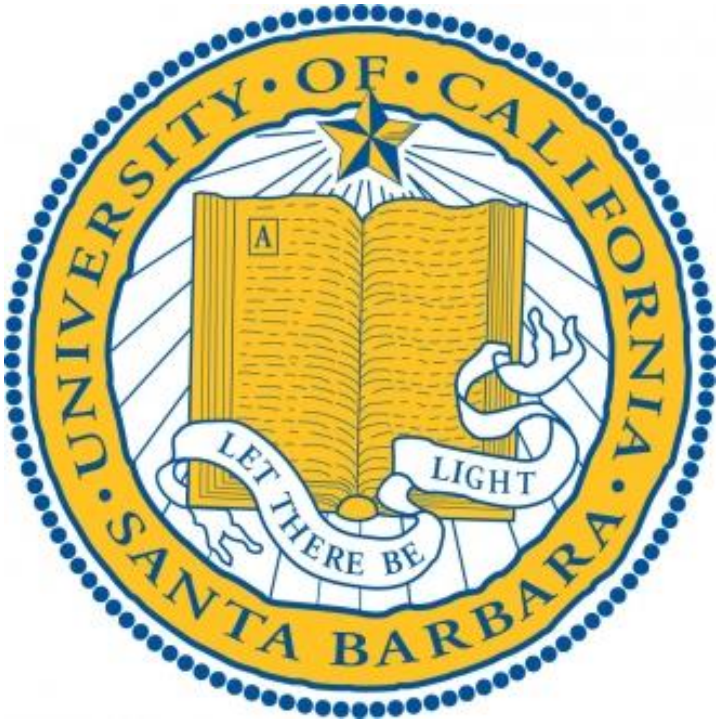


Survival Analysis Project Report

Time Return to Drug Use

Member: Pancheng(Kevin) Qu, Xiangheng Shi, Yiwei Fei, Athena Nguyen

Abstract

In today, drug is still a serious problem that many people think it is hard to stop people from retaking the drug. Many drug users can not control themselves to be away from drugs, so our group project topic is to discuss the effects of race, term of treatment and drug use history on the time return to drug use. We have 628 observations and will use Cox Hazard proportional model.

Sources and Background Information

The dataset is from Table 1.3 of Hosmer, D.W. and Lemeshow, S. and May, S. (2008) Applied Survival Analysis: Regression Modeling of Time to Event Data: Second Edition, John Wiley and Sons Inc., New York, NY. The dataset includes 628 observations and 12 variables.

Variables:

1. ID (Identification Code): the identification code is range from 1 – 628.

2. Age (Age at Enrollment): the unit of age is in Years

3. Beck (Beck Depression Score at Admission): the score is range from 0 to 54

4. Hercoc (Heroin/Cocaine Use During 3 Months Prior to Admission): code as 1 = Heroin & Cocaine, 2 = Heroin Only, 3 = Cocaine Only, 4 = Neither Heroin nor Cocaine (not significant)

5. Ivhx (IV Drug Use History at Admission): code as 1 = Never, 2 = Previous, 3 = Recent

6. Ndrugtx (Number of Prior Drug Treatments) : number ranged from 0 to 40

7. race (Subject's Race): code as 0 = White, 1 = Other

8. treat (Treatment Term): code as 0 = Short, 1 = Long

9. site (Treatment Site): code as 0 = A, 1 = B

10. los (Length of Treatment Measured from Admission): unit in Days

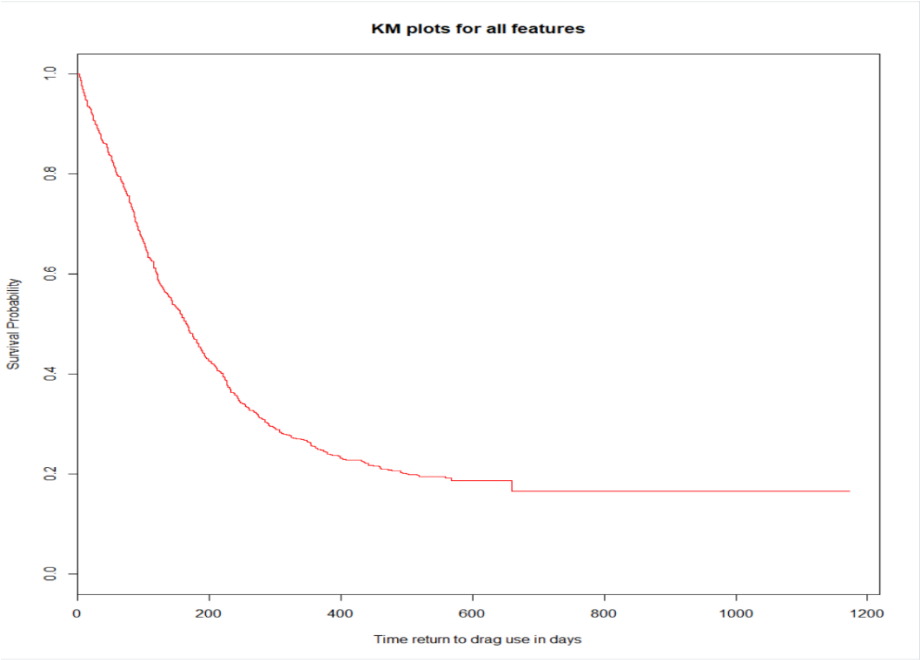
11. time (Time to Return to Drug Use Measured from Admission): unit in Days

