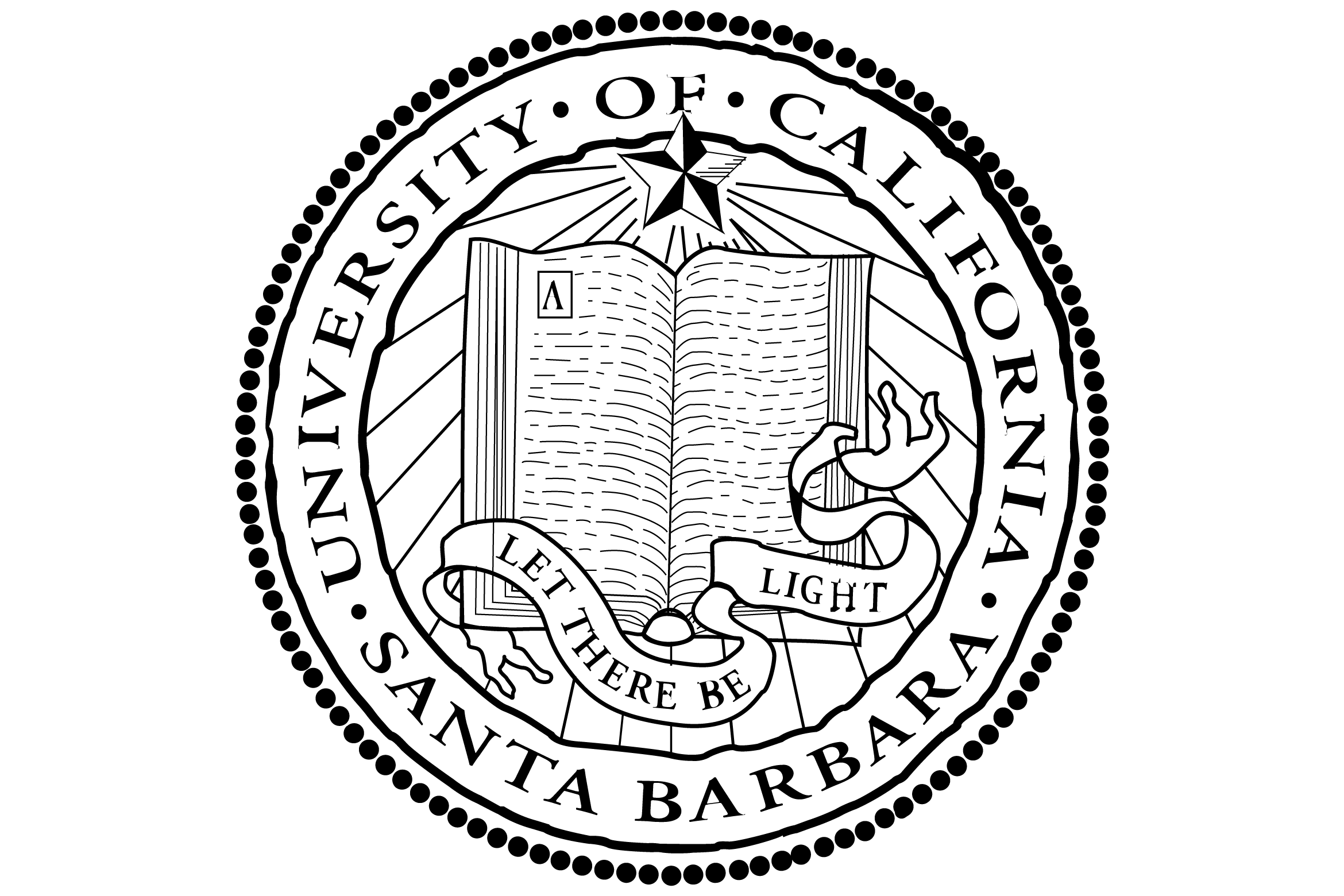
****

University of California, Santa Barbara

PSTAT 175 Final Project

**Drug Relapse Analysis**

Kaiyi Yang, Kerry Wang, Zhongyun Zhang

Instructor: Adam Tashman

5 December 2019

**Abstract**

*Drug abuse is gradually becoming a significant problem in the modern US society. Our group is interested in how different factors including race, prior drug use history, site, intravenous drug use and treatment duration would affect the rate of people returning to drug use after treatment. Hence, we build a Cox Proportional Hazard model to analyze a dataset of 628 drug users.*

**Data Source and Background Information**

The title of data is UIS Study. Our dataset is chosen from the example of the book: Applied Survival Analysis: Regression Modeling of Time to Event Data by authors John Wiley and Sons Inc. The data frame has 628 observations with 7 variables.

We have five fixed covariates: Heroin/cocaine use during 3 months prior to admission, IV drug use history at admission, treatment randomization assignment, subject's race: white or non-white, and treatment site: where the treatment located.

The variables are:

• hercoc (Heroin/cocaine use during 3 months prior to admission), code 1 = heroin and cocaine, 2 = heroin only, 3 = cocaine only, 4 = neither heroin nor cocaine.

• ivhx (IV drug use history at enrollment), code 1 = never, 2 = previous, 3 = recent.

• treat (treatment randomization assignment), code 0 = short, 1 = long.

• site (Treatment Site), code 0 = location A, 1 = location B.

• race (participant race), 0 for white and 1 for other races.

