Pstat 175 Final Project

12.04.2019

Instructor : Adam Tashman

Jasmine Wang Junyue Wang

Abstract

The latest Government study of recidivism reported that 83% of state prisoners were arrested at some point in the 9 years following their release. A majority of those were rearrested within the first 3 years, and more than 50% get rearrested within the first year. In our research, we focus on finding the most relevant factors, such as financial aid, race and education which influence convicts' rearrest time after first release.

Data source and Background

We obtained our data set, Rossi, from http://socserv.mcmaster.ca/jfox/Books/Companion/data/Rossi.txt. The data records 432 convicts who were released from Maryland state prisons in the 1970s and indicated their situations one year after release. The event of our interest is arrest, with 1 representing the convict was rearrested one year after release and 0 representing not arrested or censored.

In the data set, we have 8 fix covariances: "fin": whether the convicts receive financial aid or not. Half of the released convicts were assigned randomly to an experiment with financial aid and the other half not. "Race": the convicts are black or not. "wexp": Did they have full-time work experience before incarceration? "mar": whether the convicts were married at the time of release. "age": the convicts' age at the time of release. "paro": whether the convicts were released on parole. "prio": number of convictions prior to current incarceration. Eventually, we have "edu": the level of education of the convicts. The details of variables are listed below:

- week: week of first arrest after release or censoring; all censored observations are censored at 52 weeks.
- arrest: 1 for event and 0 for censoring.

- fin (financial aid): 1 for receiving financial aid and 0 for not receiving financial aid.
- race: coded 1 if convicts are black and 0 if not.
- age: in years at time of release. In this case, we separate variable “age” into 3 groups, 0 for convicts under 20, 1 for convicts between 20 to 27, 2 for convicts over 27.
- wexp: full-time work experience before incarceration, 1 represents yes and 0 represents no.
- mar: marital status at the time of release: 1 for married and 0 for not married.
- paro: 1 represents the convict was released on parole and 0 represents not.
- prio: number of convictions prior to current incarceration. In this case, we separate variable “prio" into 2 groups. 0 for convicts had less than 3 convictions before current incarceration, and 1 for convicts had more or equal to 3 convictions previously.
- educ: level of education. 2 = 6th grade or less; 3= 7th to 9th grade; 4 = 10th to 11th grade; 5 = 12th grade; 6 = some college.
- Notes: We have grouped the variable “age” and “prio” according to summary data. We found it makes more sense to use the grouped data instead of discrete one since we are looking for a general idea of what effect does each individual variable have on convicts’ survival time.

Research and Question

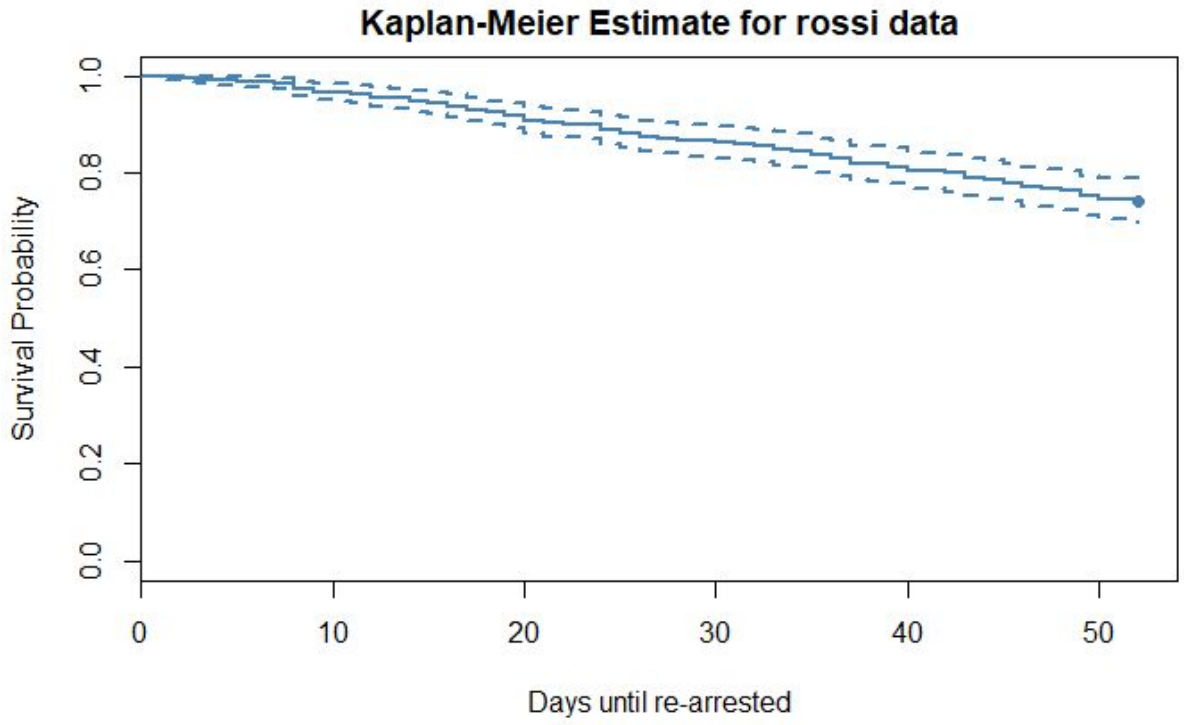
In our project, we are interested in whether receiving financial aid, race, age, having previous full-time experience, marital status, released on parole, number of previous convictions, and education level affect the convicts’ status one year after release and how much effect do these variables have (ie. coefficients analysis). Prior to our analysis, we thought that convicts were less likely to be rearrested if they received financial aid after release / had

previous full-time work experience / released on parole / had earned higher degrees. The final decision will be made in conclusion part.

Furthermore, we want to understand if any interactions exist between fixed covariates, and we are interested in finding the covariate that had the highest effect to our model.

In general, about 80% of test subjects survived one year after release.

Here is the KM curves for all variables in general:



Research Tools

The purpose of building the model is to evaluate simultaneously the effect of several factors on survival. Therefore, we want to use the Cox Proportional Hazards model for analysis. It is a “robust” model, so that the results from using the Cox model will closely approximate the results for the correct parametric model.