12. censor (Returned to Drug Use): code as 1 = Returned to Drug Use, 0 = Otherwise

	ID	age	beck	hercoc	ivhx	ndrugtx	race	treat	site	los	time	censored
1	1	39	9	4	3	1	0	1	0	123	188	1
2	2	33	34	4	2	8	0	1	0	25	26	1
3	3	33	10	2	3	3	0	1	0	7	207	1
4	4	32	20	4	3	1	0	0	0	66	144	1
5	5	24	5	2	1	5	1	1	0	173	551	0
6	6	30	32.55	3	3	1	0	1	0	16	32	1
7	7	39	19	4	3	34	0	1	0	179	459	1
8	8	27	10	4	3	2	0	1	0	21	22	1
9	9	40	29	2	3	3	0	1	0	176	210	1
10	10	36	25	2	3	7	0	1	0	124	184	1

Research Question:

We are interested in whether difference between race(race), treatment term (treat), or drug use history(ivhx) would have significant influence on time from treatment to return to drug use. In addition, we want to see if there are any interactions between any of these covariates: treat, race and drug use history. Here are Kaplan-Meier plots for all features:



How to choose variables to analyze:

From the dataset, we find that we have 12 variables including ID, time and censored. After we exclude those three variables. We think beck variable which means beck depression score at admission is not valid variable to analyze because the we don't even know the meaning of depression score and it is too subjective.

In addition, for the variables age, ndrugtx and los, we don't know how to categorize them. For example, it is vague to separate people to young and old by age since there is no clear definition how many age is old or young. Ndrugtx and los variables also have too many values and hard to stratify.

Now we have 5 variables---hercoc, ivhx, race, treat and site left. We want to see which of them are significant. We decide to use forward and backward selection to determine whether they are significant.

```
## Step:  AIC=6621.46
## time ~ race + treat + ivhx
##
##         Df Sum of Sq      RSS    AIC
## <none>               23807997 6621.5
## + site   1     56503 23751494 6622.0
## + hercoc 3     96240 23711757 6624.9
```

```
## Step:  AIC=6621.46
## time ~ ivhx + race + treat
##
##         Df Sum of Sq      RSS    AIC
## <none>               23807997 6621.5
## - ivhx   2    250130 24058127 6624.0
## - race   1    176905 23984902 6624.1
## - treat  1    220915 24028912 6625.2
```

```
## Analysis of Deviance Table
##  Cox model: response is newuis.surv
## Terms added sequentially (first to last)
##
##         loglik   Chisq Df Pr(>|Chi|)
## NULL   -2959.6
## ivhx   -2954.1 10.8859  2   0.004327 **
## race   -2951.4  5.4913  1   0.019112 *
## treat  -2948.7  5.3165  1   0.021125 *
## site   -2947.5  2.4984  1   0.113959
## hercoc -2946.7  1.5922  3   0.661164
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

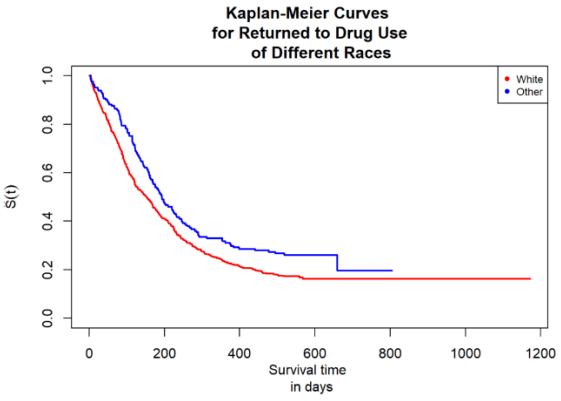
We also use ANOVA table (LRT) to check covariates' significance. Ivhx, race, and treat's p-values are all lower than 0.05, showing that they are significant. According to the forward and backward selections and ANOVA table, we find only ivhx, race, and treat are significant. Thus, we decide to include only these three covariate in our model.

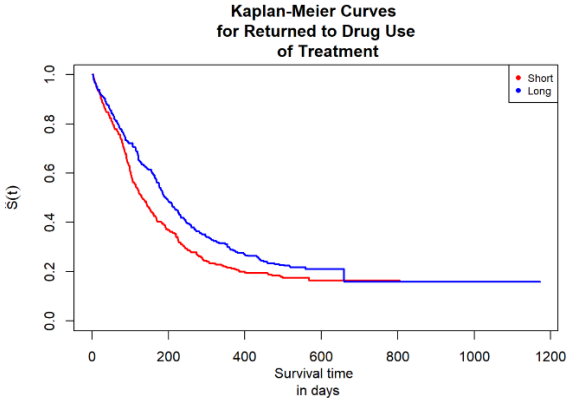
Kaplan-Meier Estimation Curves:

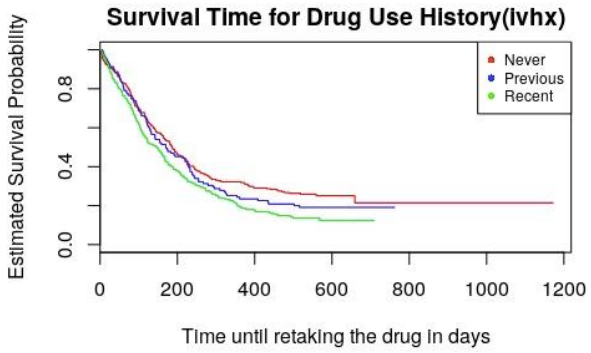
After choosing the variable we want to analyze, we want to plot Kaplan-Meier plots for all three variables, treat, ivhx and race individually. Clearly we can see that from the race Kaplan-Meier plot, the white people have worse performance on longer period back to drug use. From

the treat aspect, the long term treatment patients can hold longer than short term treatment patient. In the end, the ivhx Kaplan-Meier plot shows that people who have recent drug use history have worst performance and people who never have drug use history have best performance.

In addition, we also suspect there is a significant outlier since we find all three Kaplan-Meier plots have a long tail. We want to find this outlier and exclude it.







Cleaning data and excluding the outlier:

Check the outlier from outlier from race Kaplan-Meier plot