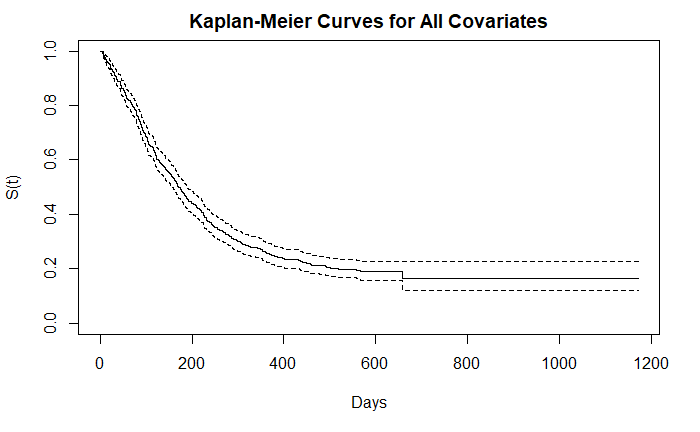
• time (Time to Return to Drug Use), measured from admission, unit: days.

• censor (Returned to Drug Use), code 1 = Returned to Drug use, 0 = Otherwise.

**Research Question**

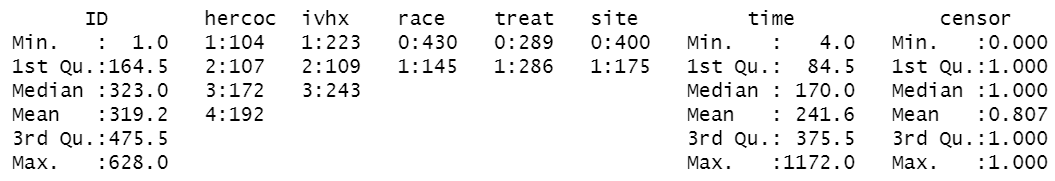
In this project, our group studies whether Heroin/cocaine use during 3 months prior to admission, IV drug use history, treatment length, treatment site or race of subjects will lead to subjects’ returning to drug use, and how these factors affects the survival time.

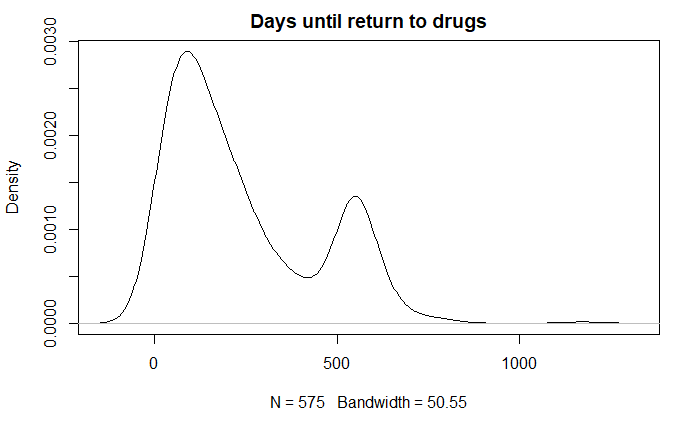
We are also interested in comparing survival time of people from different groups at the same survival probability. Moreover, we want to look for factors that will affect the survival probability differently at different phases of drug rehabilitation..



**Data Exploration**

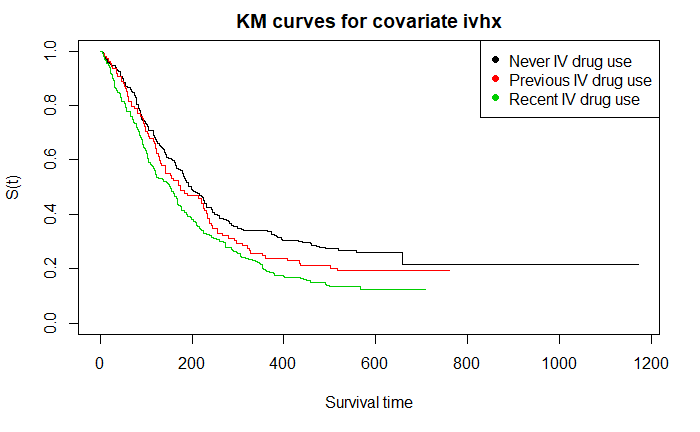
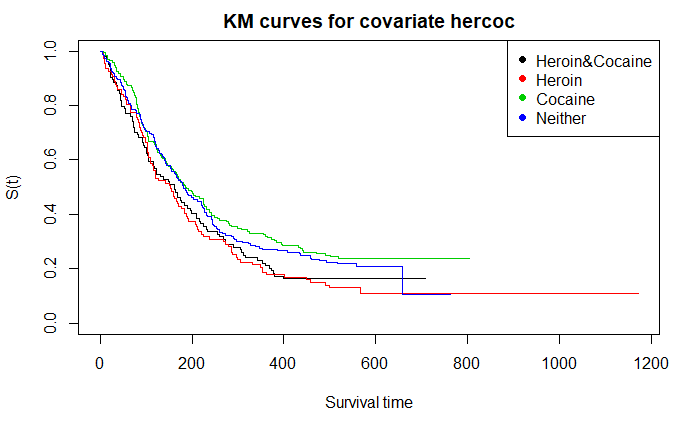
Before doing the Cox PH model, we use the summary function to see details in the data. There are 575 valid data out of 628. From the summary, we can see 223 out of 575 subjects never had intravenous drug use experience. 430 out of 575 subjects are white. 192 out of 575 people had neither Heroin nor Cocaine 3 months prior to admission. Moreover, the treatment randomization assignment for short and long time has almost equal split. About 70% of drag users are assigned to site A, and the rest 30% are assigned to site B.

