Final project

Abstract

Sociologist are always interested in crime rate, and they can observe some people are more likely to be recidivist. Education, economy issue, broken marriage can make some people into recidivist. In this project, we will build a Cox Proportional Hazards model to investigate the relationship between rearrest time of some covariates for a group of recidivist, using the dataset of 432 male prisoners.

Data source

The Rossi data set in the *carData* package contains data from an experimental study of recidivism of 432 male prisoners, who were observed for a year after being released from prison (Rossi et al.,1980). Here are the descriptions of the variables in the data:

- *week*: week of first arrest after release, or censoring time。
- *arrest*: the event indicator, equal to 1 for those arrested during the period of the study and 0 for those who were not arrested
- *fin*: a factor, with levels “yes” if the individual received financial aid after release from prison, and “no” if he did not; financial aid was a randomly assigned factor manipulated by the researchers.

- *age*: in years at the time of release.
- *race*: a factor with levels “black” and “other”
- *wexp*: a factor with levels “yes” if the individual had full-time work experience prior to incarceration and “no” if he did not.
- *mar*: a factor with levels “married” if the individual was married at the time of release and “not married” if he was not.
- *paro*: a factor coded “yes” if the individual was released on parole and “no” if he was not.
- *prio*: number of prior convictions.
- *educ*: education, a categorical variable coded numerically, with codes 2 (grade 6 or less), 3 (grades 6 through 9), 4 (grades 10 and 11), 5 (grade 12), or 6 (some post-secondary).
- *emp1-emp52*: factors coded “yes” if the individual was employed in the corresponding week of the study and “no” otherwise.

Here are the example of the data:

week	arrest	fin	age	race	wexp	mar	paro	prio	educ
20	1	no	27	black	no	not married	yes	3	3
17	1	no	18	black	no	not married	yes	8	4
25	1	no	19	other	yes	not married	yes	13	3
52	0	yes	23	black	yes	married	yes	1	5
52	0	no	19	other	yes	not married	yes	3	3

For example, the first person was arrested in week 20 of the study, while the fourth individual was never rearrested, thus he has a censoring time 52.

summary of data

```
summary(Rossi[,1:10])
```

```
##       week           arrest        fin           age          race
##  Min.   : 1.00   Min.   :0.0000   no :216   Min.   :17.0   black:379
##  1st Qu.:50.00   1st Qu.:0.0000   yes:216   1st Qu.:20.0   other: 53
##  Median :52.00   Median :0.0000             Median :23.0
##  Mean   :45.85   Mean   :0.2639             Mean   :24.6
##  3rd Qu.:52.00   3rd Qu.:1.0000             3rd Qu.:27.0
##  Max.   :52.00   Max.   :1.0000             Max.   :44.0
##   wexp              mar       paro          prio             educ
##  no :185   married    : 53   no :165   Min.   : 0.000   Min.   :2.00
0
```

```
##  yes:247   not married:379   yes:267   1st Qu.: 1.000   1st Qu.:3.00
0
##                                       Median : 2.000   Median :3.00
0
##                                       Mean   : 2.984   Mean   :3.47
7
##                                       3rd Qu.: 4.000   3rd Qu.:4.00
0
##                                       Max.   :18.000   Max.   :6.00
0
```

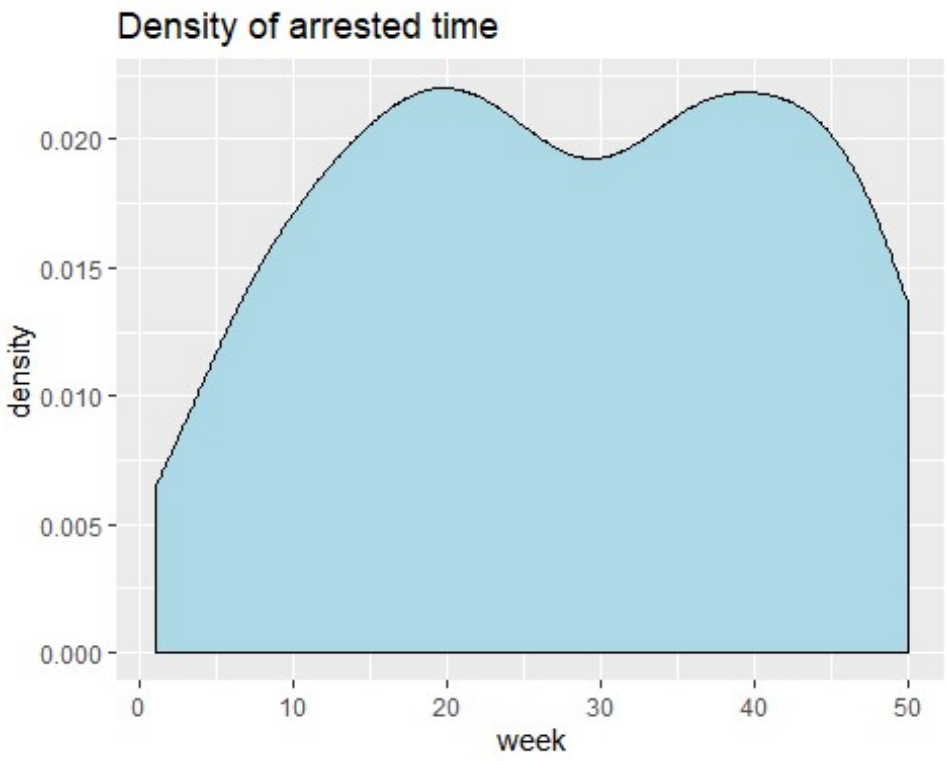
By using data summary function, we find that 87.7 of the people are black race, 57.2 of the people have full time job before incarceration. 50 of the individuals are randomly financial aided by the researcher, and only 12.3 of the people are married.

```
quantile(Rossi$week,probs = c(seq(0, 0.3, 0.03),0.5,1))
```

```
##    0%    3%    6%    9%   12%   15%   18%   21%   24%   27%   30%
50%  100%
##  1.00  9.00 16.00 20.00 26.00 33.65 37.58 43.51 49.00 52.00 52.00 52
.00 52.00
```

From the quantile result of the arrest week, we find that at least 73% of the people haven't been arrest, and about 12 of the individuals was arrested before 26 week. We could plot the density of the arrest time (of whom were arrested) here:

```
df = filter(Rossi, week < 52)

ggplot(df, aes(x = week)) + geom_density(fill = "light blue") + ggtitle
("Density of arrested time")
```



Kaplan-Mayer estimation and log-rank test

Next, we plot the Kaplan-Meier survival curves to visually analyze the effects of each catagorical covariate race, work experience, financial aid, marriage on arrested time. For each of the group, we draw the KM-estimator and it's confidence interval, which is derived from the greenwood formula:

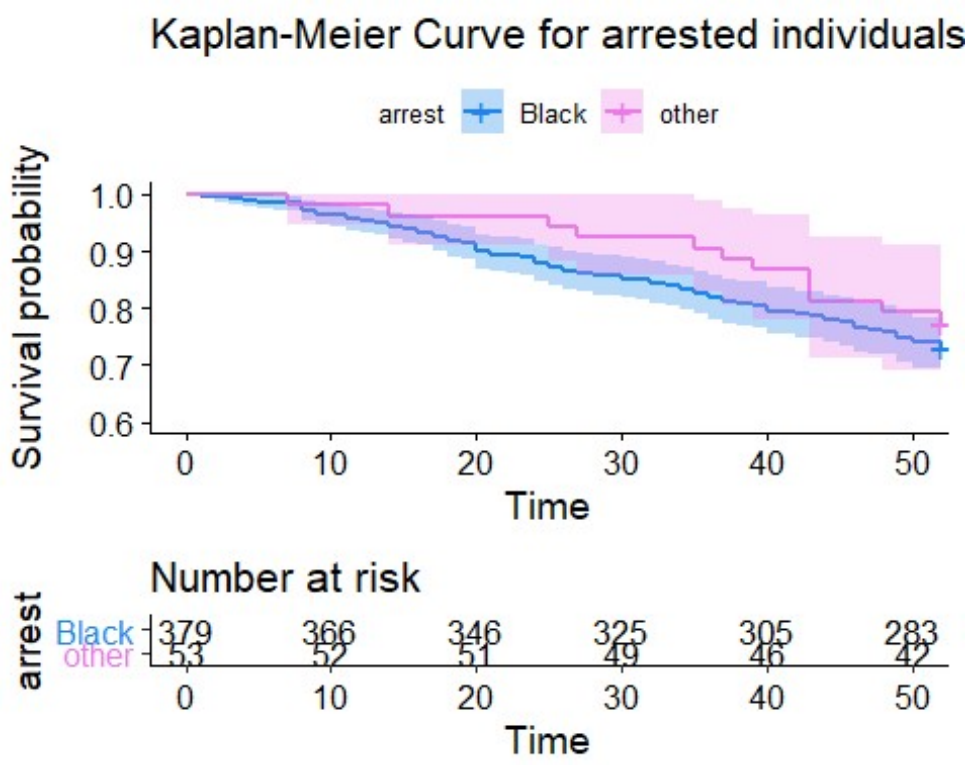
$$var(\hat{S}_t) = \hat{S}_t^2 \sum_{t_i<=t} \frac{d_i}{n_i(n_i - d_i)}$$

With this variance we could derive the confidence interval for the KM estimator. For each of the group, we will perform a log-rank test to get the p-value to test whether the two groups are different in KM-estimator.

Looking at the plots below, we can conclude that all three covariates do affect marriage length. Couples with non African American husband tend to have longer marriage. Couples with same ethnicity are less likely to get divorced than couples with different race. As to the husband educational, an interesting finding is that husband with middle education level are more likely to get divorced than the other groups, which may be caused by different sample sizes of each group.

- Race Draw the KM-estimator and it's confidence interval

```
data1 = filter(Rossi, week <= 52)[,1:10]
ggsurvplot(survfit(Surv(data1$week, data1$arrest)~race, data=data1),con
f.int=TRUE,risk.table=TRUE, ylim = c(0.6,1),
           legend.labs=c("Black", "other"), legend.title="arrest",
           palette=c("dodgerblue2", "orchid2"),
           title="Kaplan-Meier Curve for arrested individuals",
           risk.table.height=.30)
```

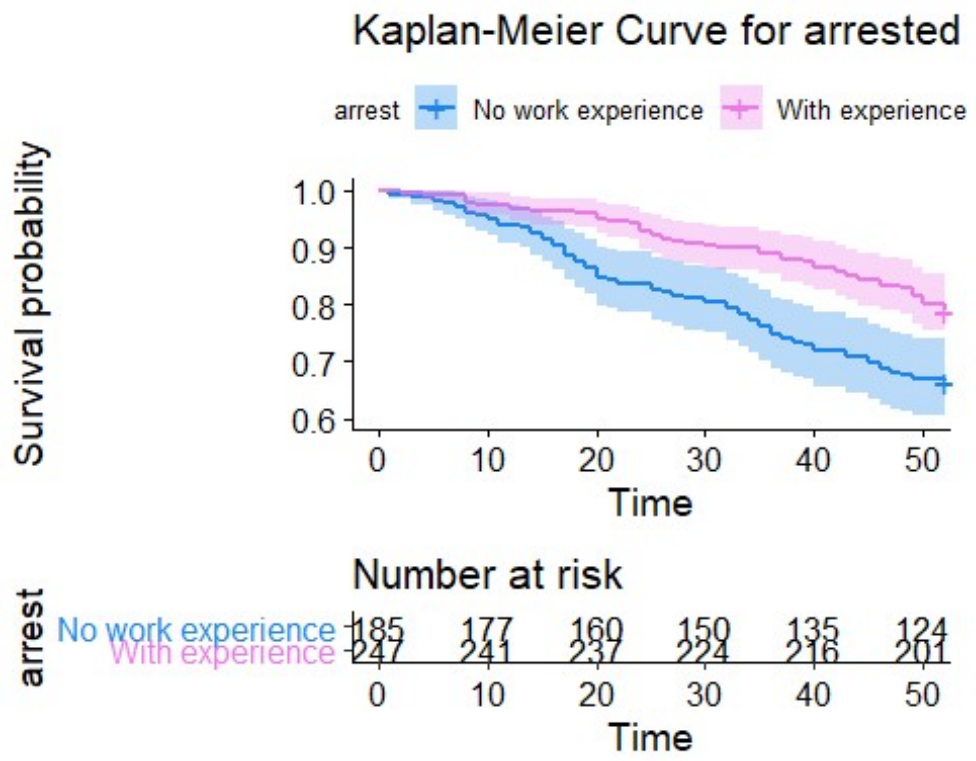


From the Kaplan-mayer estimator we could see that there is a small difference between black and other race. Then we perform the log-rank test:

```
survdiff(Surv(data1$week, data1$arrest)~race, data=data1)

## Call:
## survdiff(formula = Surv(data1$week, data1$arrest) ~ race, data = dat
a1)
##
##                N Observed Expected (O-E)^2/E (O-E)^2/V
## race=black 379      102     99.3    0.0739     0.576
## race=other  53       12     14.7    0.4990     0.576
##
##  Chisq= 0.6  on 1 degrees of freedom, p= 0.4
```

The result have a p-value of 0.4, we could not reject the null hypothesis that there is no difference between two groups.