```
> newuis[(newuis[,7]==0)&(newuis[,11]>1000),]
     ID age beck hercoc ivhx ndrugtx race treat site los time censored
112 112  35   11      2    1       3    0     1    0  51 1172        0
```

Check the outlier from outlier from treat Kaplan-Meier plot

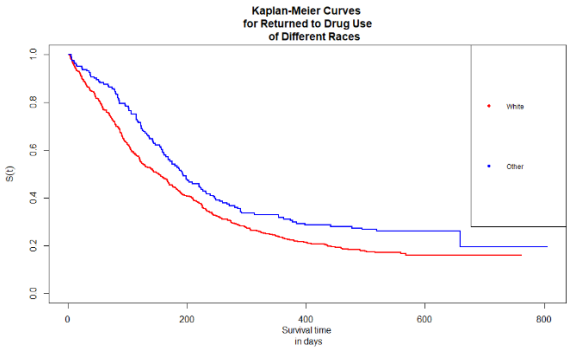
```
> newuis[(newuis[,8]==1)&(newuis[,11]>1000),]
     ID age beck hercoc ivhx ndrugtx race treat site los time censored
112 112  35   11      2    1       3    0     1    0  51 1172        0
```

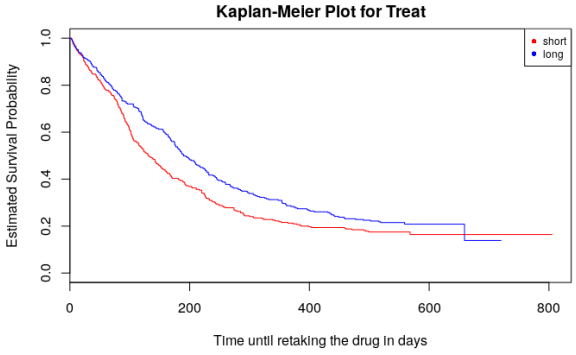
Check the outlier from outlier from ivhx Kaplan-Meier plot

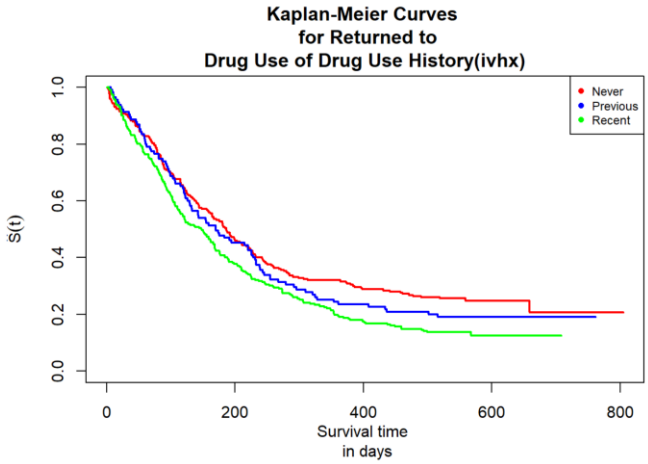
```
> newuis[(newuis[,5]==1)&(newuis[,11]>1000),]
     ID age beck hercoc ivhx ndrugtx race treat site los time censored
112 112  35   11      2    1       3    0     1    0  51 1172        0
```

Since we check from three Kaplan-Meier plots and ID 112 is an obvious outlier, we decide to exclude it from the data and then continue our analysis.

New Kaplan-Meier plots for three variables:







After removing the outlier, we find that all three Kaplan-Meier plots do not have tails and looks better now.

Log rank test:

When we finish the Kaplan-Meier plots for three variables, we still want to use log rank test to see if they have significant different effects on time return to drug use. From P value we find that the race, treat and ivhx have significant different effects since their P value is less than 0.05. It is reasonable because their Kaplan-Meier plots have obvious differences.