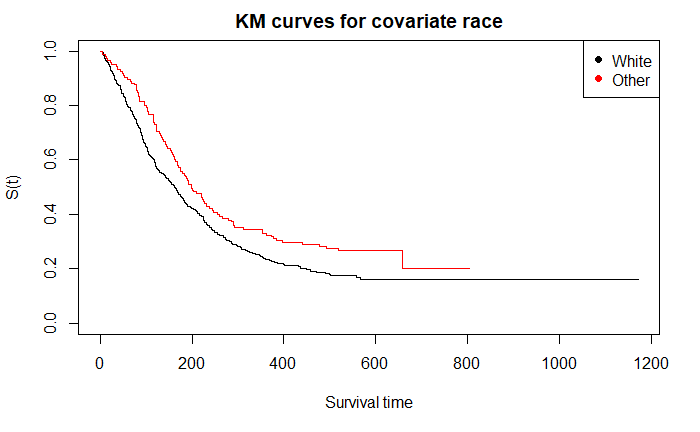


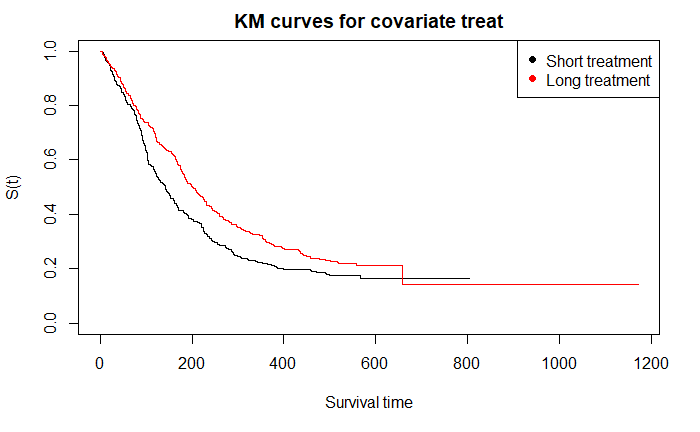


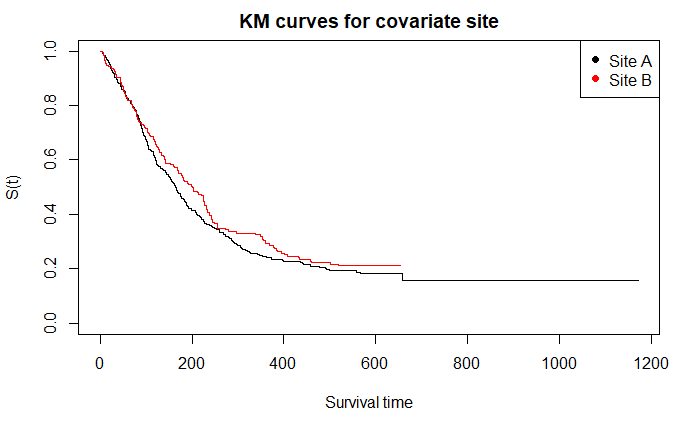
**Kaplan-Meier Estimation Curves**

Now, we plot the Kaplan-Meier survival curve by using these five covariates individually: hercoc, ivhx, race, treat, site. As we can see from the plot, type of drug subjects used seem to have insignificant different to the probability of relapse, since these four curves tangle with each other. Subjects who never had IV drug use history at admission tended to have longer survival time with slightly higher probability not returning to drug use. White people tend to be more likely to return to drug use. People in site A and B have nearly same rate of treatment success. Subjects with longer treatment tend to have slightly higher rate not returning to drug use than subjects with short treatment. However, after around day 650, two survival curves tend to be the same.