- Working experience Draw the KM-estimator and it's confidence interval



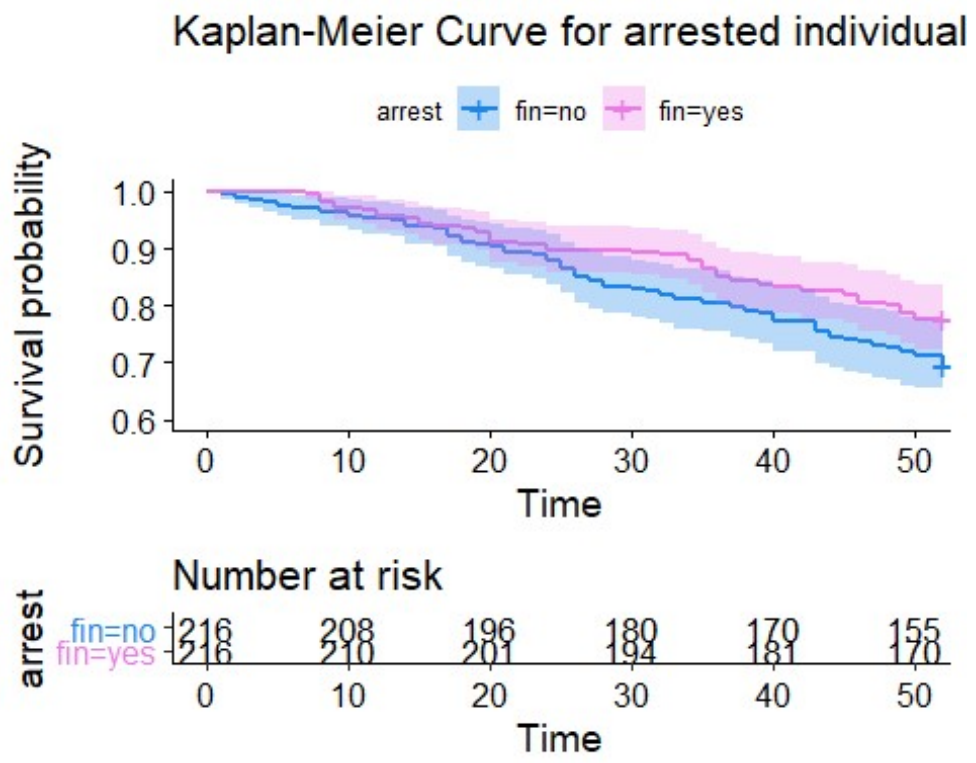
From the Kaplan-mayer estimator we could see that there is a significant difference between with working experiences and without. Then we perform the log-rank test:

```
survdiff(Surv(data1$week, data1$arrest)~wexp, data=data1)

## Call:

## survdiff(formula = Surv(data1$week, data1$arrest) ~ wexp, data = dat

a1)

##

##             N Observed Expected (O-E)^2/E (O-E)^2/V
```

```
## wexp=no  185        62      45.6       5.91       9.91

## wexp=yes 247        52      68.4       3.94       9.91

##

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

You can see the p-value is 0.002, which is enough to reject the null hypothesis.

- Financial aid Draw the KM-estimator and it's confidence interval

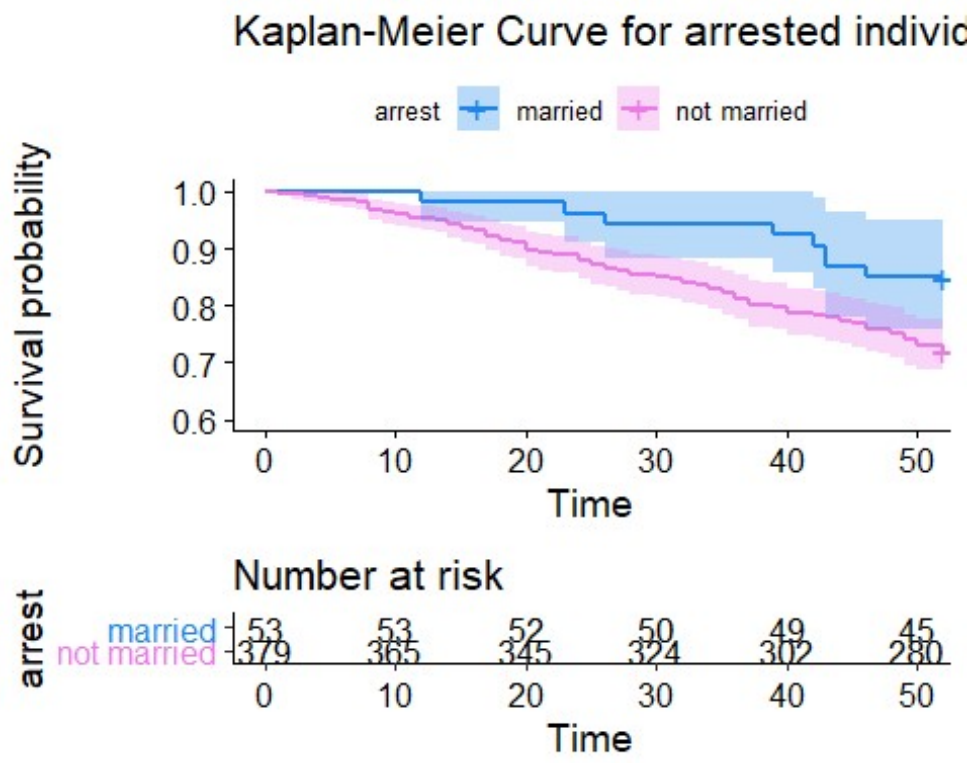



```
survdiff(Surv(data1$week, data1$arrest)~fin, data=data1)

## Call:

## survdiff(formula = Surv(data1$week, data1$arrest) ~ fin, data = data

1)

##

##            N Observed Expected (O-E)^2/E (O-E)^2/V
```

```
## fin=no  216        66      55.6       1.96       3.84

## fin=yes 216        48      58.4       1.86       3.84

##

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

With the p-value 0.05, we could reject the null hypothesis under $\alpha = 0.1$.

- Marriage Draw the KM-estimator and it's confidence interval




```
survdiff(Surv(data1$week, data1$arrest)~mar, data=data1)

## Call:

## survdiff(formula = Surv(data1$week, data1$arrest) ~ mar, data = data

1)

##

##                     N Observed Expected (O-E)^2/E (O-E)^2/V
```

```
## mar=married      53        8     15.2     3.394      3.94
## mar=not married 379      106     98.8     0.521      3.94
##
##  Chisq= 3.9  on 1 degrees of freedom, p= 0.05
```

With the p-value 0.05, we could reject the null hypothesis under $\alpha = 0.1$.

Now with those results we could say those categorical variables have some effect on arresting time more or less. Then we could build our Cox proportional hazard model.

Cox proportional harzard model

To build the Cox proportional hazard model on the time-constant covariates, we need to first determine how many variables we need to include in this model. We will first build a full model, then use backward elimination with AIC criterion to find the model with minimum AIC.