Data Exploration

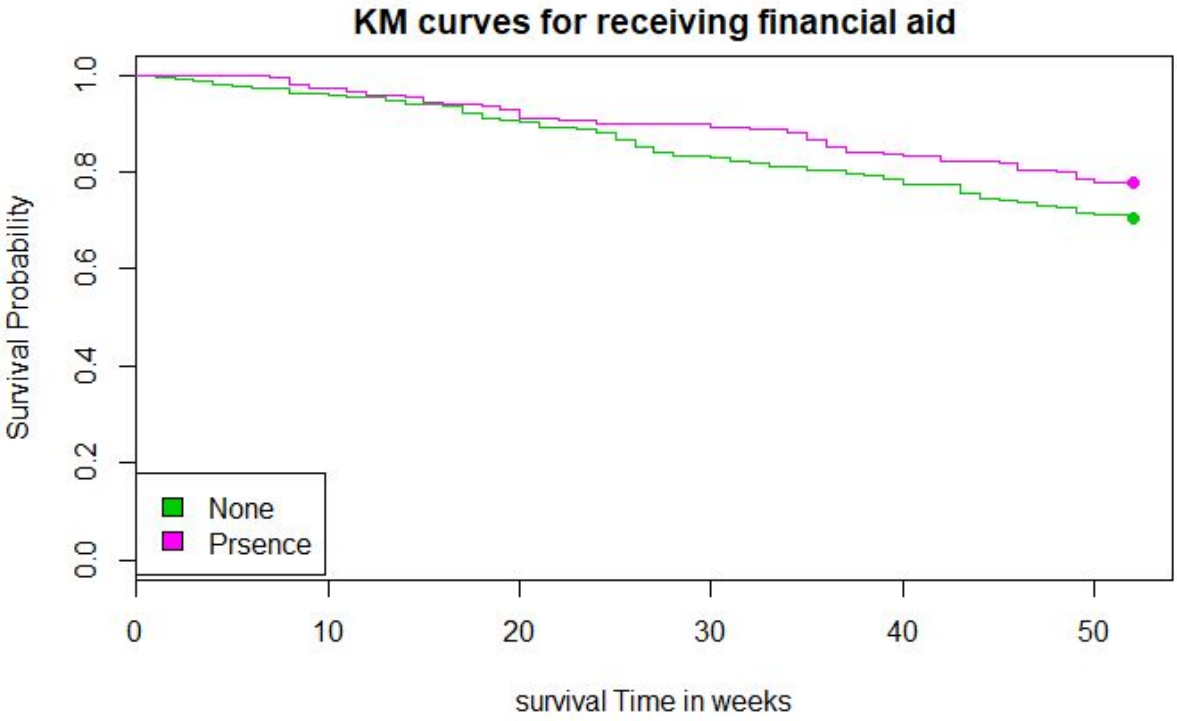
We summarize the data at first to have a general idea. Out of 432 observations, half of the convicts were randomly chosen to receive financial aid and the other half were not. 127 convicts are younger than 20 years old, and 205 of them are between 20 to 27, and 100 among 432 convicts are older than 27 years old. 53 convicts are not black and 379 of them are black. 247 out of 432 observations had previous full time work experience and the remaining 185 convicts did not. 379 convicts were not married and 53 convicts were married. 267 out of 432 convicts were released on parole and 165 were not. 251 convicts had less than 3 convictions before current incarceration and the remaining 181 had more than or equal to 3 previous convictions. 24 convicts are on the education level of 6th grade or less, 239 convicts on the level of 7th to 9th grade, 119 of them on the level of 10th to 11th grade; 39 of them finished high school and only 11 out of 432 convicts went to college.

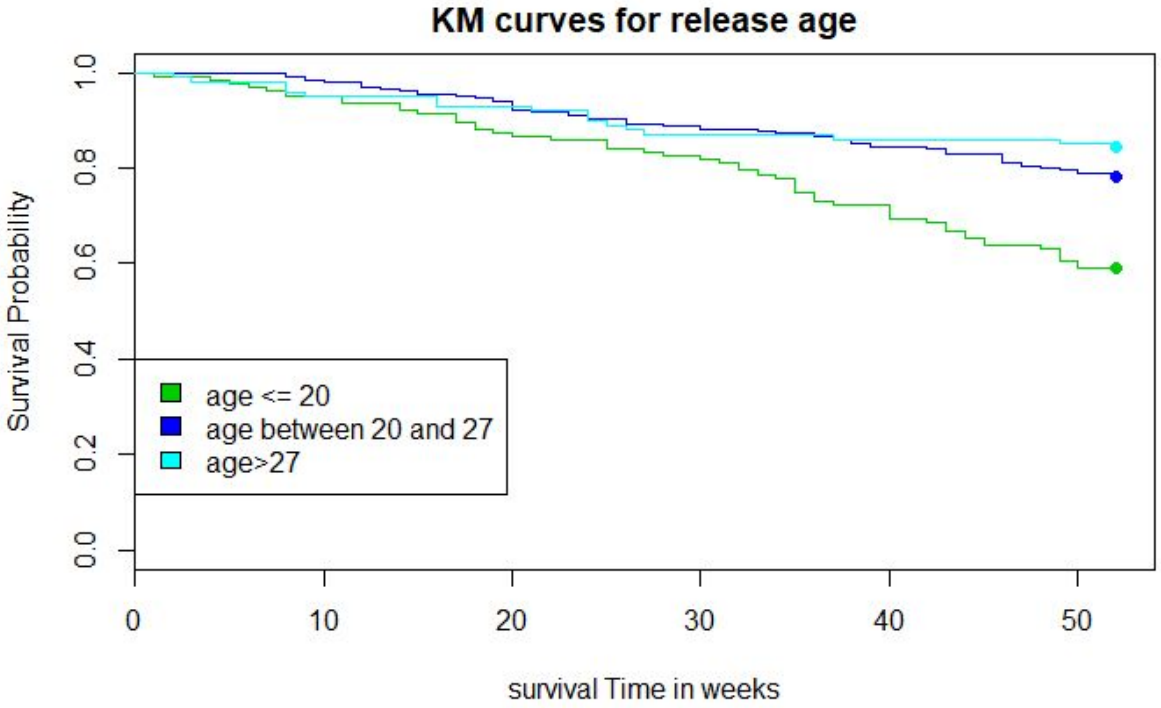
```
      week        arrest  fin     age     race    wexp    mar     paro    prio    educ
 Min.   : 1.00   0:318   0:216   0:127   0: 53   0:185   0:379   0:165   0:251   2: 24
 1st Qu.:50.00   1:114   1:216   1:205   1:379   1:247   1: 53   1:267   1:181   3:239
 Median :52.00                   2:100                                           4:119
 Mean   :45.85                                                                   5: 39
 3rd Qu.:52.00                                                                   6: 11
 Max.   :52.00
```