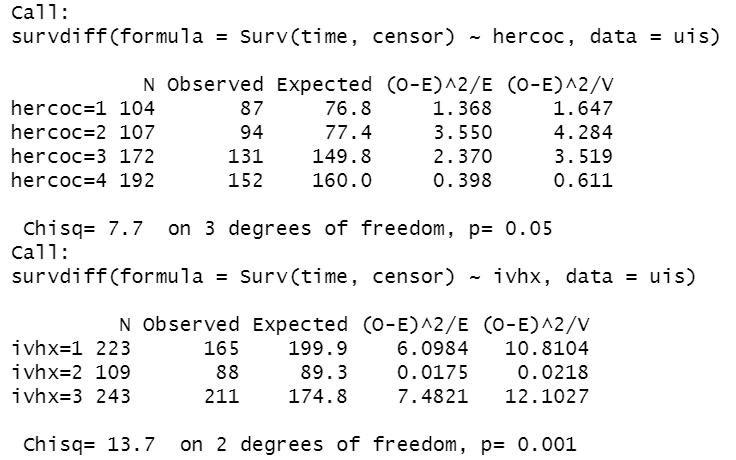


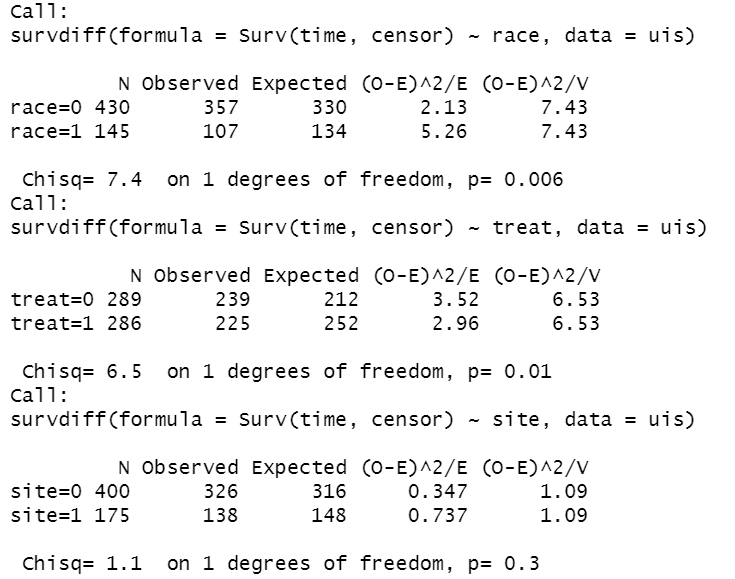




**Log-rank Test**

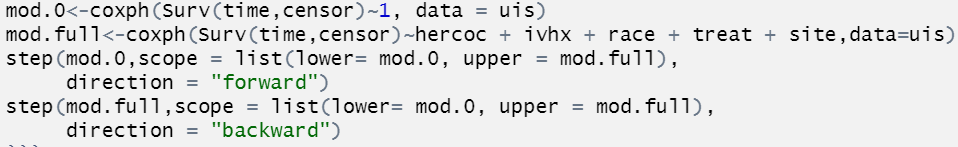
After visualizing the Kaplan-Meier curves, we utilize log rank test to see if every variable had significant effect on the rate of subjects reuse drug. We can see from the data that ivhx, race and treat have small p-value 0.001, 0.006, 0.01 respectively, which are less than 0.05. Variable hercoc and site have p-value greater than 0.05, which means that they have no significant effect on subjects’ drug reuse. Therefore, we will only include ivhx, race, and treat in our further study.

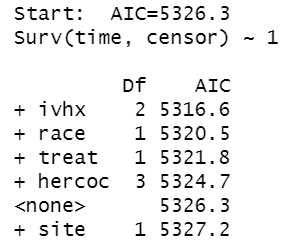
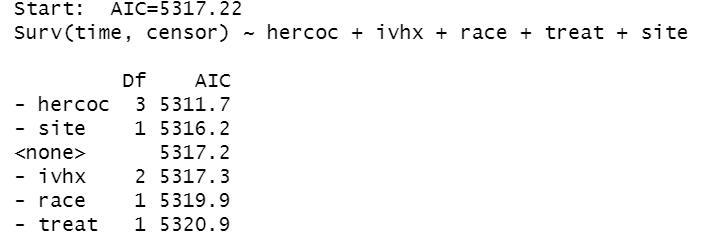




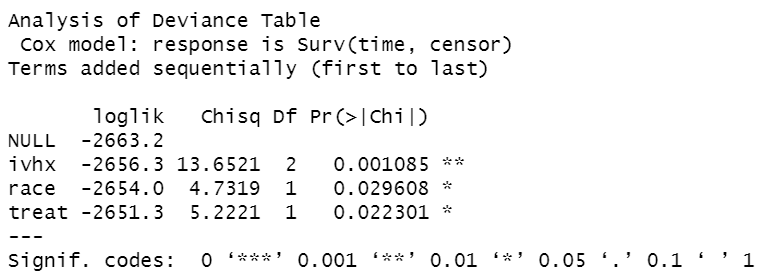
**Model Building**

Before we start to build Cox Proportional Hazard model, we want to use forward selection and backward selection to choose right combination of covariates. To start with, we use forward selection with AIC in R, which gradually adding one variable to the model. The model starts with no variable in and started to add ivhx, treat, race into it. It ends at calling three variables meaning that we should include all of them into our model. Moreover, for backward selection, it starts with full model and reduces one variable in the next model. It also turns out to hold the same result with forward selection, which we should include ivhx, treat, race in our model.



Furthermore, we start to do the likelihood ratio test to choose covariate. We first put individual variable in to the model. Then, include all variables the model. We can see that p-value for the likelihood ratio tests indicate that all these three variables are significant.



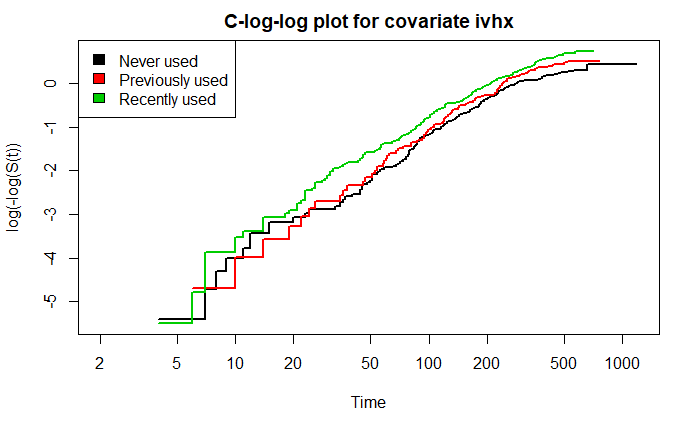
Therefore, we choose the most significant variable ivhx and then add race to build a compare model. As we can see, we should add race into the model. Then, we also add treat into the model and find that it is significant. In the end, we get the same result as the forward selection and backward selection that ivhx, race, treat should be add into the model.

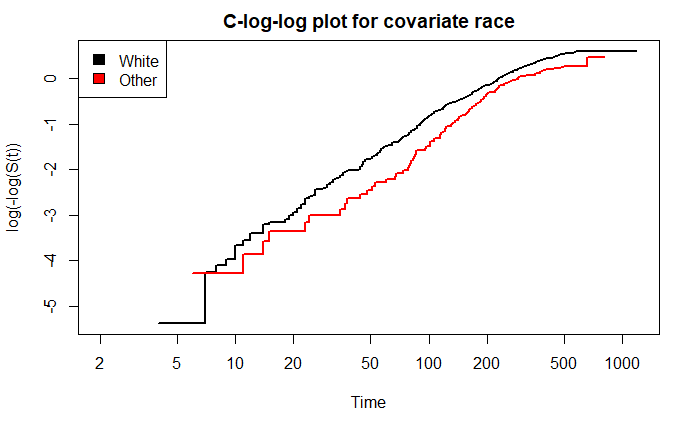
**Model Checking**

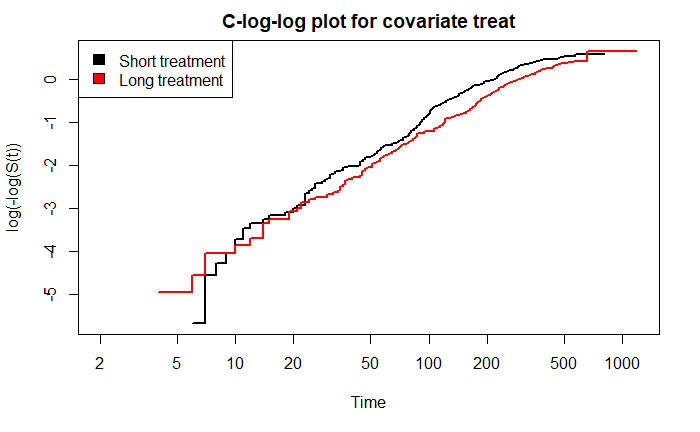
We need to do the residual tests and make C-log-log plot to check if covariates ivhx, race, and treat satisfy the Cox PH assumption before we do the Cox PH model.

