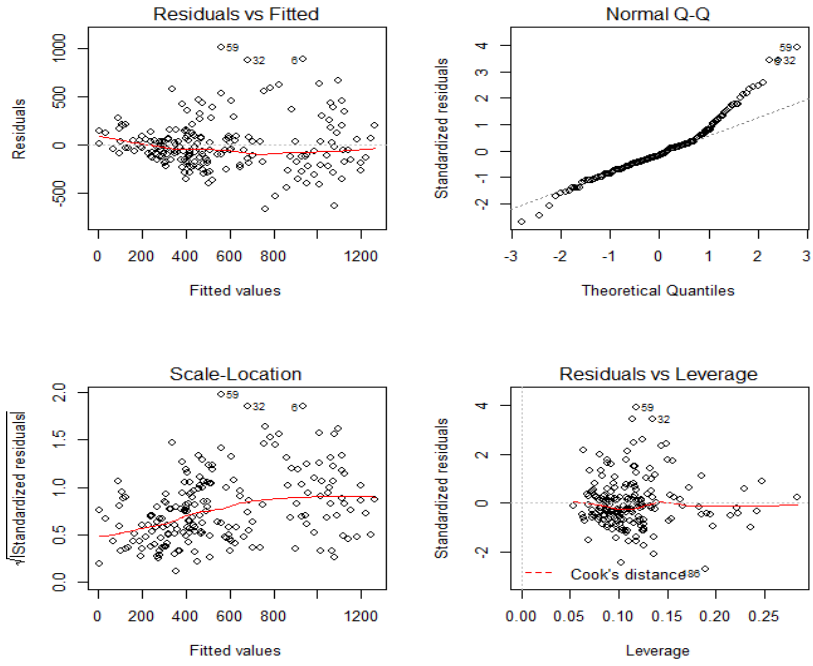
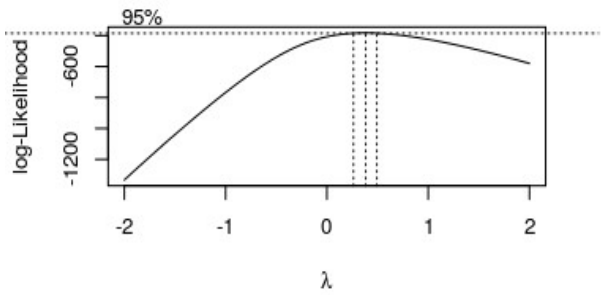


Figure 4. 11

Model Building: First we build a linear model by using all the covariates excepts case and entry_dt. This gives a significant model with an R^2 value of 0.5491. Our linear model should satisfy some certain assumptions. Dependent variable should be approximately normally distributed. Residuals should have constant variance. There shouldn't be any multicollinearity among covariates. Also, error terms should be uncorrelated. There shouldn't be any outliers which affect the model greatly. Finally, there shouldn't be unnecessary covariates in the model. Because having variables more than necessary doesn't add accuracy to the model and increases the variance. When we look at the plot of the model1 plot(Figure 4.12), normal-QQ plot suggests that there is normality problem and scale-location plot suggests that there is constant variance problem.



Normality assumption: To test this claim from the plot we apply Shapiro-Wilk test. In this test null hypothesis is that the variable is normally distributed. When we apply this test, it gives an approximately zero p-value, which means that our dependent variable is not normally distributed. To fix this problem we will make a transformation. To decide which transformation we will use Box-Cox method. This method suggests a lambda value for the transformation y^{λ}, as you can see in the Figure 4.13. Box-Cox transformation suggests λ value 1/3 for this model. When we do the transformation we get 0.085 as p-value in Shapiro-Wilk test. We solved the normality problem. We build another model, model2, with all the covariates and transformed time variable. R^2 value of this new model is 0.5719. But there are many insignificant variables in the model. To be able to pick the best set of variables we will apply backward, forward and stepwise methods and we will compare those models with 5-fold cross validation.