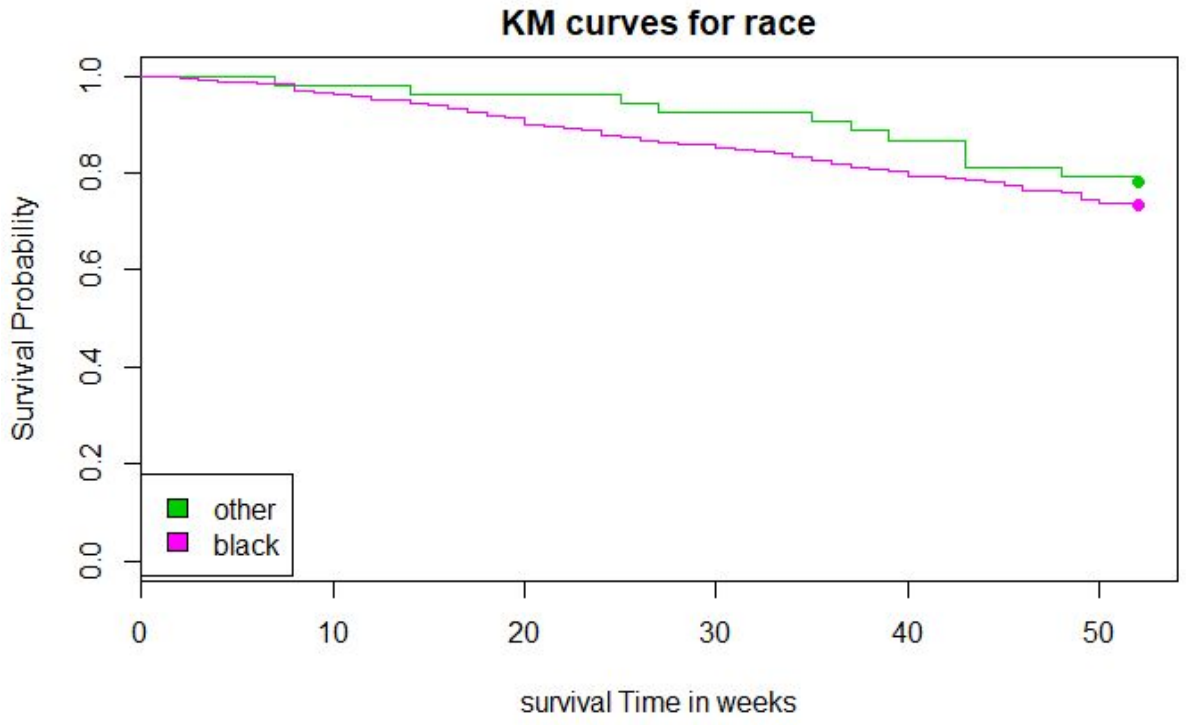
Kaplan-Meier estimation curves

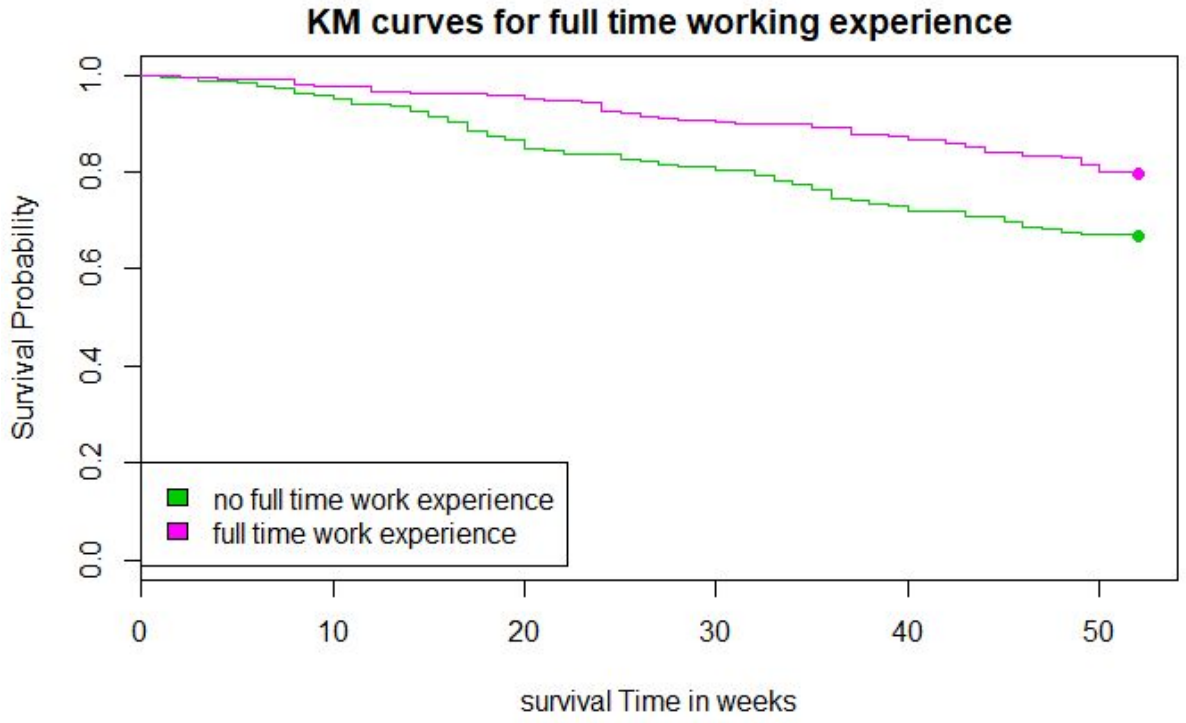
We then try to find out the effects of each covariate on survival time (in weeks) after release by plotting the Kaplan-Meier estimation curves. Convicts who received financial aid are more likely to survive longer than those who did not. Convicts who are older than 27 are less likely to go back to prison within one year. Non-black convicts tend to have higher survival rates. If convicts had previous full time work experience, they tend to have higher survival probability. Convicts who were not married are less likely to survive longer than those who were. Convicts

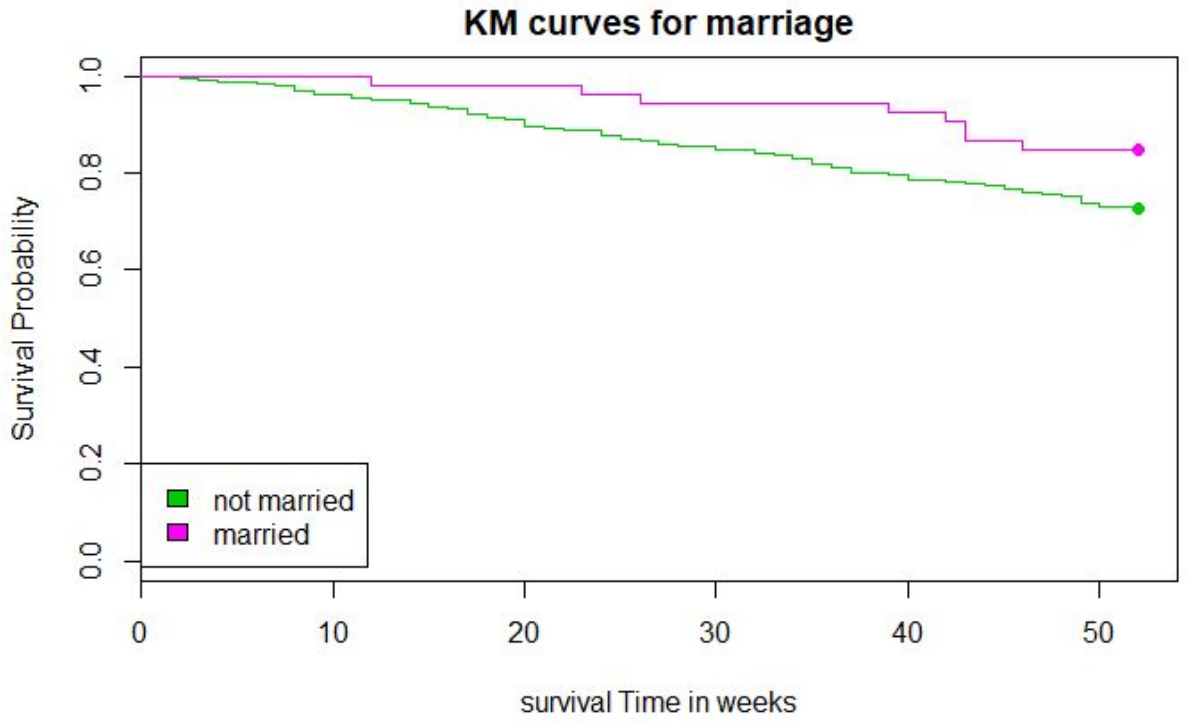
who had less than 3 convictions are more likely to have higher survival rate. Convicts who earned some college degrees are less likely to be arrested again. However, there is barely any difference in survival rate between convicts who were released on parole and who were not.