**C-log-log Plot**

Now, we start to plot the C-log-log plot to check if covariates satisfy Cox PH. As we can see from the graph, three models have some minor cross overs. We want to arrive at the conclusion: all these three variables pass c-log-log test that we do not need to stratify. However, we are concerned with these cross overs, then we decide to use cox.zph( ) function to justify our conclusion.

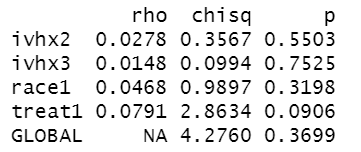






**Residual Test**

cox.zph() is a function in R that performs statistical tests on the PH assumption based on Schoenfeld residuals. As we can see from the result, ivhx, race, and treat all have p-value greater than 0.05. In this way, we decide not to do stratification and these three variables all satisfy the Cox PH assumption.

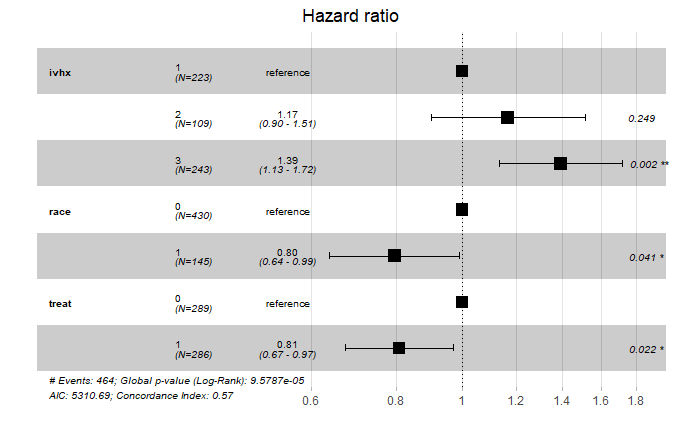


**Interaction Term**

After doing the Cox PH assumption check, we want to see if there are any interaction terms. Three potential interaction candidates are ivhx \* race, ivhx \* treat, race \* treat. By using anova, we can see that none of three interaction terms are significant because they have p-values greater than 0.05. We do not need to stratify any of the variables. Therefore, our final model is still Surv~ivhx + race + treat.

**Hazard Ratios and C.I.**

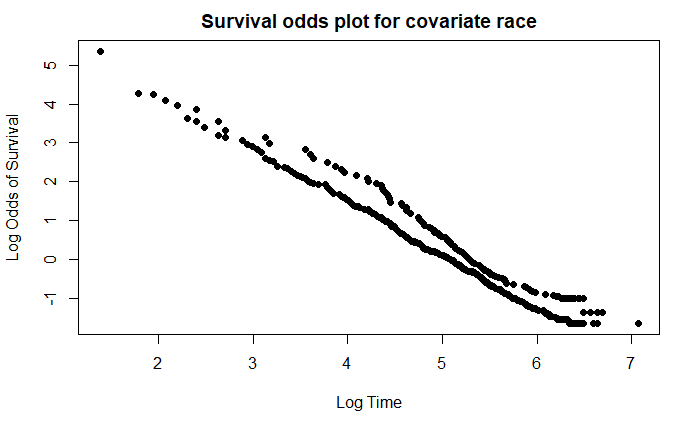
Now, we use ggforest() to create a graph to view the hazard ratio and confidence interval of different covariates in each group.