5-fold cross-validation: To be able to compare models I will perform k-fold cross validation as explained before with "caret" package. However, while trying to find the mean of the sum of squared errors across 5-folds I got an error: "`factor cond has new levels 4`", which means that model cannot predict observation with cond=4 because there was no such factor level in the training set for one of the folds. When I inspected the data I saw that cond variable is 4 only for one observation out of 195 observations. Therefore, I modified model2 by simply erasing that observation. This didn't change anything in the model but now I can compute mean sse of the model, which is 80.11.

Variable selection: Afterwards I performed forward, backward and stepwise variable selection methods, naming resulting models as forward, backward and stepwise respectively. Mean sse of backward is 76.85 where mean sse of stepwise is 77.48. Backward has both smallest number of variables and smallest mean sse, therefore we will use backward as our final model till now. We will refer it as model3 from now on.

Uncorrelated Errors: To check uncorrelated error assumption for model3 we will use Durbin-Watson test from "lmtest" package. The null hypothesis for this test is that there is no autocorrelation among errors. Our p-value is 0.4982 which is bigger than 0.05, therefore we do not reject the null hypothesis. There is no autocorrelation.

Heteroscedasticity: We should check whether residuals have a constant variance or not. We will use ncv test in package "car". The null hypothesis is that residuals have constant variance. Our p-value is 0.18 which is greater than 0.05. Therefore, we do not reject the null hypothesis. We do not have heteroscedasticity problem.

Multicollinearity: Multicollinearity problem arises when variables are too much dependent on each other. When one variable is a linear function of the other we can have many different solutions to the regression which are very different than each other. These different models work well on the training set and work very poorly on the test data. To check whether we have this problem in our data we look at vif values. Since none of the vif values is greater than 10, we conclude that there is no multicollinearity problem.

Influential points: To check whether we have any influential point which affects the model

too drastically we use influence.measures() function. According to the summary, we suspect that 19,59,89 and 186 might be influential. I deleted them and built a new model named model4, I did variable selection from this model and found that backward selection of this model gives the best sse and the least variable number, let's call it backward2. While comparing model 3 and backward2, sse of the model without outliers is less than the one with outliers. Also, R^2 improved a little. Also, it turns out that without outliers site variable is not necessary anymore. So, there is big influence caused by outliers and we will use backward2 instead of model3 so backward2 is our final model till now.

Interaction effect: Sometimes, the behavior of some variable can be affected by another variable. In that case, we should add an interaction term to the model. To check which variables are interacting with each other we use interaction plots. If lines are parallel in those plots it means that there is no interaction, otherwise, we suspect that there is interaction. After checking we built another model by adding those interactions to the final model we have till now and did variable selection again. Among these forward6 has the best R^2 value, but sse of it is slightly more than the model without interaction. Therefore, there is no need to add the interaction term to the model. So backward2 is still the final model we use.

Transformation of a variable: Since R^2 of the final model is not too high, I suspected that there might be more that I can do. While deleting the variable entry_dt, I assumed that it is irrelevant but I didn't check this assumption. If the subject is dead, entry_dt doesn't say much but if it is not dead, if entry_dt is old time variable will be more, if entry_dt is late then time variable will be small, although two people might be equally healthy. To check this I plotted entry_dt variable graph, Figure 10.14 and Figure 10.15 for status=0 and status=1, since entry_dt is in order in the data just plotting time for the observations will give us the relations. As I suspected for status=1(dead people), there is no relation but for status=0, there is a positive linear relation between time and date. But to use that variable I should transform the coded date into a proper scale. Entry_dt is coded as dddyy, last two digits give the year and first three gives the day of the year. By taking modulus of 100 of this variable, I extracted the year. And by doing integer division by 100 I found the day of the year. Then I found the number of days passed since the first day of 1968 and called this new variable newdate. Since there is an obvious interaction between this variable and status I will add status* newdate variable with newdate variable to the model.