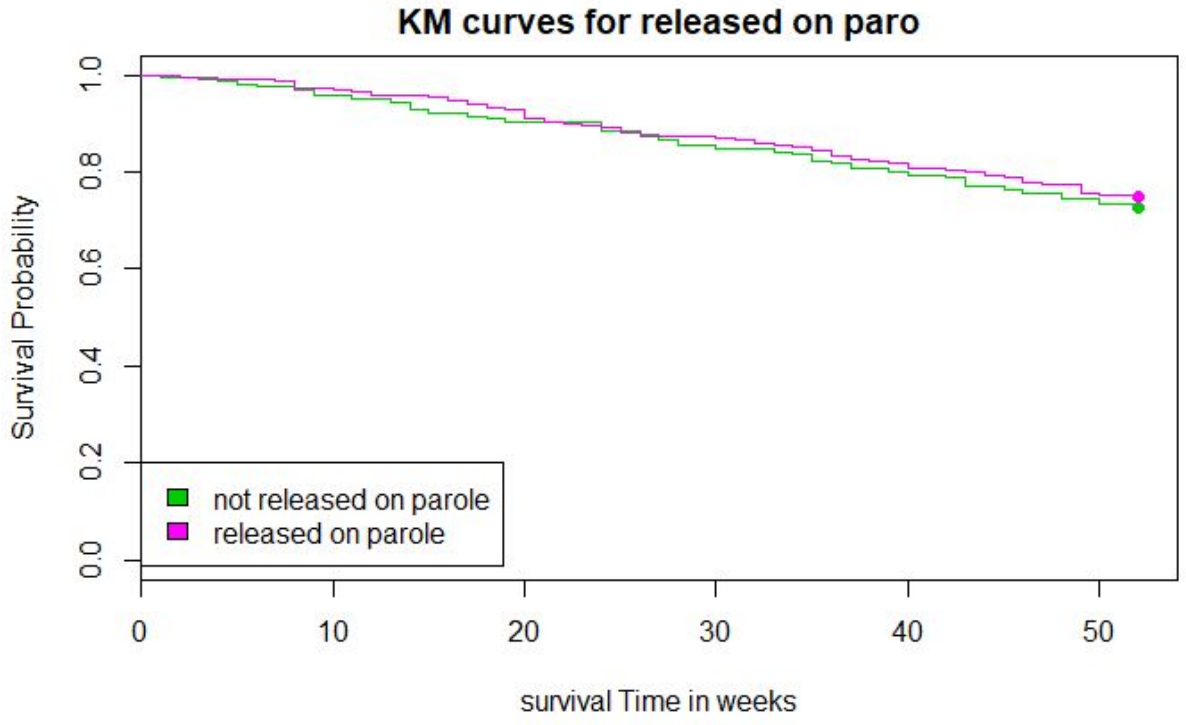


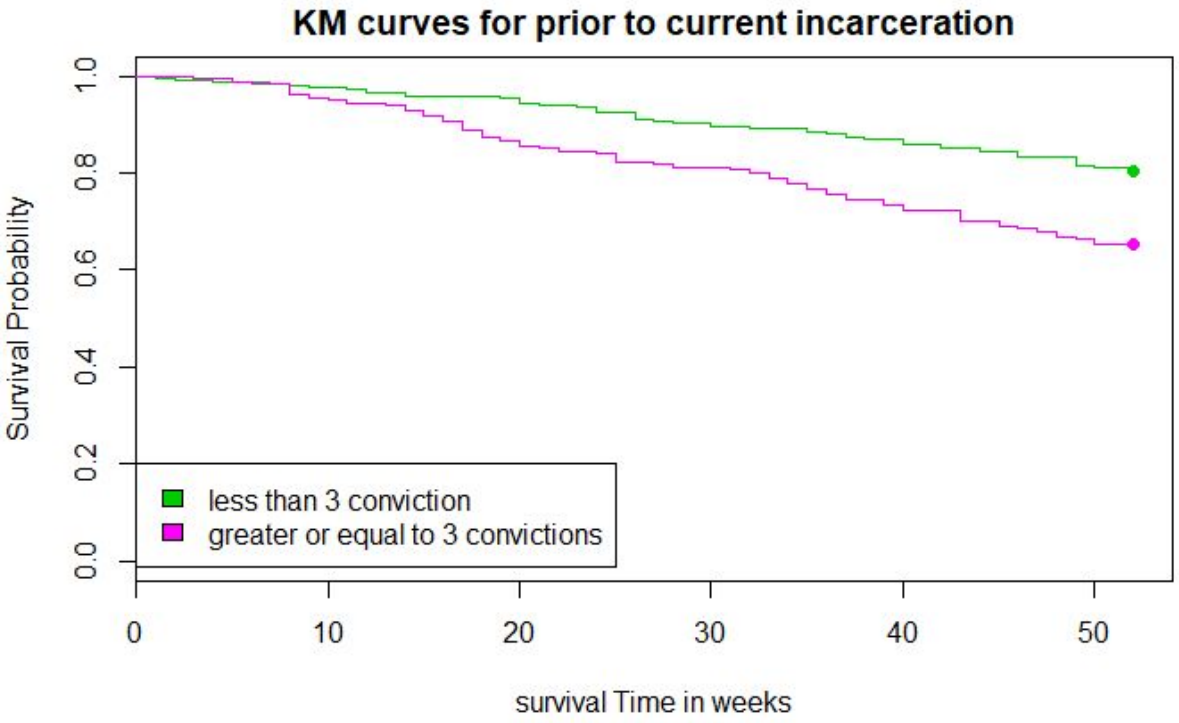


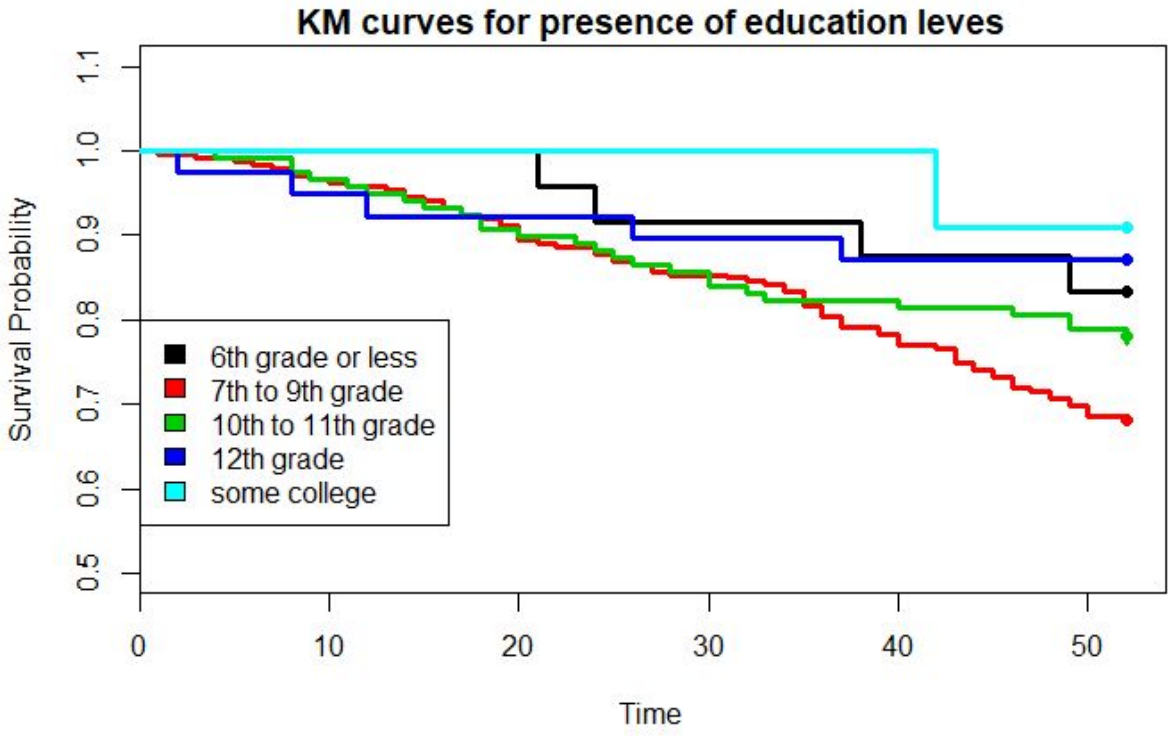

KM curves for race
Survival Probability
0.0
0.2
0.4
0.6
0.8
1.0
other
black
0
10
20
30
40
50
survival Time in weeks


KM curves for full time working experience
Survival Probability
0.0
0.2
0.4
0.6
0.8
1.0
no full time work experience
full time work experience
0
10
20
30
40
50
survival Time in weeks


KM curves for marriage
Survival Probability
0.0
0.2
0.4
0.6
0.8
1.0
not married
married
0
10
20
30
40
50
survival Time in weeks


KM curves for released on paro
Survival Probability
0.0
0.2
0.4
0.6
0.8
1.0
not released on parole
released on parole
0
10
20
30
40
50
survival Time in weeks


KM curves for prior to current incarceration
Survival Probability
0.0
0.2
0.4
0.6
0.8
1.0
less than 3 conviction
greater or equal to 3 convictions
0
10
20
30
40
50
survival Time in weeks


KM curves for presence of education leves
Survival Probability
0.5
0.6
0.7
0.8
0.9
1.0
1.1
6th grade or less
7th to 9th grade
10th to 11th grade
12th grade
some college
0
10
20
30
40
50
Time

Log rank test

We next conduct log-rank test on each covariate. We observe that the P-value for most covariates is smaller than or close to 0.05 except “race” and “paro”, which means all variables have significant effects on the event except “race” and “paro”. So we choose to keep “fin”, “age”, “wexp”, “mar”, “prio”, “educ” in our rough model.

```
Call:
survdiff(formula = rossi.surv ~ rossi$age)

               N Observed Expected (O-E)^2/E (O-E)^2/V
rossi$age=0 127       52     30.8     14.56     20.09
rossi$age=1 205       46     55.8      1.72      3.39
rossi$age=2 100       16     27.4      4.73      6.26

 Chisq= 21.2  on 2 degrees of freedom, p= 3e-05


Call:
survdiff(formula = rossi.surv ~ rossi$mar)

               N Observed Expected (O-E)^2/E (O-E)^2/V
rossi$mar=0 379      106     98.8     0.521      3.94
rossi$mar=1  53        8     15.2     3.394      3.94

 Chisq= 3.9  on 1 degrees of freedom, p= 0.05


Call:
survdiff(formula = rossi.surv ~ rossi$wexp)

               N Observed Expected (O-E)^2/E (O-E)^2/V