Full model:

```
Coxfull = coxph(formula = Surv(week, arrest) ~ fin + age + race + wexp
+ mar + paro + prio + educ , data = Rossi)
summary(Coxfull)

## Call:
## coxph(formula = Surv(week, arrest) ~ fin + age + race + wexp +
##     mar + paro + prio + educ, data = Rossi)
```

```
##
##   n= 432, number of events= 114
##
##                     coef exp(coef) se(coef)      z Pr(>|z|)
## finyes         -0.35963   0.69794  0.19180 -1.875  0.06079 .
## age            -0.05768   0.94395  0.02187 -2.638  0.00835 **
## raceother      -0.34554   0.70784  0.30907 -1.118  0.26356
## wexpyes        -0.11439   0.89191  0.21311 -0.537  0.59145
## marnot married  0.42496   1.52953  0.38209  1.112  0.26605
## paroyes        -0.08991   0.91401  0.19568 -0.459  0.64589
## prio            0.08469   1.08838  0.02919  2.902  0.00371 **
## educ           -0.18578   0.83046  0.13153 -1.412  0.15782
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
##                exp(coef) exp(-coef) lower .95 upper .95
## finyes            0.6979     1.4328    0.4792    1.0164
## age               0.9440     1.0594    0.9044    0.9853
## raceother         0.7078     1.4128    0.3862    1.2972
## wexpyes           0.8919     1.1212    0.5874    1.3543
## marnot married    1.5295     0.6538    0.7233    3.2344
## paroyes           0.9140     1.0941    0.6229    1.3413
## prio              1.0884     0.9188    1.0279    1.1525
## educ              0.8305     1.2042    0.6417    1.0747
##
```

```
## Concordance= 0.656  (se = 0.026 )
## Likelihood ratio test= 35.35  on 8 df,   p=2e-05
## Wald test            = 33.74  on 8 df,   p=5e-05
## Score (logrank) test = 35.1  on 8 df,   p=3e-05
```

Then we check the tests for proportional-hazards assumption. It could be obtained from *cox.zph()*, which computes a test for each covariate, along with a global test for the model as a whole. The function tests PH assumption based on Schoenfeld residuals, to test for independence between residuals and time:

```
cox.zph(Coxfull)
```

```
##                     rho   chisq        p
## finyes           0.0198  0.0464 0.829381
## age             -0.2682 11.4821 0.000703
## raceother        0.0924  0.9614 0.326824
## wexpyes          0.2501  8.3553 0.003846
## marnot married -0.0773  0.7693 0.380440
## paroyes         -0.0372  0.1633 0.686140
## prio            -0.0503  0.3291 0.566197
## educ            -0.2225  4.6328 0.031367
## GLOBAL               NA 22.1904 0.004575
```

As the variables working experience (wexp) , education (educ) and age (age) have a small p value (which indicate non-proportional hazard), we need to change our oringinal model. As (wexp) is a catogorical variable and do have a significant

influence in the KM estimator in two groups, we could use stratication for this variable.

```
fit1 = coxph(formula = Surv(week, arrest) ~ fin + strata(wexp) + race +
educ + paro + age + mar + prio , data = data1)
```

Then we will use backward AIC criterion to decide the final cox model:

```
AIC(fit1)

## [1] 1173.596

fit2 = coxph(formula = Surv(week, arrest) ~ fin + strata(wexp) + race +
educ + age + mar + prio , data = data1)
AIC(fit2)

## [1] 1171.799
```

Variable paro could be eliminated.

```
AIC(fit2)

## [1] 1171.799

fit3 = coxph(formula = Surv(week, arrest) ~ fin + strata(wexp)  +educ +
 age + mar + prio , data = data1)
AIC(fit3)

## [1] 1171.066
```

Variable race could be eliminated.

```
AIC(fit3)
```

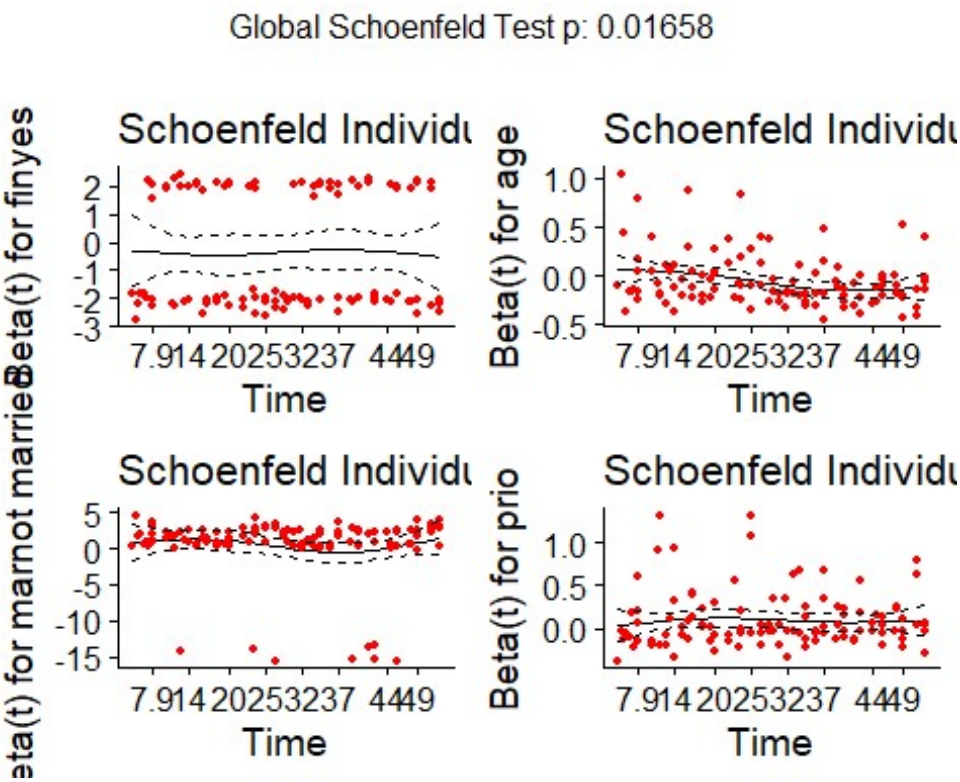
```
## [1] 1171.066

fit4 = coxph(formula = Surv(week, arrest) ~ fin + strata(wexp)  + age +
 mar + prio , data = data1)
AIC(fit4)

## [1] 1170.991
```

Variable education could be eliminated. Check AIC of other model and find that is the smallest AIC model. So our final cox model would be

```
fit4

## Call:
## coxph(formula = Surv(week, arrest) ~ fin + strata(wexp) + age +
##     mar + prio, data = data1)
##
##                     coef exp(coef) se(coef)      z       p
## finyes          -0.36306   0.69554  0.19050 -1.906 0.05667
## age             -0.05632   0.94524  0.02181 -2.582 0.00982
## marnot married   0.49948   1.64786  0.37969  1.316 0.18834
## prio             0.09068   1.09492  0.02843  3.189 0.00143
##
## Likelihood ratio test=22.56  on 4 df, p=0.0001553
## n= 432, number of events= 114
```