I repeated all steps explained before for this new model and ended up with a model, model3_2. Sse of model 3_2 is better and also R^2 of this model is 7% more than earlier model backward2, so our final model will be model3_2.

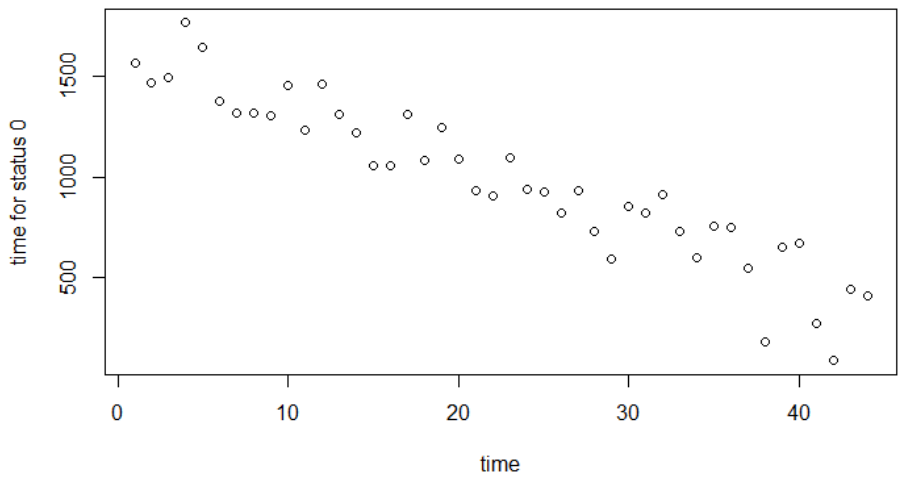


Figure 4. 14

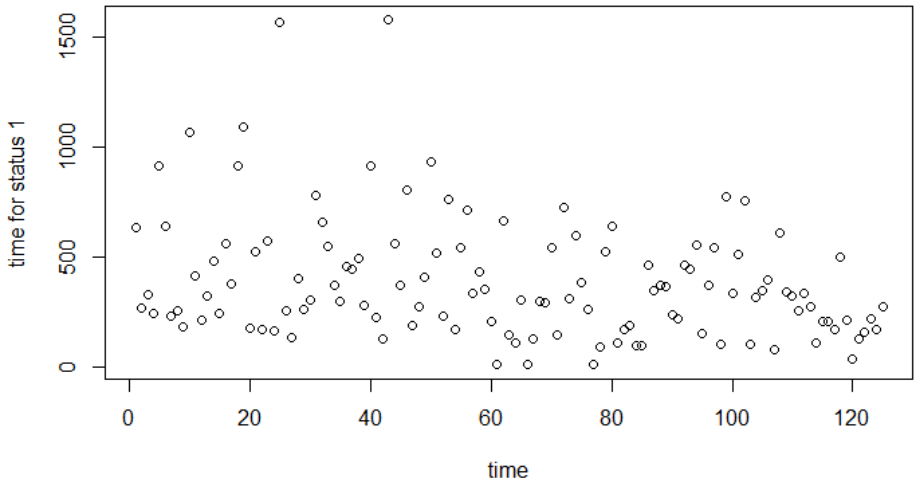


Figure 4. 15

5. DISCUSSION & CONCLUSION:

We ended up with the model;

time ~ status + cond + entry_dt + sex + n_stage + t_stage + status:entry_dt

According to this final model, institution, age, site and treatment are all insignificant variables, which means that we didn't actually need them to predict time variable. So there is no significant effect of receiving different treatments on survival time. Also among the variables in the model status, cond and entry_date are the most significant ones. Sex, n_stage and t_stage are relatively less important.

Also, we should have noticed and done variable transformation at the beginning by carefully implementing explanatory data analysis which would have saved us lots of time. Also even after using all of the explanatory variables we ended up with an R^2 of 62 percent, which is not good enough. Either we lack some important variables which affects survival time or nature of our data is not very compatible with linear regression. Different modeling techniques can be used in the future to get better results.