rossi$wexp=0 185       62     45.6      5.91      9.91
rossi$wexp=1 247       52     68.4      3.94      9.91

 Chisq= 9.9  on 1 degrees of freedom, p= 0.002


Call:
survdiff(formula = rossi.surv ~ rossi$race)

               N Observed Expected (O-E)^2/E (O-E)^2/V
rossi$race=0  53       12     14.7    0.4990     0.576
rossi$race=1 379      102     99.3    0.0739     0.576

 Chisq= 0.6  on 1 degrees of freedom, p= 0.4
```

```
Call:
survdiff(formula = rossi.surv ~ rossi$prio)

               N Observed Expected (O-E)^2/E (O-E)^2/V
rossi$prio=0 251       51     69.3      4.84      12.4
rossi$prio=1 181       63     44.7      7.50      12.4

 Chisq= 12.4  on 1 degrees of freedom, p= 4e-04


 Call:
 survdiff(formula = rossi.surv ~ rossi$paro)

                N Observed Expected (O-E)^2/E (O-E)^2/V
 rossi$paro=0 165       46     43.1     0.202     0.326
 rossi$paro=1 267       68     70.9     0.122     0.326

  Chisq= 0.3  on 1 degrees of freedom, p= 0.6


Call:
survdiff(formula = rossi.surv ~ rossi$fin)

              N Observed Expected (O-E)^2/E (O-E)^2/V
rossi$fin=0 216       66     55.6      1.96      3.84
rossi$fin=1 216       48     58.4      1.86      3.84

 Chisq= 3.8  on 1 degrees of freedom, p= 0.05


 Call:
 survdiff(formula = rossi.surv ~ rossi$educ)

                N Observed Expected (O-E)^2/E (O-E)^2/V
 rossi$educ=2  24        4     6.81     1.157      1.24
 rossi$educ=3 239       77    61.54     3.884      8.49
 rossi$educ=4 119       27    31.59     0.668      0.93
 rossi$educ=5  39        5    10.78     3.102      3.45
 rossi$educ=6  11        1     3.28     1.581      1.64

  Chisq= 10.5  on 4 degrees of freedom, p= 0.03
```

Model Building

Before we build Cox PH model, we decided to use backward elimination method to select the covariates that fit the best. The final output from R indicates that we only need to keep “fin”, “age”, and “prio” in our model.

```
Step:  AIC=1327.28
Surv(week, arrest) ~ fin + age + prio

       Df    AIC
<none>    1327.3
- fin   1 1327.9
- prio  1 1334.3
- age   2 1339.2
Call:
coxph(formula = Surv(week, arrest) ~ fin + age + prio, data = rossi)
```