scaled Schoenfeld residuals plot

```
ggcoxzph(cox.zph(fit4))
```



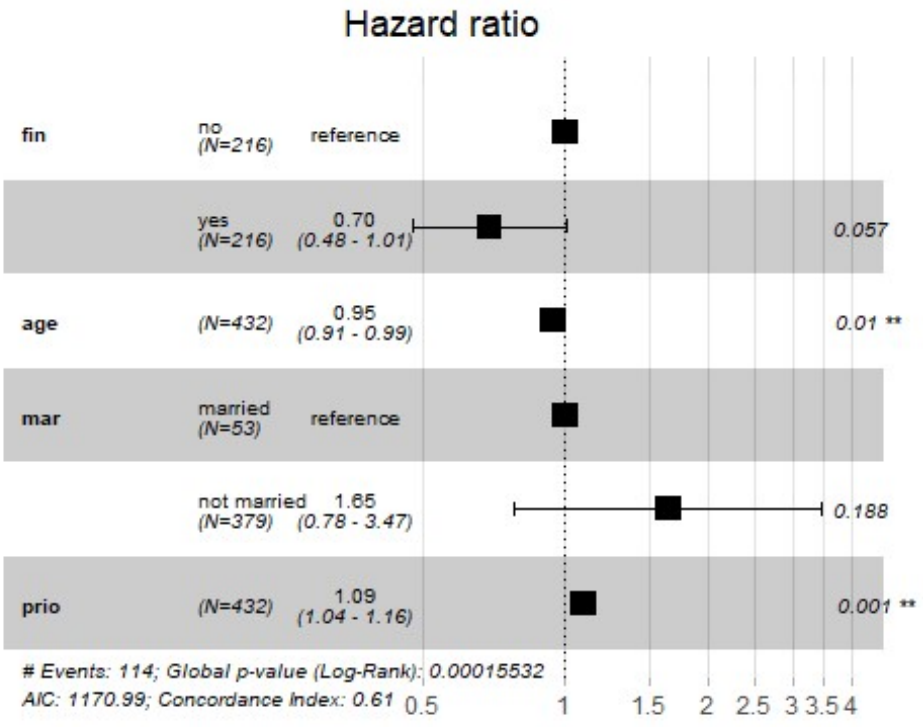
These graphs are analyzed with smoothing spline (solid line in graphs). The broken lines represent 2 times standard-error around the fit. Systematic departures from a horizontal line are indicative of non-proportional hazards, in which age plot is clearly this case. The assumption of proportional hazards appears to be supported for the covariates fin and prio, while it appears to be a trend in the plot for age, and the age effect is declining with time.

Hazard ratios and C.I.

Using *ggforest()* function can visualize the hazard ratio and their confidence interval for each covariate.

```
fit4 = coxph(formula = Surv(week, arrest) ~ fin + strata(wexp)  + age +
 mar + prio , data = data1)
```

```
attr(fit4$terms,"dataClasses") = attr(fit4$terms,"dataClasses")[-3]

ggforest(fit4,data = data1)

## Warning: Removed 2 rows containing missing values (geom_errorbar).
```

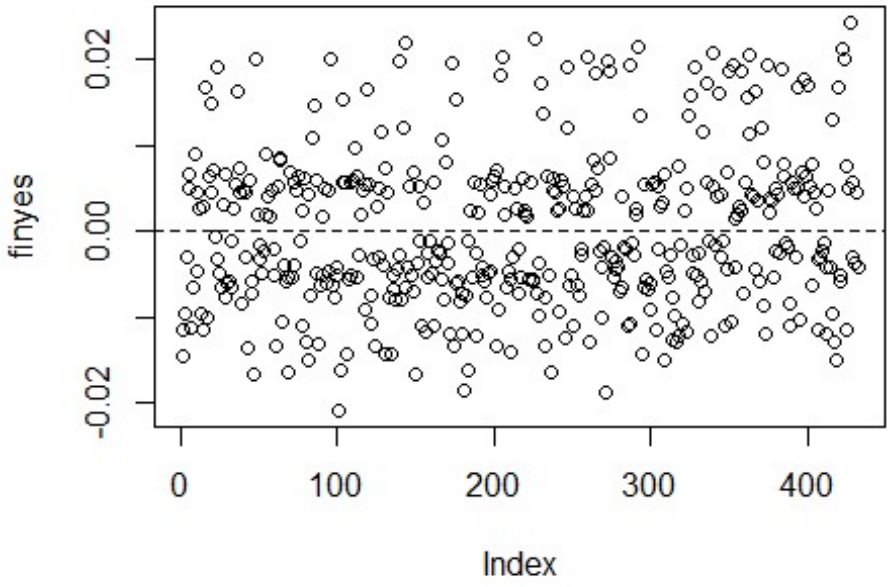


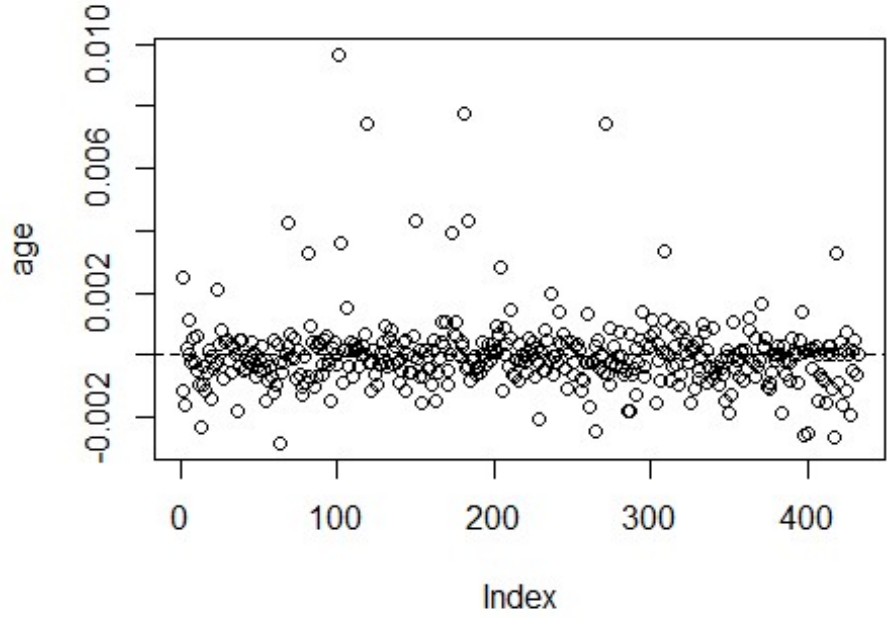
For the information in the hazard ratio chart above, we know that hazard ratio of financial aid yes is centered at 0.70 and its 95% confidence interval is between 0.48 and 1.01. That shows the financial aid could reduce the risk of been arrested. The hazard ratio of not married is 1.65 with 95% confidence interval 0.78 to 3.47 compared to married individuals. That shows the marriage could also reduce the risk of been arrested. The hazard ratio of prior convition time is 1.09 with 95% confidence interval 1.04 to 1.16. That shows the more crime history of one person, the higher the risk to be arrested.