```
> survdiff(formula=uis.surv~race,data = newuis)
Call:
survdiff(formula = uis.surv ~ race, data = newuis)

         N Observed Expected (O-E)^2/E (O-E)^2/V
race=0 466      388      359      2.27      7.81
race=1 161      120      149      5.50      7.81

 Chisq= 7.8  on 1 degrees of freedom, p= 0.005
```

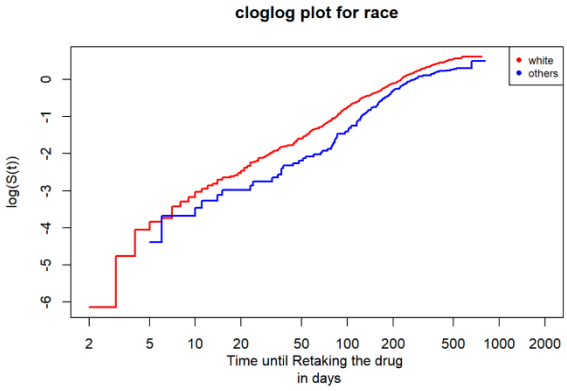
```
> survdiff(formula=uis.surv~treat,data = newuis)
Call:
survdiff(formula = uis.surv ~ treat, data = newuis)

          N Observed Expected (O-E)^2/E (O-E)^2/V
treat=0 320      265      237      3.40      6.42
treat=1 307      243      271      2.97      6.42

 Chisq= 6.4  on 1 degrees of freedom, p= 0.01
```

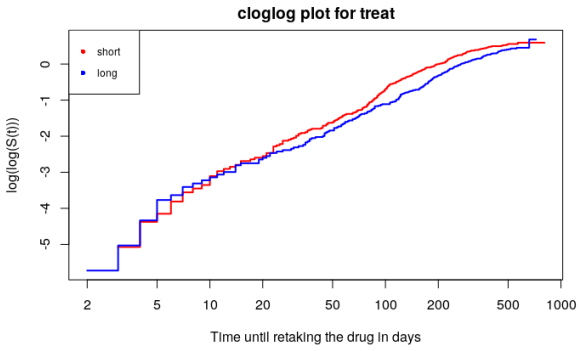
```
> survdiff(formula=uis.surv~ivhx,data = newuis)
Call:
survdiff(formula = uis.surv ~ ivhx, data = newuis)

          N Observed Expected (O-E)^2/E (O-E)^2/V
ivhx=1 250      188    219.3    4.4656    7.9233
ivhx=2 115       93     95.4    0.0593    0.0733
ivhx=3 262      227    193.3    5.8648    9.5418

 Chisq= 10.5  on 2 degrees of freedom, p= 0.005
```

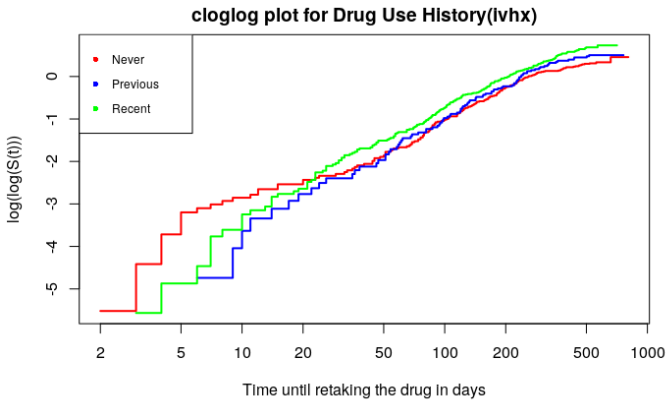
PH assumption check:

CLog-log plots

We use clog-log plot to check if the variable satisfies the PH assumptions. From the plots we can see that race clearly satisfies PH assumptions. However, for the treat and ivhx variable, we see there are some crosses between curves. To determine if it satisfies the PH assumptions, we need to use residual test which is cox.zph() to test again.







Residual Tests:

We can use cox.zph() function to test the PH assumption for treat and ivhx variables again. From cox.zph() test for treat, we find that although the P value of treat is 0.0717 which is close to 0.05 but they are still greater than 0.05. Therefore, we conclude that treat variable satisfies the PH assumption and we do not need to stratify it.

```
> cox.zph(cox0)
                rho chisq      p
newuis$treat 0.0802  3.24 0.0717
```

Then we did the same thing to ivhx variable, From the cox.zph() result, the P value of Global is 0.776 which is greater than 0.05. Therefore, we conclude ivhx also satisfies the PH assumption and we don't need to stratify it.