We then double check our result with anova table based on the rough model we built with “fin”, “age”, “wexp”, “mar”, “prio”, and “educ”. From the anova table, we get the same result as using backward elimination method.Therefore, we decide to include “fin”, “age” and “prio” in our Cox PH model.

```
Analysis of Deviance Table
 Cox model: response is Surv(week, arrest)
Terms added sequentially (first to last)

      loglik   Chisq Df Pr(>|Chi|)
NULL -675.38
fin  -673.46  3.8371  1   0.050131 .
age  -664.16 18.6010  2  9.138e-05 ***
wexp -662.97  2.3828  1   0.122679
mar  -662.42  1.1071  1   0.292710
prio -658.47  7.8873  1   0.004978 **
educ -655.67  5.6021  4   0.230901
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Model Checking

After building our Cox PH model, we want to check if these three variables satisfy the Cox PH assumption. In order to test our assumption, we are going to use Residual tests and draw C-log-log plots.

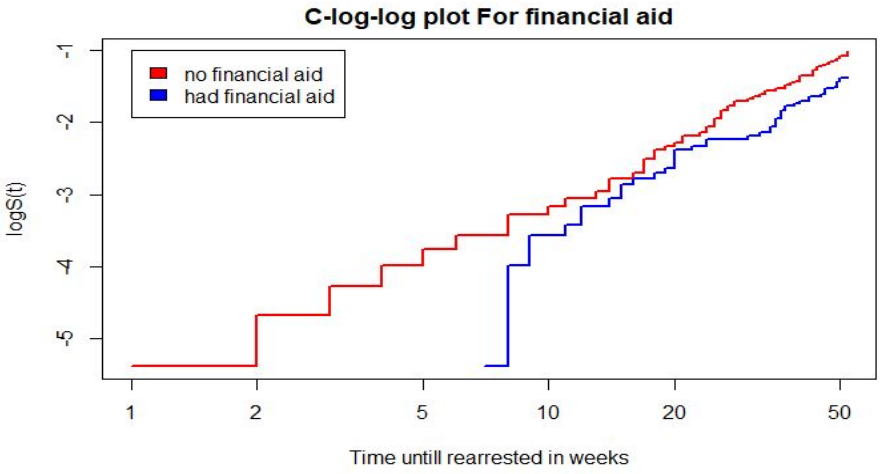
Residual Tests

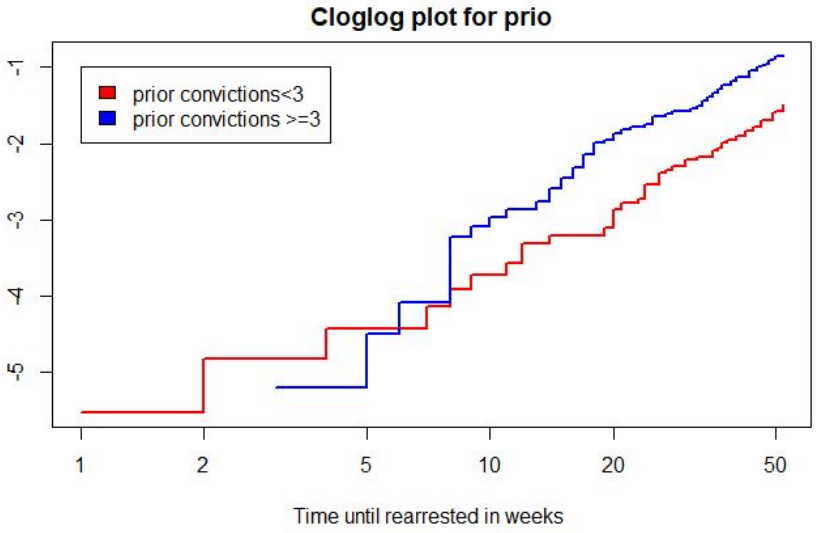
In this case, we use the R command cox.zph( ), for PH assumption, to test the independence between each residual and time. From our report below, it is obvious to see that for “age2”, its p value is 0.025, which is smaller than 0.05. Hence, we conclude that the variable “age” fails the PH assumption and we need to stratify it later.

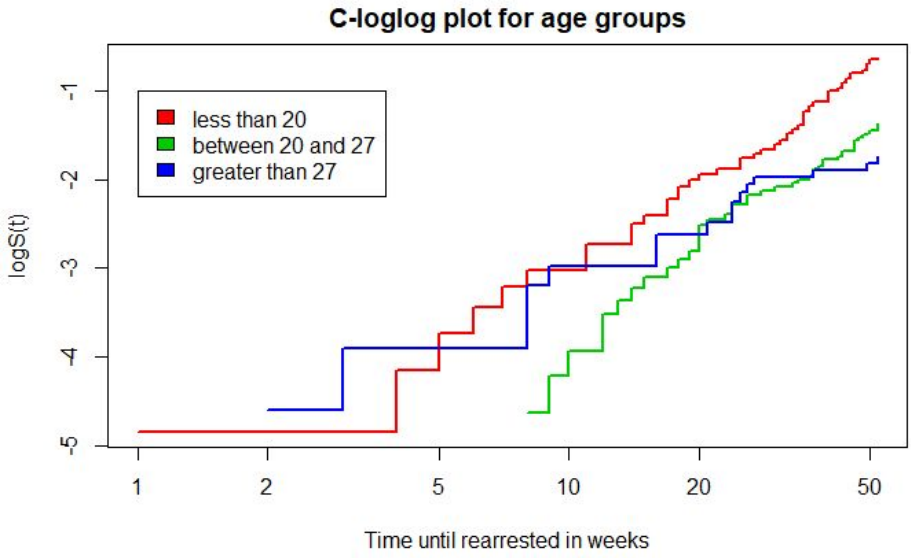
	rho	chisq	p
fin1	-0.0107	0.0133	0.908
age1	-0.0275	0.0876	0.767
age2	-0.2058	5.0237	0.025
prio1	-0.0973	1.0812	0.298
GLOBAL	NA	6.2699	0.180

C-log-log plots

Then we plot C-log-log graphs for each variable to reassure Cox PH assumptions. From the plots below, it is clear that the curves cross each other in “age” plot and therefore we stratify “age” variable since it fails CoxPH assumption check. Although the graph for “fin” diverges at first and “prio” has some minor intersections before 10 weeks, which slightly violates the assumptions, we decide to follow the residual test results and not to stratify them because we are testing the overall effects of all covariates on survival time and residual test is much stronger. In a word, we only stratify the “age” variable.


C-log-log plot For financial aid
no financial aid
had financial aid
logS(t)
-1
-2
-3
-4
-5
1
2
5
10
20
50
Time untill rearrested in weeks


Cloglog plot for prio
prior convictions<3
prior convictions >=3
-1
-2
-3
-4
-5
1
2
5
10
20
50
Time until rearrested in weeks


C-loglog plot for age groups
less than 20
between 20 and 27
greater than 27
logS(t)
-1
-2
-3
-4
-5
1
2
5
10
20
50
Time until rearrested in weeks

Interaction terms

In considering interaction terms of our model, we have three possible interactions: “fin*prio”, “strata(age)*fin” and “strata(age)*prio”. We observed that the p value of each term is bigger than 0.05. Therefore, we conclude that none of the interaction term is significant and our final model is:

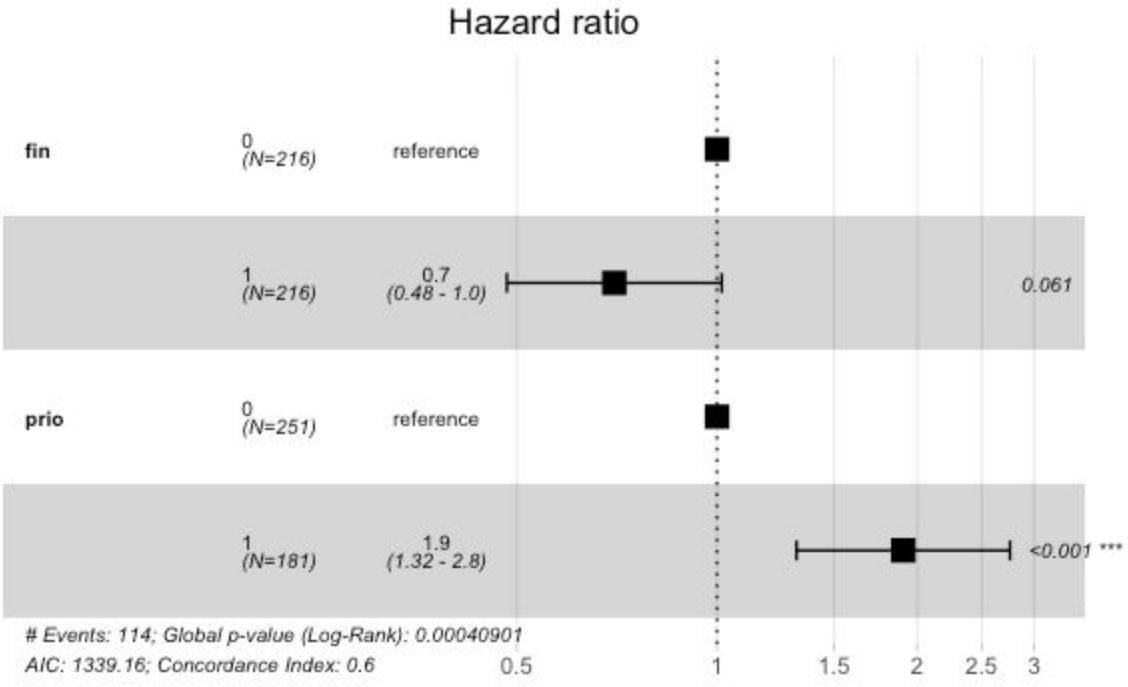
Surv~fin+prio+strata(age).