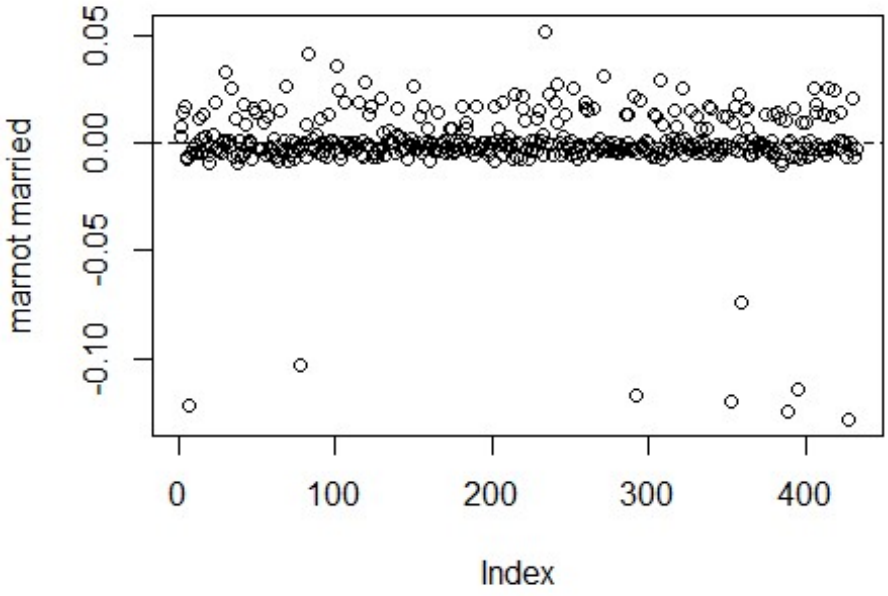
Check influential cases

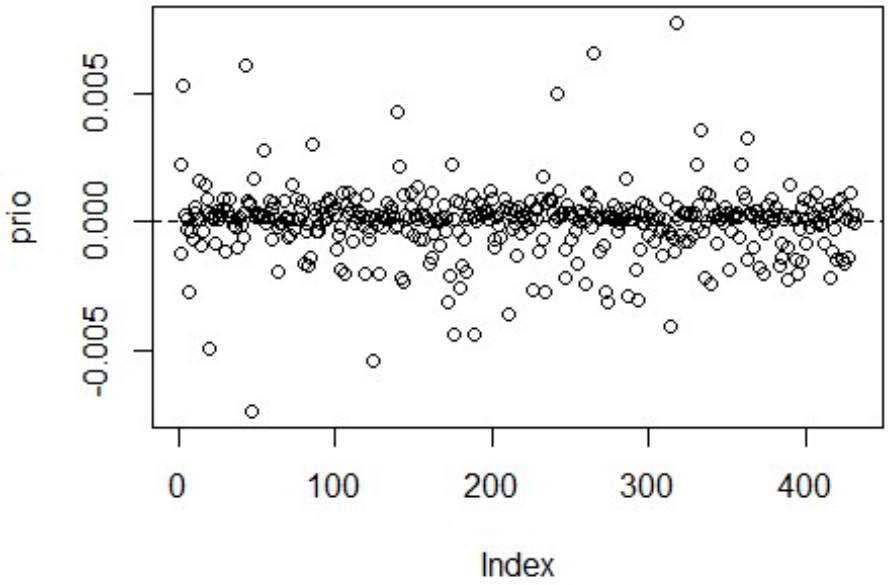
For the stratified model regressing time to rearrest on financial aid, age, marriae and number of prior offenses:

```
dfbeta <- residuals(fit4, type="dfbeta")

for (j in 1:4) {

plot(dfbeta[, j], ylab=names(coef(fit4))[j])

abline(h=0, lty=2)

}
```


finyes
0.02
0.00
-0.02
0
100
200
300
400
Index


age
0.010
0.006
0.002
-0.002
0
100
200
300
400
Index


marnot married
0.05
0.00
-0.05
-0.10
0
100
200
300
400
Index

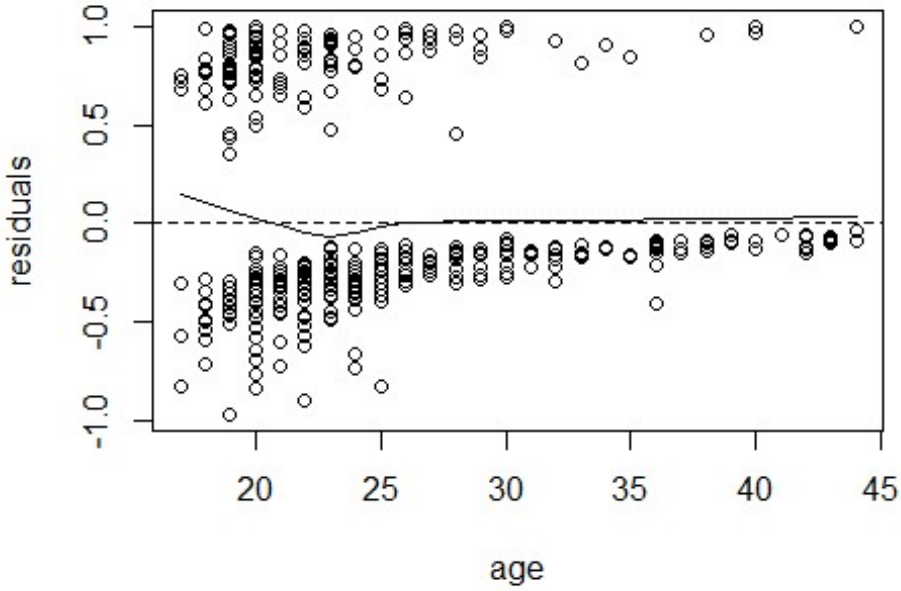

prio
0.005
0.000
-0.005
0
100
200
300
400
Index

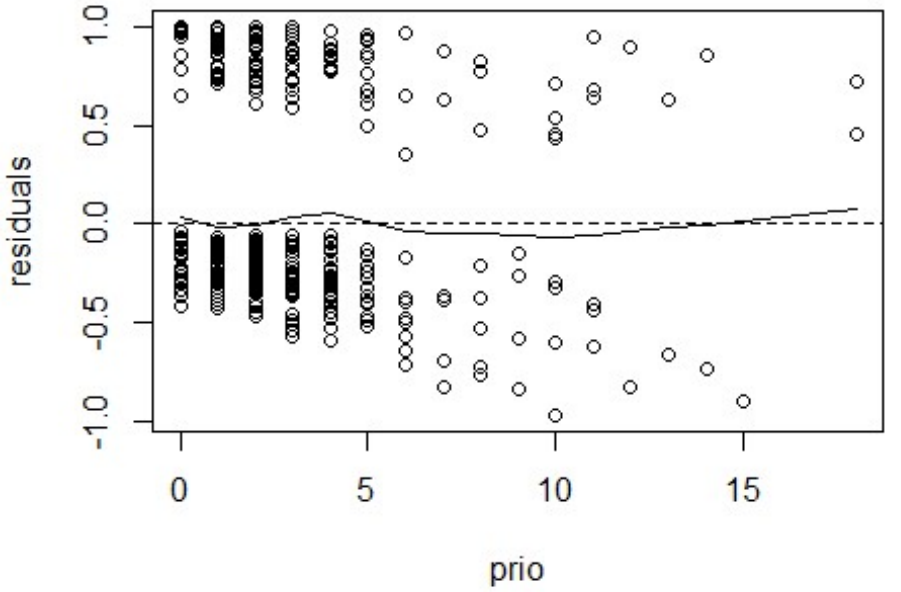
Comparing the magnitudes of the largest dfbeta values to the regression coefficients, we could conclude that none of the cases are dramatically influenced by some individual cases, even though some of the dfbeta values for age and marriage are large compared with the others.