```
> cox.zph(cox1)
                  rho chisq     p
newuis$ivhx1 -0.0346 0.611 0.434
newuis$ivhx2  0.0164 0.136 0.712
newuis$ivhx3      NA   NaN   NaN
GLOBAL            NA 1.104 0.776
```

Interaction Effects:

This step is to check the interactions between out three variables. We need to check treat*race, race*ivhx, ivhx*treat and the interaction for three variables---race*ivhx*treat. Our

null hypothesis is the reduced model is preferred (without interaction term). Alternative hypothesis is the full model is preferred (with interaction term). First of all, we can do treat*race.

```
> fitSC.int1 = coxph(Surv(time,censored)~race*treat,data=newuis)
> anova(fitSC.int1)
Analysis of Deviance Table
 Cox model: response is Surv(time, censored)
Terms added sequentially (first to last)

           loglik  Chisq Df Pr(>|Chi|)
NULL       -2957.8
race       -2953.7 8.1357  1    0.00434 **
treat      -2950.9 5.6824  1    0.01714 *
race:treat -2949.8 2.1867  1    0.13921
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Since the P value is greater than 0.05, we conclude that the interaction between treat and race is not significant.

The next thing we do is to check interaction between race and ivhx. Since the P value is 0.22290 which is greater than 0.05, we fail to reject the null hypothesis and conclude that there is no interaction between race and ivhx is significant.

```
> fitSC.int2 = coxph(Surv(time,censored)~race*ivhx,data=newuis)
> anova(fitSC.int2)
Analysis of Deviance Table
 Cox model: response is Surv(time, censored)
Terms added sequentially (first to last)

           loglik  Chisq Df Pr(>|Chi|)
NULL       -2957.8
race       -2953.7 8.1357  1    0.00434 **
ivhx       -2949.7 8.0435  2    0.01792 *
race:ivhx -2948.2 3.0021  2    0.22290
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

The third thing we do is to check interaction between ivhx and treat. Since the P value is 0.706185 which is still greater than 0.05. We conclude that there is no significance for interaction between ivhx and treat.

```
> fitSC.int3 = coxph(Surv(time,censored)~treat*ivhx,data=newuis)
> anova(fitSC.int3)
Analysis of Deviance Table
 Cox model: response is Surv(time, censored)
Terms added sequentially (first to last)

           loglik  Chisq Df Pr(>|Chi|)
NULL       -2957.8
treat      -2954.6 6.4023  1   0.011397 *
ivhx       -2949.8 9.4647  2   0.008806 **
treat:ivhx -2949.5 0.6958  2   0.706185
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

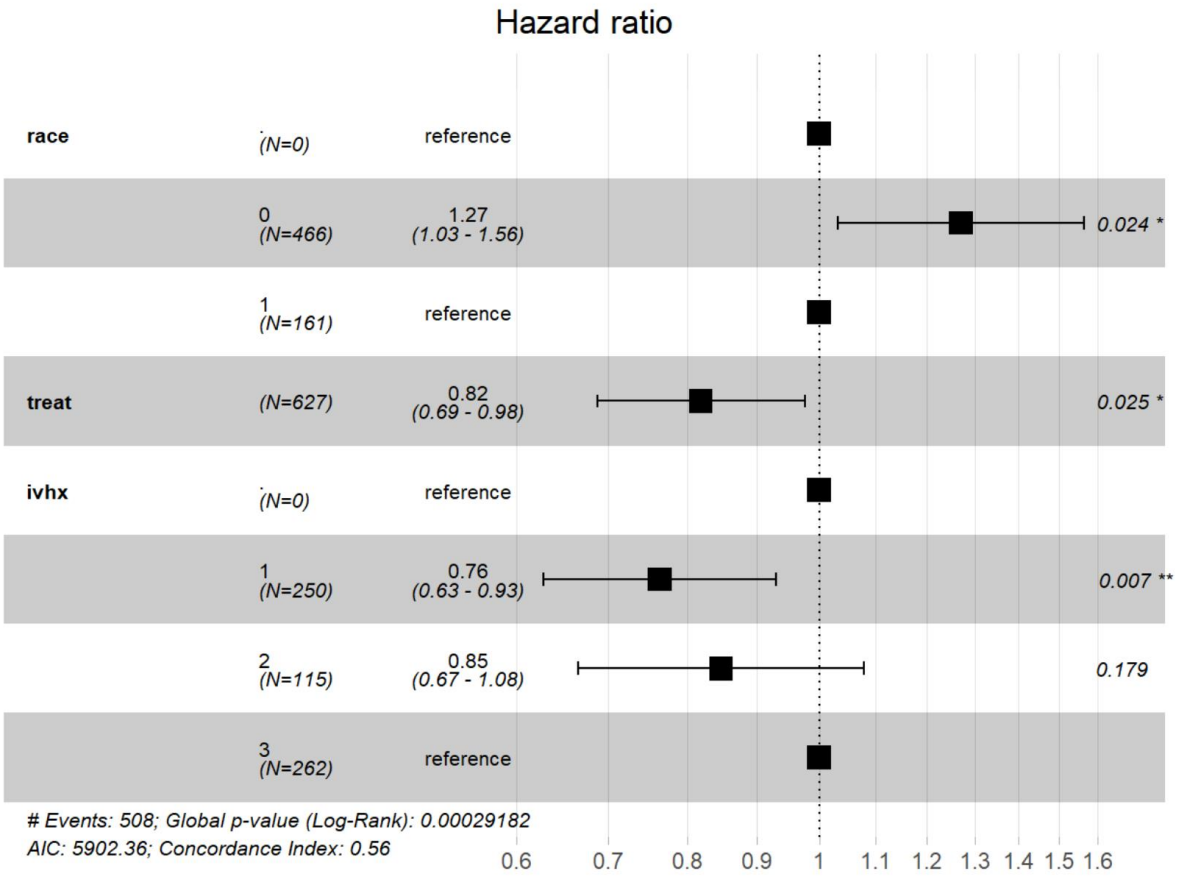
Then we have to check the last thing which is the interaction between three variables. From the last P value which is 0.319256 and it is greater than 0.05. We conclude that there is no significant interaction between these three variables.

```
## Analysis of Deviance Table
##  Cox model: response is newuis.surv
## Terms added sequentially (first to last)
##
##                  loglik   Chisq Df Pr(>|Chi|)
## NULL            -2957.8
## ivhx            -2952.6 10.3985  2   0.005521 **
## race            -2949.7  5.7807  1   0.016203 *
## treat           -2947.2  4.9998  1   0.025351 *
## ivhx:race       -2945.6  3.0958  2   0.212693
## ivhx:treat      -2945.3  0.6276  2   0.730649
## race:treat      -2944.4  1.8611  1   0.172503
## ivhx:race:treat -2943.2  2.2835  2   0.319256
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Therefore, since we do not have interaction between all three variables and all variables satisfy the PH assumptions. Thus, we determine our final model is:

```
> coxph(formula= uis.surv~treat+race+ivhx,data=newuis)
Call:
coxph(formula = uis.surv ~ treat + race + ivhx, data = newuis)

          coef exp(coef) se(coef)     z      p
treat -0.2058    0.8140   0.0893 -2.30 0.0212
race0  0.2333    1.2628   0.1058  2.21 0.0274
race1      NA        NA   0.0000    NA     NA
ivhx1 -0.2770    0.7580   0.1001 -2.77 0.0057
ivhx2 -0.1652    0.8477   0.1234 -1.34 0.1805
ivhx3      NA        NA   0.0000    NA     NA

Likelihood ratio test=21.69  on 4 df, p=2e-04
n= 628, number of events= 508
```

Hazard ratio and CI:



Looking at the hazard ratio chart, we know that hazard ratio of white people is centered at 1.27 and its 95% confidence interval is between 1.03 and 1.56. It indicates that white people

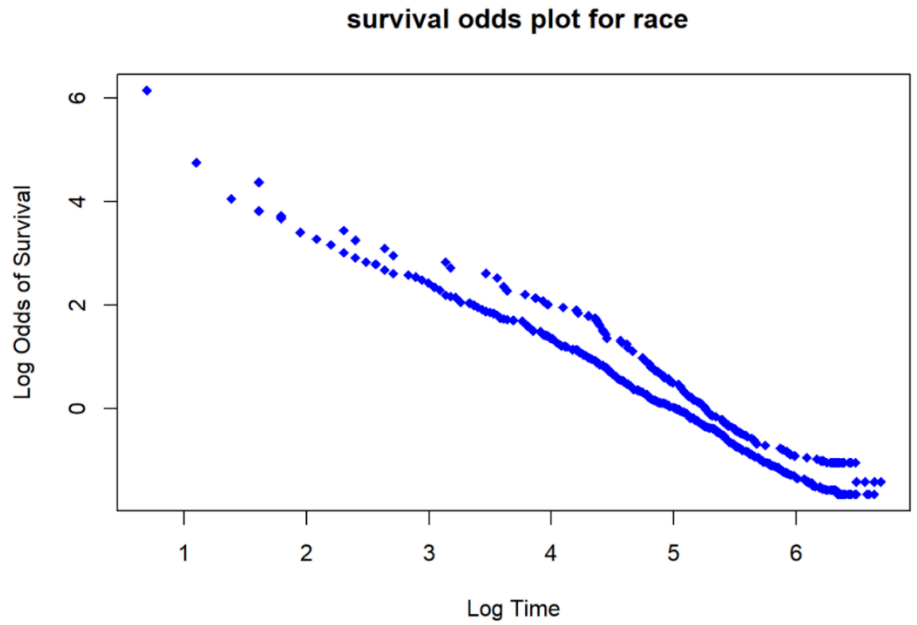
have 27% more likelihood to reuse the drugs than other races.

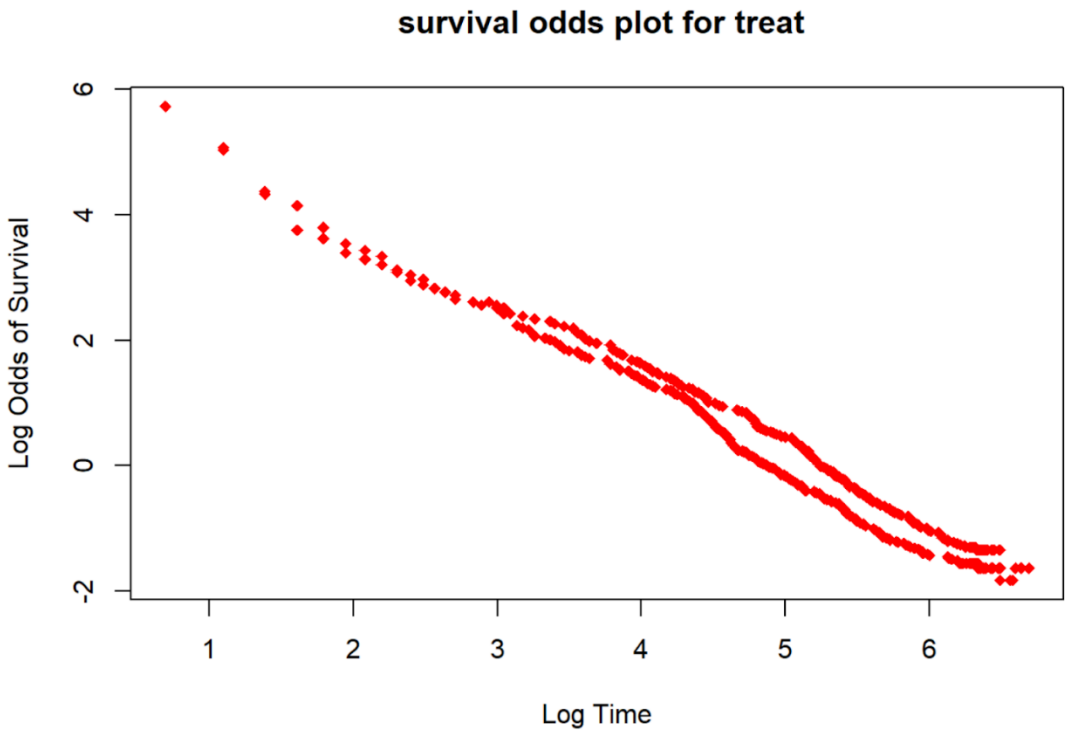