```
Analysis of Deviance Table
 Cox model: response is Surv(week, arrest)
Terms added sequentially (first to last)

          loglik   Chisq Df Pr(>|Chi|)
NULL     -675.38
fin      -673.46  3.8371  1  0.0501315 .
prio     -667.58 11.7665  1  0.0006031 ***
fin:prio -667.31  0.5463  1  0.4598366
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1


Analysis of Deviance Table
 Cox model: response is Surv(week, arrest)
Terms added sequentially (first to last)

                loglik  Chisq Df Pr(>|Chi|)
NULL            -551.47
fin             -550.05 2.8343  1    0.09227 .
fin:strata(age) -547.98 4.1314  2    0.12673
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1


Analysis of Deviance Table
 Cox model: response is Surv(week, arrest)
Terms added sequentially (first to last)

                 loglik  Chisq Df Pr(>|Chi|)
NULL             -551.47
prio             -546.79 9.3608  1   0.002217 **
prio:strata(age) -546.68 0.2070  2   0.901663
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Hazard Ratios and Confidence Intervals

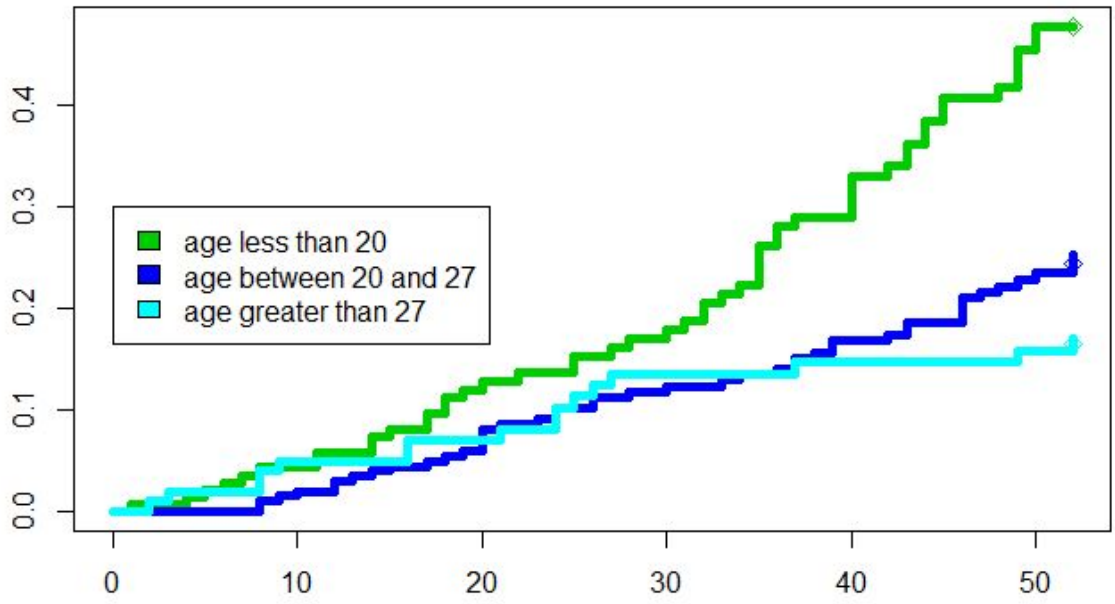
In order to compute Hazard Ratios and Confidence intervals for each covariates of different groups, we use R command: ggforest( ) to visualize the result.

From the chart below, we observe that the hazard ratio of convicts receiving financial aids is centered at point 0.7, and the 95% confidence interval is (0.48, 1.0), which indicates that convicts who received financial aid at the time of release have 30% less likelihood to be arrested again after first release than those who did not receive financial aid. Similarly, the hazard ratio of convicts who had more than or equal to 3 convictions previously is centered at point 1.9 and the related 95% Confidence Interval is (1.32, 2.8), which represents that convicts who had more than or equal to 3 convictions before this incarceration had 90% more likelihood to get back to prison again than convicts who had less than 3 convictions before.



Baseline Hazard Rates

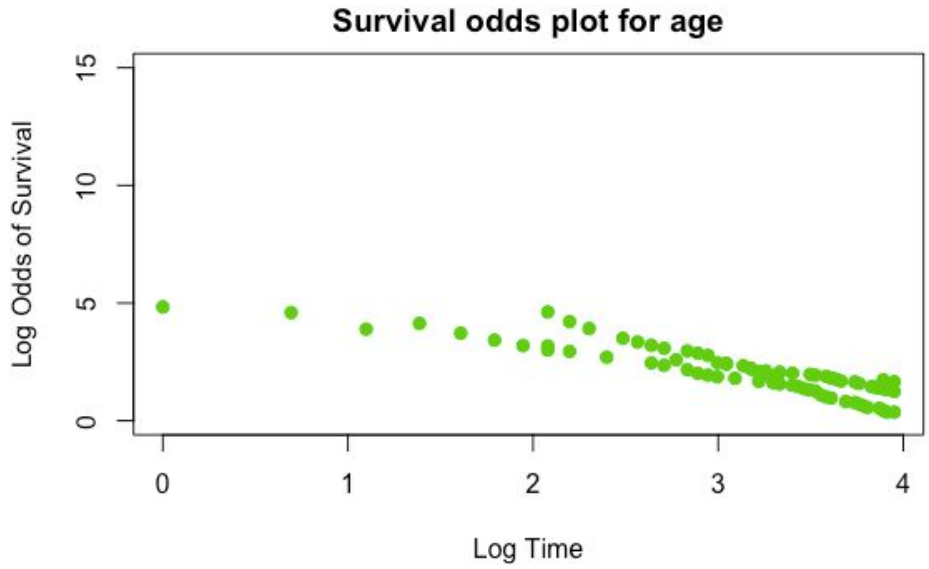
Then we draw the baseline hazard plot for the stratified variable “age”. From the plot below, we can observe how baseline hazard is plotted for each age group, and it is obvious to conclude that the group of convicts who are less than 20 years old has higher baseline hazard rate than other age groups. Therefore, it reveals that convicts who are 20 or younger are more likely to be arrested again after their first release.