Check the nonlinearity

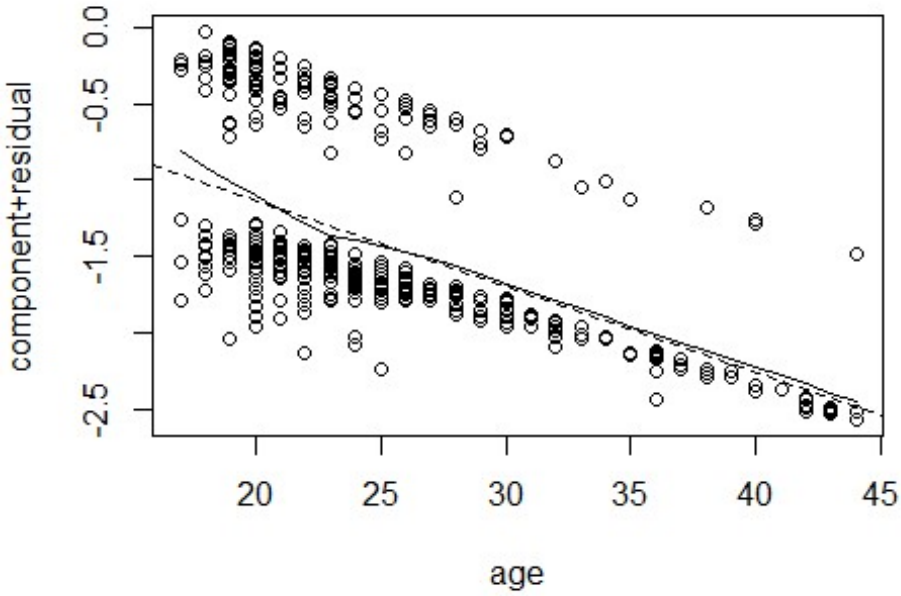
We could use martingale residuls and component-plus-residual and plot it against covariates to detect nonlinearity. Because financial aid and marriage are dichotomous factor, we only need check with the covariates number of prior arrests and age.

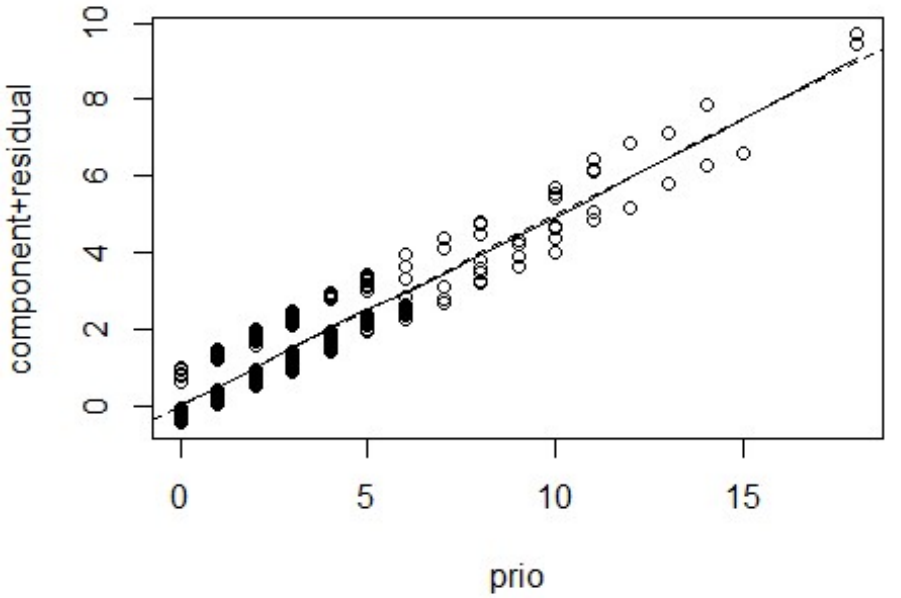
```
res <- residuals(fit4, type="martingale")
X <- as.matrix(Rossi[, c("age", "prio")]) # matrix of covariates
for (j in 1:2) { # residual plots
plot(X[, j], res, xlab=c("age", "prio")[j], ylab="residuals")
abline(h=0, lty=2)
lines(lowess(X[, j], res, iter=0))
}
```


residuals
1.0
0.5
0.0
-0.5
-1.0
20
25
30
35
40
45
age


residuals
1.0
0.5
0.0
-0.5
-1.0
0
5
10
15
prio

```r
b <- coef(fit4)[c(2,3)] # regression coefficients

for (j in 1:2) { # component-plus-residual plots

plot(X[, j], b[j]*X[, j] + res, xlab=c("age", "prio")[j],

ylab="component+residual")

abline(lm(b[j]*X[, j] + res ~ X[, j]), lty=2)

lines(lowess(X[, j], b[j]*X[, j] + res, iter=0))

}
```

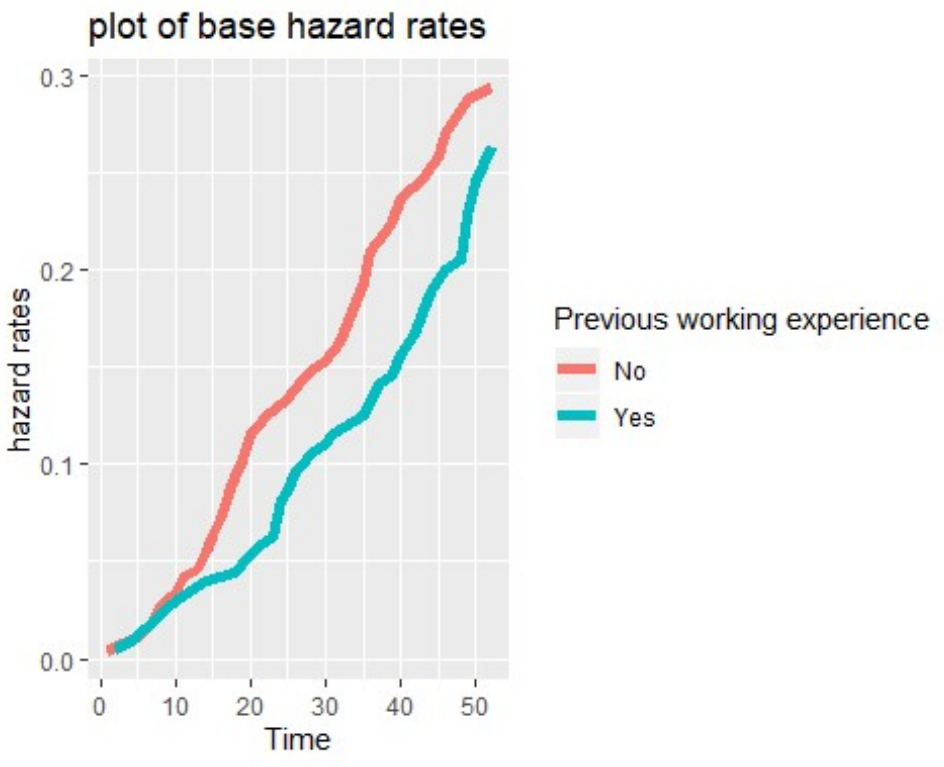

component+residual
0.0
-0.5
-1.5
-2.5
20
25
30
35
40
45
age


component+residual
0
2
4
6
8
10
0
5
10
15
prio

We obtained martingale-residual plots (the first two) and component-plus-residual plots (the latter two) for the covariates age and prio. In martingale-residual plots, the broken line is $y = 0$, and in component-plus-residual plots the broken line is fitted by linear least-squares. And the solid lines in all of the plots are fitted by local linear regression. Check the difference between solid line and the broken line, we can conclude that there is slight nonlinearity in our data.