The hazard ratio of ivhx1 is 0.76 (95% CI is [0.63,0.93]) compared to ivhx3. It indicates that ivhx1 have 24% less likelihood to reuse drug compared with ivhx3. Moreover, the hazard ratio of ivhx2 is 0.85 (95% CI is [0.67,1.08]) compared to ivhx0, which indicates ivhx2 have 15% less likelihood to reuse drug compared with ivhx3.

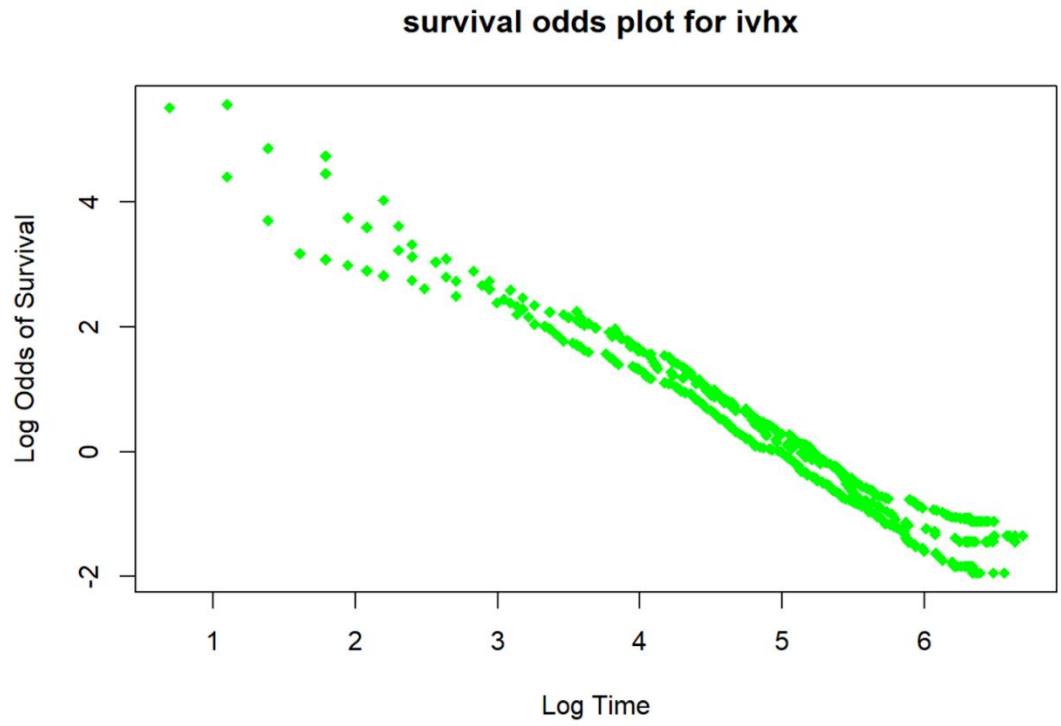
The hazard ratio of long treatment is centered at 0.82 and its 95% confidence interval is (0.69,0.98). It indicates that long treatment has 18% less likelihood to reuse drug than short treatment.

Extension: AFT model

For the extension section, we build an AFT model. We want to use log-logistic distribution in our AFT model. Therefore, we check the survival odds plots for different covariates. The straight lines indicate that we can use log-logistic distribution. Thus, we use log-logistic distribution in our AFT model.







Since the lines are straight, we can use the log-logistic distribution.