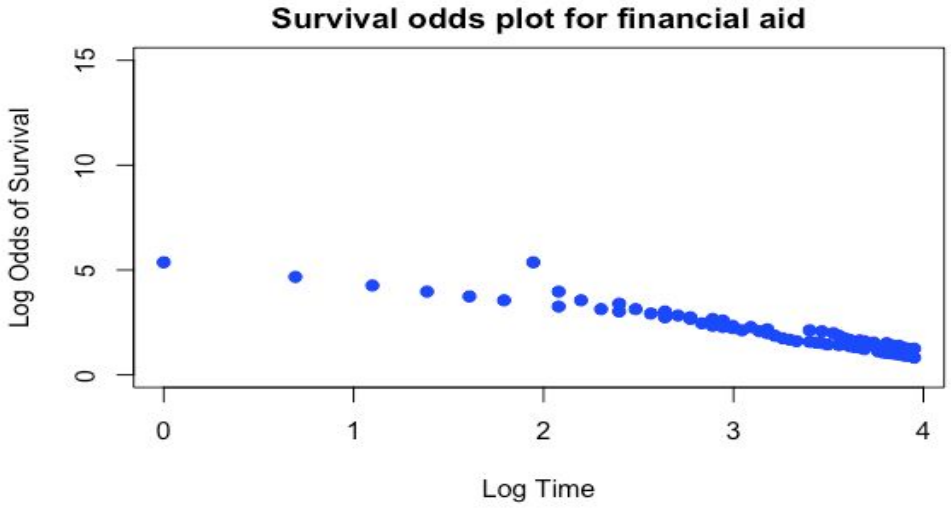


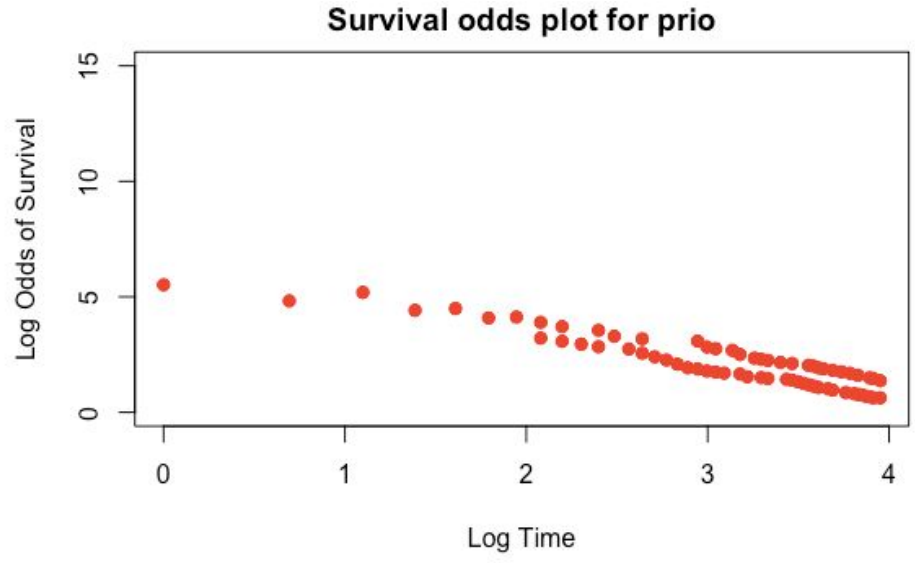
Extension -- AFT model

We now try to extend our analysis of the final model by building AFT model with log-logistic distribution. We use survival odds plots for “fin”, “age”, and “prio” to check AFT assumptions. A central AFT model assumption is that covariates have multiplicative effects on survival time that is consistent over time. According to the AFT assumption, when controlled

effect is consistent across the lifespan, points in this plot should approximate a straight line. The plots below indicate that we could use log-logistic distribution to build our AFT model.







```
Call:
survreg(formula = coxph(Surv(week, arrest) ~ fin + prio + age,
    data = rossi), data = rossi, dist = "loglogistic")
              Value Std. Error     z       p
(Intercept)  4.3614     0.1627 26.81 < 2e-16
fin1         0.2427     0.1462  1.66  0.0970
prio1       -0.4437     0.1480 -3.00  0.0027
age1         0.4747     0.1616  2.94  0.0033
age2         0.7109     0.2189  3.25  0.0012
Log(scale)  -0.4233     0.0868 -4.88 1.1e-06

Scale= 0.655

Log logistic distribution
Loglik(model)= -681.4   Loglik(intercept only)= -696.7
        Chisq= 30.53 on 4 degrees of freedom, p= 3.8e-06
Number of Newton-Raphson Iterations: 4
n= 432
```

From the plot above, we observe that the estimated acceleration factor γ for receiving financial aid and not receiving financial aid is about 1.275, which means weeks of first arrest after release for convicts who received financial aid accelerated 1.275 times than those who did not receive financial aid. Thus, the probability for convicts who did not receive financial aid surviving x weeks of first arrest after release is the same as convicts who received financial aid surviving $1.275x$ weeks.

Similarly, the estimated acceleration factor γ for comparing convicts who had 3 or more convictions before and had less than 3 convictions previously, is about 0.642, which shows that weeks of first arrest after release for convicts who had 3 or more convictions before accelerated 0.642 times than those who had less than 3 convictions. Hence, the probability for convicts who had less than 3 previous convictions surviving x weeks of first arrest after release is the same as convicts had 3 or more convictions before surviving $0.642x$ weeks.

We could also observe that compared to convicts who are less than or equal to 20 years old, the estimated acceleration factor γ for convicts between 20 and 27 is about 1.608, and for

convicts older than 27 is approximately 2.036. Therefore, the probability for convicts who are less than or equal to 20 surviving x weeks of first arrest after release is the same as convicts between 20 to 27 surviving $1.608x$ weeks and convicts who are older than 27 surviving $2.036x$ weeks out of prison before arrest.

Hence, by applying AFT model with log-logistic distribution, we conclude that the covariate “age” has the highest effect on survival time.

Conclusion

Based on the right-censored data of 432 observations and related variables, we first apply Kaplan-Meier estimation curves to check whether each covariate has an effect on the event, and we use log-rank test to eliminate variables should not be considered at the first place. We then use backward elimination method to build our model and double check it with anova table. The original guess that “convicts were less likely to be rearrested if they received financial aid after release / had previous full-time work experience / released on parole / had earned higher degrees” was partially correct and our final model says that only “fin”,”age”,”prio” will influence survival rates. After checking our model with residual test and C-log-log plots, we decided to stratify the variable “age” and keep the others unchanged. We then tested for any possible interaction terms and concluded that none of these should be considered in our model. Our final model will be $Surv\sim fin+prio+strata(age).$ Applying ggforest( ) function gives us the estimation of the coefficients and 95% confidence intervals for the non-stratified models and we draw the baseline hazard plot for the stratified variable age. In the extension part, we build the AFT model and find out that “age” has the highest effect on the event. In conclusion, we set up our model with the variables “fin”, “prio”, and “strata(age)”, and determine that convicts who are under 20 years old, having more than or equal to 3 previous convictions, without receiving any financial aid after release, are more likely to be arrested again.