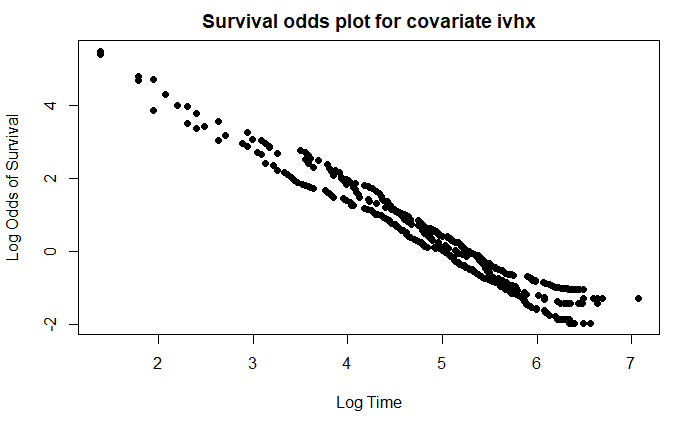


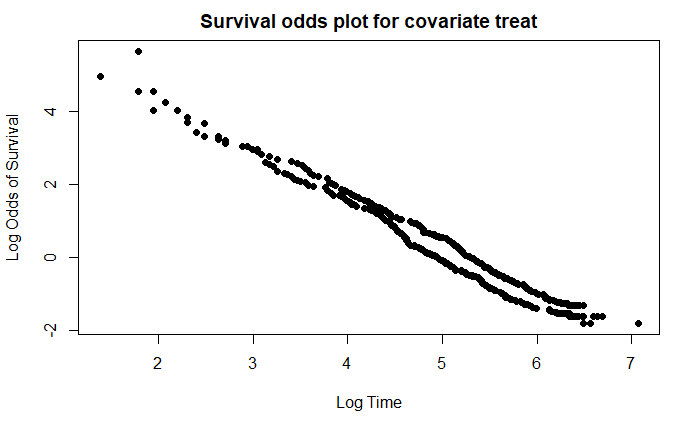
As we look at the hazard ratio in the graph, we can see that hazard ratio for subjects who have previous IV drug use history(ivhx2) are centered at 1.17, and it has a 95% confidence interval between 0.90 and 1.51. Hence, subjects with previous IV Drugs use history at admission is 17% more likely to return to drug use than subjects with no IV Drugs use history. For subjects who recently have IV drug use history at admission(ivhx3) has hazard ratio centered at 1.39, with a 95% confidence interval between 1.13 and 1.72. This means that subjects with recent IV Drugs use history at admission is 39% more likely to return to drug use than subjects with no IV Drugs use history. For hazard ratio of race, non-white(race1) are centered at 0.8, with 95% confidence interval between 0.64 and 0.99. Which means that people who are non-white is 20% less likely to return to drug use than white people. Furthermore, hazard ratio for subjects who had longer treatment(treat1) is centered at 0.81, had 95% confidence interval of 0.67 to 0.97. Subjects who get longer treatment is 19% less likely to return to drug use than those who get shorter treatment.

**Extension 1: AFT Model**

We try to build AFT model in this part. We first draw the survival odds plot for race, ivhx, treat to determine which distribution we are going to use. For each variable, the plot shows a straight line, so we adopt the log-logistic distribution.

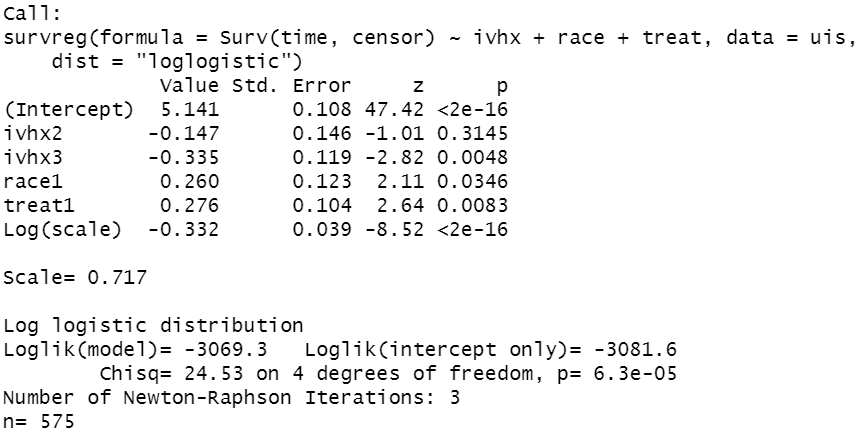






According to the summary table, the estimated acceleration factor between subjects with previous IV drug use history (ivhx=2) and subjects with no IV drug use history(ivhx=1) is 0.86 (e^-0.147). This indicates that the probability of people who never had IV drug use keeping away from relapse for x days is the same with the probability of those who had IV drug use keeping away from relapse for 0.86\*x days.





The estimated acceleration factor between subjects with recent IV drug use history (ivhx=3) and subjects with no IV drug use history(ivhx=1) is 0.72 (e^-0.335). This indicates the probability of people who never have IV drug use keeping away from relapse for x days is the same with the probability of those who had recent IV drug use keeping away from relapse for time 0.72\*x days.

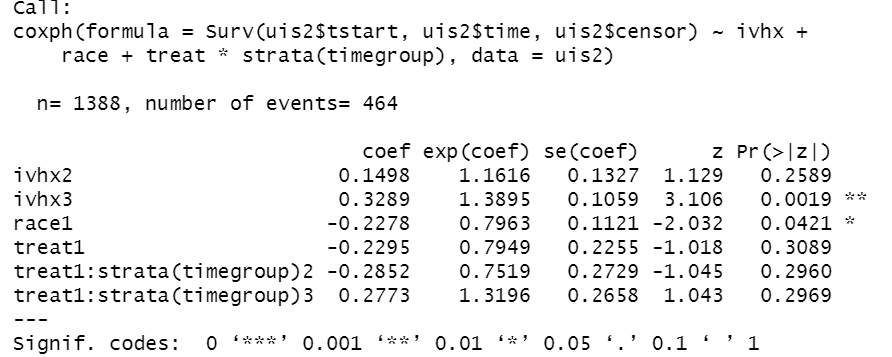
From the above two acceleration factors, we find that at the same survival probability level, the people who had previous or recent IV drug use history keep away from drug use for shorter time, compared with those who never had IV drug use. Especially for those who had recent IV drug use, they return to drug use most quickly.

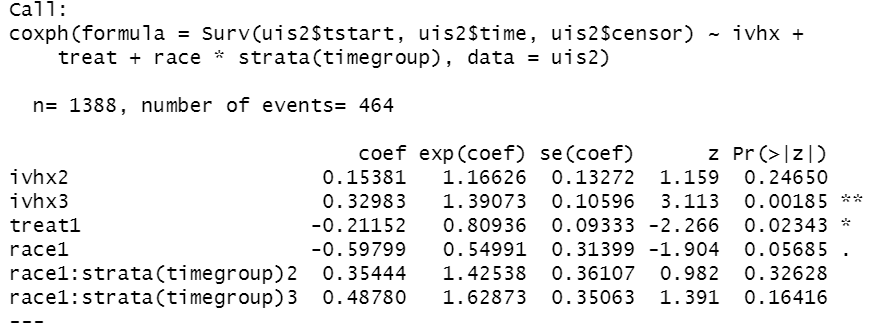
The estimated acceleration factor between white people (race=0) and other races (race=1) is 1.30 (e^0.260). This indicates the probability of white people keeping away from relapse for x days is the same with the probability of non-white keeping away from relapse for 1.30\*x days. Hence, at the same survival probability level, white people keep away from drug use for shorter time compared with subjects of other races.