```
##
## Call:
## survreg(formula = newuis.surv ~ ivhx + race + treat, data = newuis,
##     dist = "loglogistic", init = TRUE)
##               Value Std. Error     z       p
## (Intercept)  5.0693     0.1455 34.85 < 2e-16
## ivhx1        0.2448     0.1230  1.99   0.047
## ivhx2        0.1769     0.1499  1.18   0.238
## ivhx3        0.0000     0.0000    NA      NA
## race0       -0.2987     0.1269 -2.35   0.019
## race1        0.0000     0.0000    NA      NA
## treat        0.2686     0.1089  2.47   0.014
## Log(scale)  -0.2450     0.0375 -6.53 6.8e-11
##
## Scale= 0.783
##
## Log logistic distribution
## Loglik(model)= -3348   Loglik(intercept only)= -3358.2
##  Chisq= 20.41 on 6 degrees of freedom, p= 0.0023
## Number of Newton-Raphson Iterations: 4
## n= 627
```

The estimated acceleration factor γˆ comparing race0 and race1 is 0.74 (e^-0.2987). This leads to S(race1) = S(0.74*race0). Therefore, the survival time for race0(white) is "accelerated" by a factor of 0.74 compared to the race1(other) based on AFT model with log-logistic distribution.

Similarly, the estimated acceleration factor γˆ comparing long treatment and short treatment is 1.31 (e^0.2686). S(short treatment)=S(1.31*long treatment). The survival time for “long treatment” is "accelerated" by a factor of 1.31 compared to the “short treatment.”

The estimated acceleration factor γˆ comparing drug use history is a little complex. From the table, we find that S(recent use)=S(1.19*previous use) (e^0.1769=1.19). S(recent use)=S(1.28*never use) (e^0.2448=1.28). The survival time for “previous use” is "accelerated"

by a factor of 1.19 compared to the “recent use.” The survival time for “never use” is "accelerated" by a factor of 1.28 compared to the “recent use.”

In a nutshell, we find that covariate “race” has the highest effect on survival time, based on AFT model with log-logistic distribution.

Conclusion/Summary:

In the beginning of the project, after we decide our dataset which has 628 observations and 12 variables, we start to graph Kaplan-Meier plots for all features to check if the dataset has lots of censored data. After finishing this procedure, we select three significant variables---treat, race, ivhx, by backward and forward selections. From the result of ANOVA table, we find only these three variables are significant.

Then we start to graph Kaplan-Meier plots separately and we discover that all of our plots have long tails. Therefore, we suspect there is an outlier and start to check it. Luckily, we found this outlier and decide to exclude it from our dataset. The Kaplan-Meier plots after we exclude the outlier look better and no longer have long tails.

The third thing we did is to check the PH assumptions. First of all, we use clog-log plots to check if there are cross between curves or there are divergence between curves. The result is that we find ivhx and treat variable have some crosses in the plots. Therefore, we once again use cox.zph test to test their PH assumption again. Fortunately, we find they still satisfy the PH assumptions and we don’t need to stratify any variables.

The next step we did is to check the interaction between three variables including ivhx*treat, ivhx*race, race*treat and race*treat*ivhx. After using the ANOVA table to know

the P values, we find that they do not have interactions. Hence, we get our final model and generate our conclusion to questions. The result indicates that the white race people may be more likely to reuse the drugs after the treatment. The people who have shorter treatment will have higher probability to reuse the drugs. In addition, the people who have most recent use of drug history will have higher probability to use the drugs again.

For the extension, we use AFT model. From the survival odds plots, we decide to use log-logistic distribution to predict. Estimating $\gamma\hat{}$ (acceleration factor), we find that white people have an accelerated "speed" to reuse the drug compared to other races. Short treatment also has an accelerated "speed" reuse the drug compared to long treatment. Besides, people who recently use drug have an accelerated "speed" reuse the drug compared to people who never used drug. The covariate "race" has the highest effect on survival time under this model.