Baseline Hazard Rates

```
df = basehaz(fit4)

ggplot(df, aes(x = time, y = hazard, color = strata)) + geom_line(size
= 2)+ labs(title = "plot of base hazard rates", x = "Time", y = "hazard
 rates") + scale_colour_discrete(name  ="Previous working experience",
labels=c("No", "Yes"))
```



From the plot below, we clearly see that group without previous working experience has higher baseline hazard rates than another group. As a result, if an individual has no previous working experience, then he will be more likely to be arrested.

Extension: Time-Dependent Covariates

In *survival* package, the coxph() function handles time-dependent covariates by requiring that each time period for an individual appear as a separate "case", that is, as a separate row (or record) in the data set. In the Rossi data, each individual has a single row, with the weekly employment indicators as 52 columns in the data frame, with names emp1 through emp52.

For example, for the first person:

```
Rossi[1,11:62]

##   emp1 emp2 emp3 emp4 emp5 emp6 emp7 emp8 emp9 emp10 emp11 emp12 emp
13 emp14
## 1   no   no   no   no   no   no   no   no   no    no    no    no
no    no
##   emp15 emp16 emp17 emp18 emp19 emp20 emp21 emp22 emp23 emp24 emp25
emp26 emp27
## 1    no    no    no    no    no    no  <NA>  <NA>  <NA>  <NA>  <NA>
 <NA>  <NA>
##   emp28 emp29 emp30 emp31 emp32 emp33 emp34 emp35 emp36 emp37 emp38
emp39 emp40
## 1  <NA>  <NA>  <NA>  <NA>  <NA>  <NA>  <NA>  <NA>  <NA>  <NA>  <NA>
 <NA>  <NA>
##   emp41 emp42 emp43 emp44 emp45 emp46 emp47 emp48 emp49 emp50 emp51
emp52
## 1  <NA>  <NA>  <NA>  <NA>  <NA>  <NA>  <NA>  <NA>  <NA>  <NA>  <NA>
 <NA>
```

The employment indicators are missing after week 20, when individual 1 was rearrested. We use *upfold()* function in package *RcmdrPlugin.survival* to transform the data into another dataset with time-dependent covariates which *coxph()* function can use.

```
data2 = read.csv("Rossi_time.csv")
time_dep <- coxph(Surv(start, stop, arrest.time) ~fin + age + strata(we
xp) + mar + prio + emp,
data=data2)
summary(time_dep)

## Call:
## coxph(formula = Surv(start, stop, arrest.time) ~ fin + age +
##     strata(wexp) + mar + prio + emp, data = data2)
##
##   n= 19809, number of events= 114
##
##                    coef exp(coef) se(coef)      z Pr(>|z|)
## finyes         -0.34288   0.70972  0.19060 -1.799  0.07202 .
## age            -0.04626   0.95479  0.02170 -2.132  0.03299 *
## marnot married  0.35868   1.43144  0.38129  0.941  0.34686
## prio            0.08186   1.08530  0.02886  2.836  0.00457 **
## empyes         -1.32325   0.26627  0.25067 -5.279  1.3e-07 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
##                exp(coef) exp(-coef) lower .95 upper .95
## finyes            0.7097     1.4090    0.4885    1.0312
## age               0.9548     1.0473    0.9150    0.9963
## marnot married    1.4314     0.6986    0.6780    3.0222
## prio              1.0853     0.9214    1.0256    1.1485
```

```
## empyes            0.2663     3.7556    0.1629    0.4352
##
## Concordance= 0.689  (se = 0.024 )
## Likelihood ratio test= 57.64  on 5 df,   p=4e-11
## Wald test            = 48.15  on 5 df,   p=3e-09
## Score (logrank) test = 53.59  on 5 df,   p=3e-10
```

From the result we could see that the time-dependent covariate employ is very significant <0.0001, and has a very strong effect (0.26) in Cox hazard ratio. That indicates thetime-dependent employment covariate has an apparently large effect, but we must know that if a person was arrested , he could never work when he is in jail. So this must partly resulted by that casual relation.

Lagged Covariates time dependent cox model

TO deal with this situation, we can use a lagged value of employment, from the previous week for example. Thus the time-dependent employ status become time-dependent employed status. The *unfold()* function can easily provide lagged time-dependent covariates, and thus we read the data here:

```
data3 = read.csv("healthstudy2.csv")
time_dep_lag <- coxph(Surv(start, stop, arrest.time) ~fin + age + strat
a(wexp) + mar + prio + employed,data=data3)
summary(time_dep_lag)

## Call:
## coxph(formula = Surv(start, stop, arrest.time) ~ fin + age +
```

```
##     strata(wexp) + mar + prio + employed, data = data3)
##
##   n= 19377, number of events= 113
##
##                     coef exp(coef) se(coef)      z Pr(>|z|)
## finyes         -0.33895   0.71252  0.19115 -1.773 0.076187 .
## age            -0.04969   0.95152  0.02179 -2.281 0.022553 *
## marnot married  0.40604   1.50086  0.38126  1.065 0.286876
## prio            0.08958   1.09372  0.02867  3.124 0.001783 **
## employedyes    -0.78339   0.45685  0.21825 -3.589 0.000331 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
##                exp(coef) exp(-coef) lower .95 upper .95
## finyes            0.7125     1.4035    0.4899    1.0363
## age               0.9515     1.0509    0.9118    0.9930
## marnot married    1.5009     0.6663    0.7109    3.1686
## prio              1.0937     0.9143    1.0339    1.1569
## employedyes       0.4569     2.1889    0.2979    0.7007
##
## Concordance= 0.646  (se = 0.028 )
## Likelihood ratio test= 37.03  on 5 df,   p=6e-07
## Wald test            = 34.6  on 5 df,   p=2e-06
## Score (logrank) test = 35.87  on 5 df,   p=1e-06
```

The coefficient for employed indicator still has a small p-value, but the estimated effect of employment, though substantial, is much smaller than before. It changed from 0.26 to 0.46.

Conclusion and Discussion

In this project, we are analyzing the relationship between rearrest time and some covariates such as age, financial aid, marriage and previous offense times. We plotted Kaplan-Meier curve to check if some catagorical covariate affects hazard rate over time. Then we used log-rank test to confirm that some covariates do have significant influence. By fitting the Cox PH model, we could quantitively measure the influence of different covariates, and we use AIC criterion, Proportional Hazards Assumption test to select appropriate model. Then we check the influential cases and nonlinearity using martingale residual plot and partial residual plot. Then with the fianl cox model, we could estimate the coefficients and confidence intervals for the two non-stratified variables previous offence and age, and we et the baseline hazard rates for with working experience case and without case. The result is, marriage, working experience, financial aid all have a negative relationship with hazard rate, while the offence time has a positive relationship.

We then extense our model into time dependent cox PH model. We split the time period by week of each individual, with their employment status as time-dependnet variable. In that model we find that the employment status has a huge influence in hazard rate (0.26), with a very significant p value. But as nobody can work in jail, we

then change that model with the time dependnent variable into previous employed status. With this model, the employed covariates still have a big influence in hazard rate (0.46), but that is much smaller than the first model.