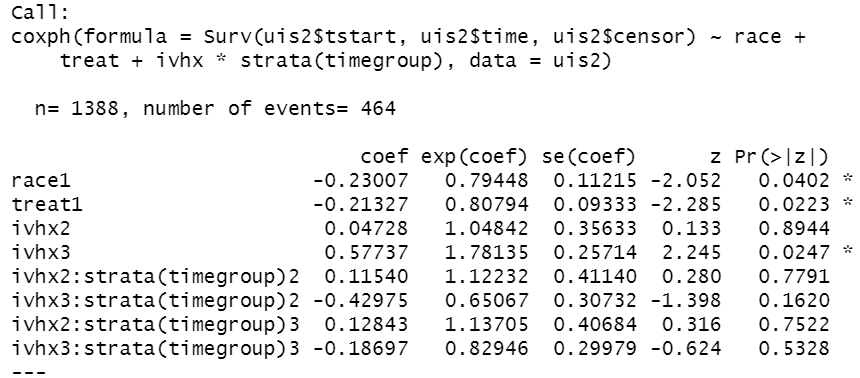
The estimated acceleration factor between long treatment (treat=1) and short treatment (race=0) is 1.32 (e^0.276). This indicates the probability of subjects assigned to short treatment group keeping away from relapse for x days is the same as the probability of subjects assigned to long treatment group keeping away from relapse for 1.32\*x days. Hence, at the same survival probability level, people who get short treatment keep away from relapse for shorter time compared with people who get long treatment.

**Extension 2: Time-Varying Parameter**

For this part, we want to find out whether the covariates’ effect on the probability of returning to drug use varies depend on time. We want to split the survival time into a few appropriate intervals, so we need to find the reasonable cutting point(s). We first look for the cutting points that evenly distributes the subjects to each time interval. We find that when we set 50th day and 150th day as cutting points and split the survival time into three intervals, the number of subjects falls into each time groups are relatively close to each other. 575 subjects has a survival time that falls between or beyond the interval 0-50 days; 495 subjects has a survival time that falls between or beyond the time interval 50-150 days; 318 subjects has a survival time that longer than 150 days. Therefore, the sample size of each time group is big enough to use Cox PH model individually.

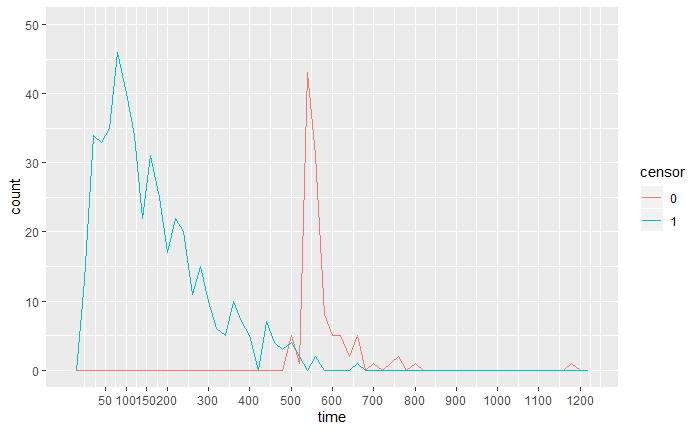






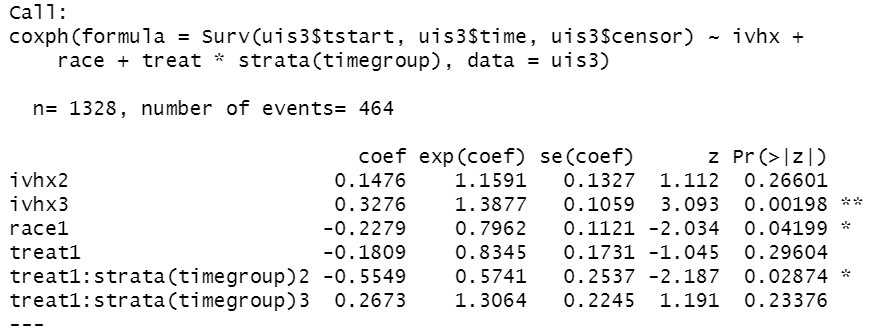
Then we stratify the time group and check if any of our variables has interaction with the time groups. We check one variable at a time, so we build three models. For each of the model, all the interaction terms are insignificant. Therefore, none of our three variables affect survival probability differently depending on time, when we cut survival time at the 50th day and the 150th day.

Since the above cutting points do not give significant interaction terms, we decide to choose better cutting point(s) by looking at the event frequency plot.



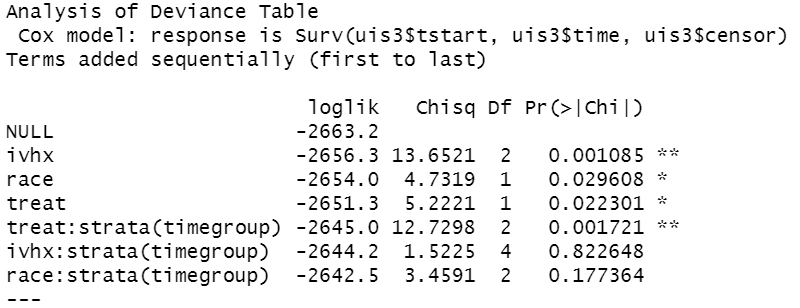
The frequency plot show the number of events and censored subjects at different time. We are looking for the points at which significant declines occur. The survival probabilities before and after such points are very different, which probably indicates a transition from one phase of drug rehabilitation to another. From the plot, we find two major declines occurred at around the 80th day and the 155th day.

Then, we split the survival time at the 80 and 155. 575 subjects has a survival time that falls between or beyond the interval 0-50 days; 440 subjects has a survival time that falls between or beyond the time interval 50-150 days; 313 subjects has a survival time that longer than 150 days. Again, the sample size of each time group is big enough to use Cox PH model individually.

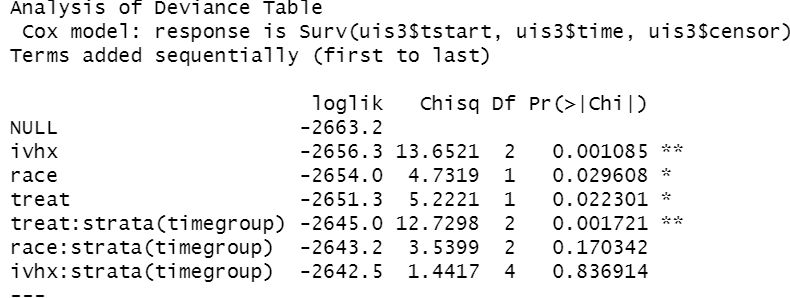


Again, we stratify the time group and look for any significant interaction between each variable and time groups. Only one interaction term treat\*strata(timegroup)2 is significant. Its p-value is 0.02874. Therefore, the effect of treatment on survival rate is likely to vary depending on time. To confirm the model, we use anova function to perform likelihood ratio test.

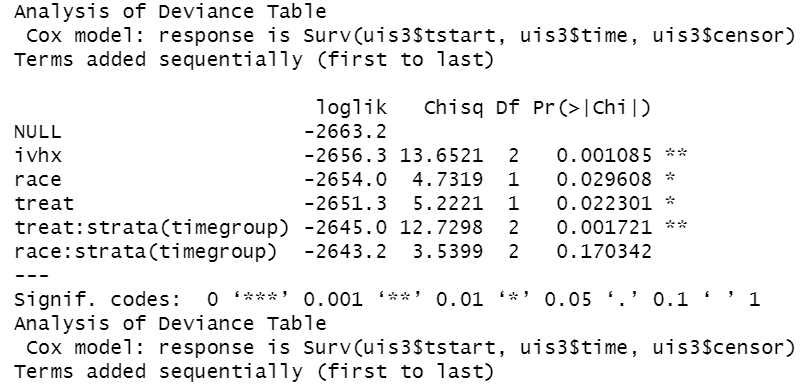
Because anova is an order-sensitive function, we need to determine the order of the variables. We first run anova with all the variables stratified by timegroup. The variable with smallest p-value should be included first. Then, with all the variables arranged in correct order, we run anova again. Finally, we drop a few variables that have large p-values until all the terms in table is significant.



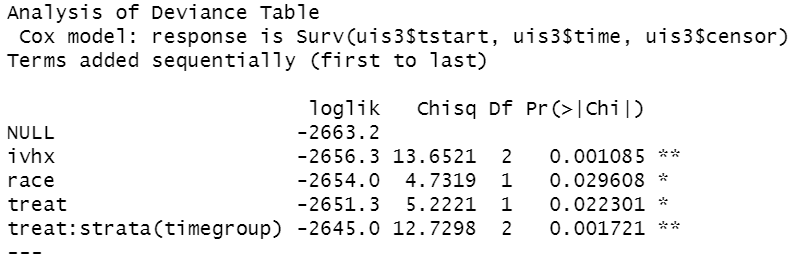
According to the p-values, the model should include variables in the order of treat, race and ivhx. We arrange three variables in the correct order and run anova again:



We find that ivhx has the largest p-value 0.836914. Hence, we first remove ivhx from the model and rerun the anova function:



There is still an insignificant term race in the table. We remove race to make sure all the terms in the table are significant.



After removing ivhx and race, the only variable included in the model is treat. This result agrees with the previous conclusion that treatment is the only variable whose effect on survival probability is time-dependent.

**Conclusion**

For this project, we choose a dataset of 628 former drug users. Besides the survival time until event and the event indicator, we are given five variables that respectively provide the subjects’ race, treatment length, treatment site, intravenous drug use history and drug type of drug used. To clean the dataset, we first delete 53 subjects with missing values. We plot the Kaplan-Meier survival curve for each variable to check if any of them significantly affect the probability of relapse. From the plots, variables ivhx, race and treat show significance and the other two variable, hercoc and site seem to be insignificant. To confirm the model, we use log-rank test, forward and backward selection and likelihood ratio tests. All of the different methods give the same conclusion. Using C-log-log plot and residual test, we confirm that the all of three variables satisfy Cox PH assumptions, so none of the them needs to be stratified. We then check for any significant interaction terms, and we find that none of interaction terms is significant. Finally, we estimate the coefficients and confidence interval for each variable and interpret the hazard ratios. We conclude that white individuals who have recent intravenous drug use and took shorter treatment are more likely to return to drug use.

For the first extension part, we build AFT model. We estimate accelerated factor for each variable. We find that at the same survival rate level, having previous and recent intravenous drug use accelerate the survival time by 0.86 and 0.715 respectively. Accepting longer treatment prolong and being non-white prolongs the survival time by 1.32 and 1.30 time respectively. For the second extension part, we examine the time-dependent parameters. We first find the appropriate cutting points to split survival time into three reasonable intervals. Then we look for any significant interaction between the variables and the time groups. The results show that only treatment has varying influence on survival probability depending on time. The conclusion is confirmed by anova table. We conclude that drug users who took longer treatment are less likely to return to drug use between the 80th day and the 155th day compared with those who took shorter treatment.