BIS 630 HW 4

Joanna Chen

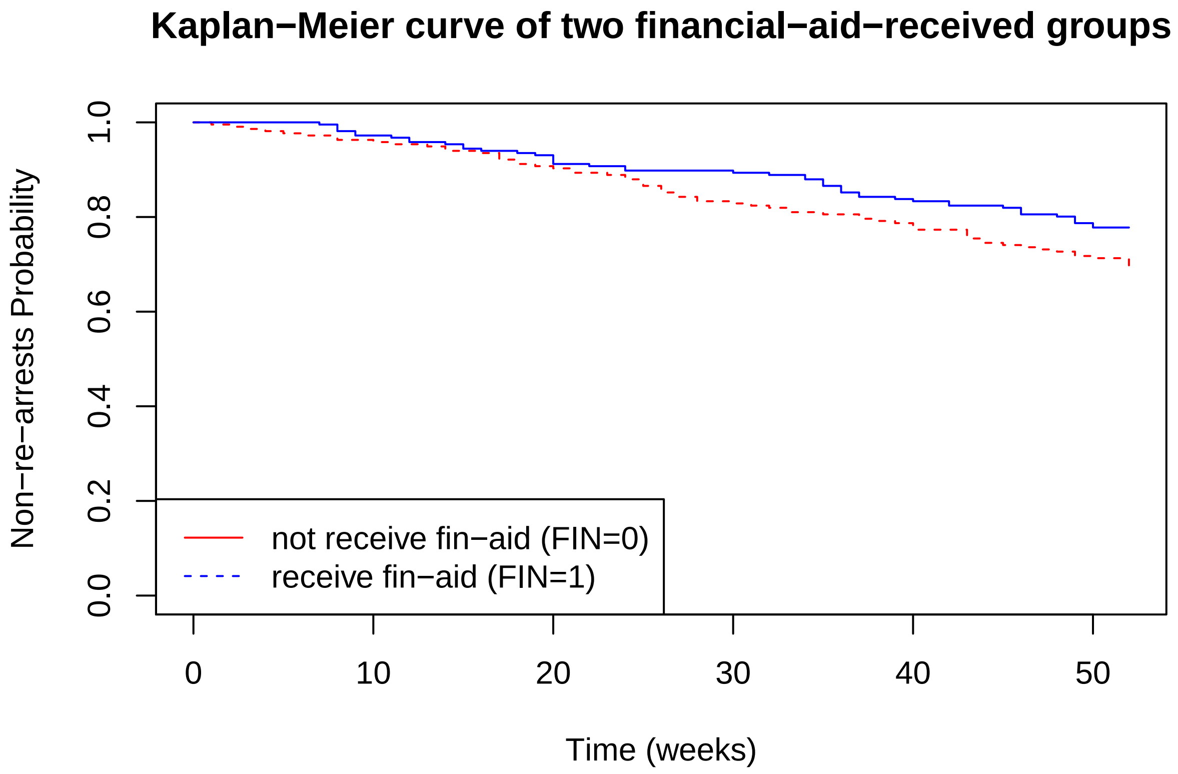
2/19/2020

```
df[,PRIO2:=ifelse(PRIO<=2,0,1)]
```

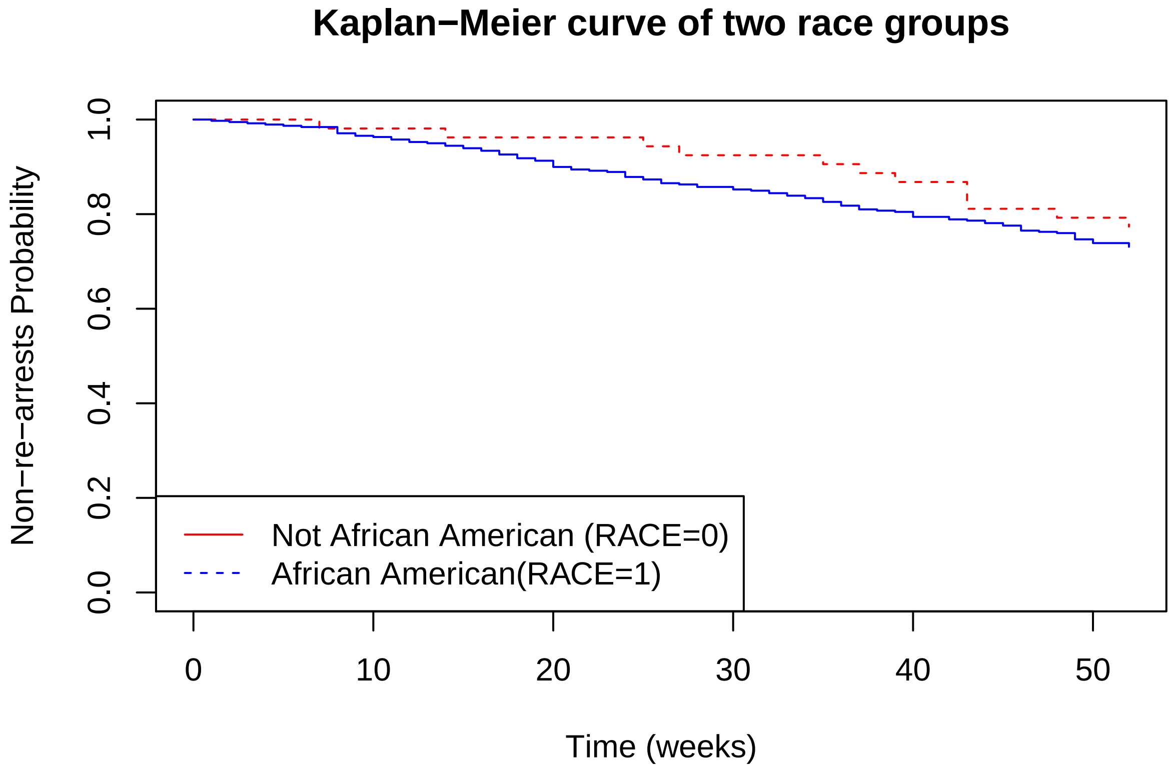
1. Exploratory Analysis:

a. Begin by plotting Kaplan-Meier survival curves for each of the 7 predictor variables under consideration.

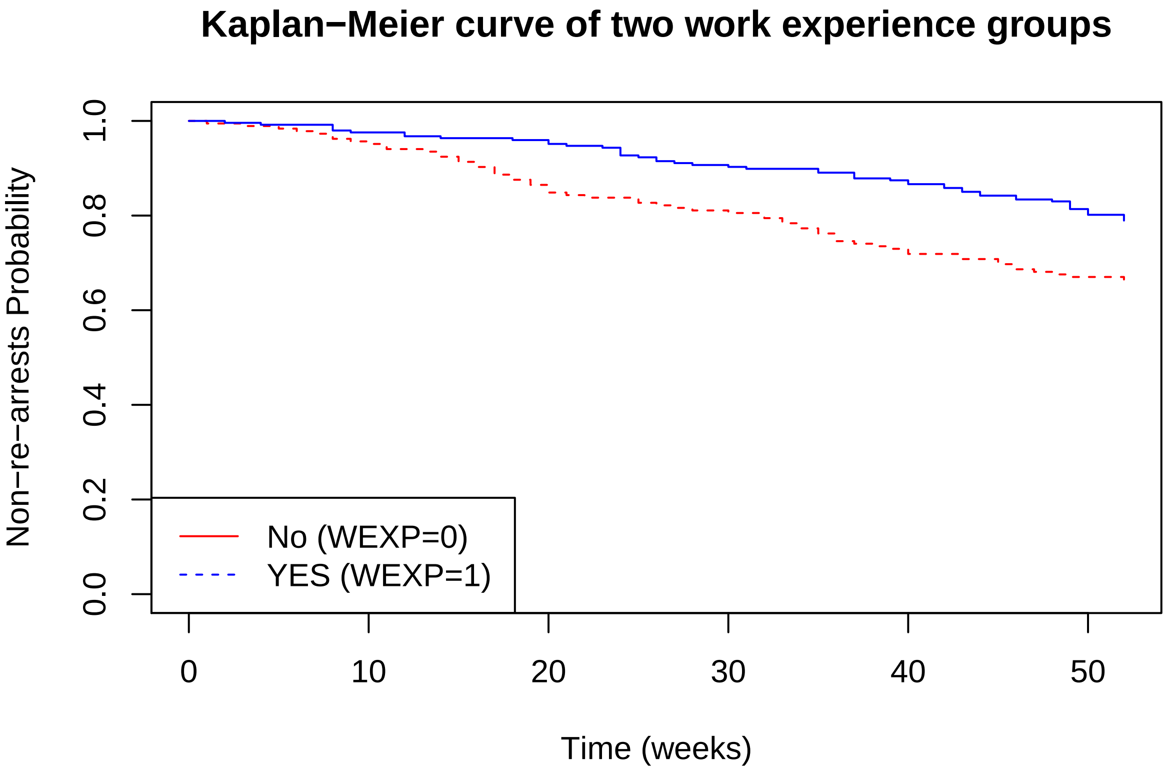
```
df$FIN = factor(df$FIN)
df$RACE = factor(df$RACE)
df$WEXP = factor(df$WEXP)
df$MAR = factor(df$MAR)
df$PARO = factor(df$PARO)
df$AGEGRP = factor(df$AGEGRP)
df$PRIO2 = factor(df$PRIO2)

fit <- survfit(Surv(WEEK,ARREST)~FIN, data = df,conf.type = "none")
plot(fit, xlab="Time (weeks)", ylab="Non-re-arrests Probability",
   conf.int=FALSE, col=c("red", "blue"), lty=c(2,1))
legend("bottomleft", c("not receive fin-aid (FIN=0)", "receive fin-aid (FIN=1) "), col=c("red", "blue")
title("Kaplan-Meier curve of two financial-aid-received groups")
```



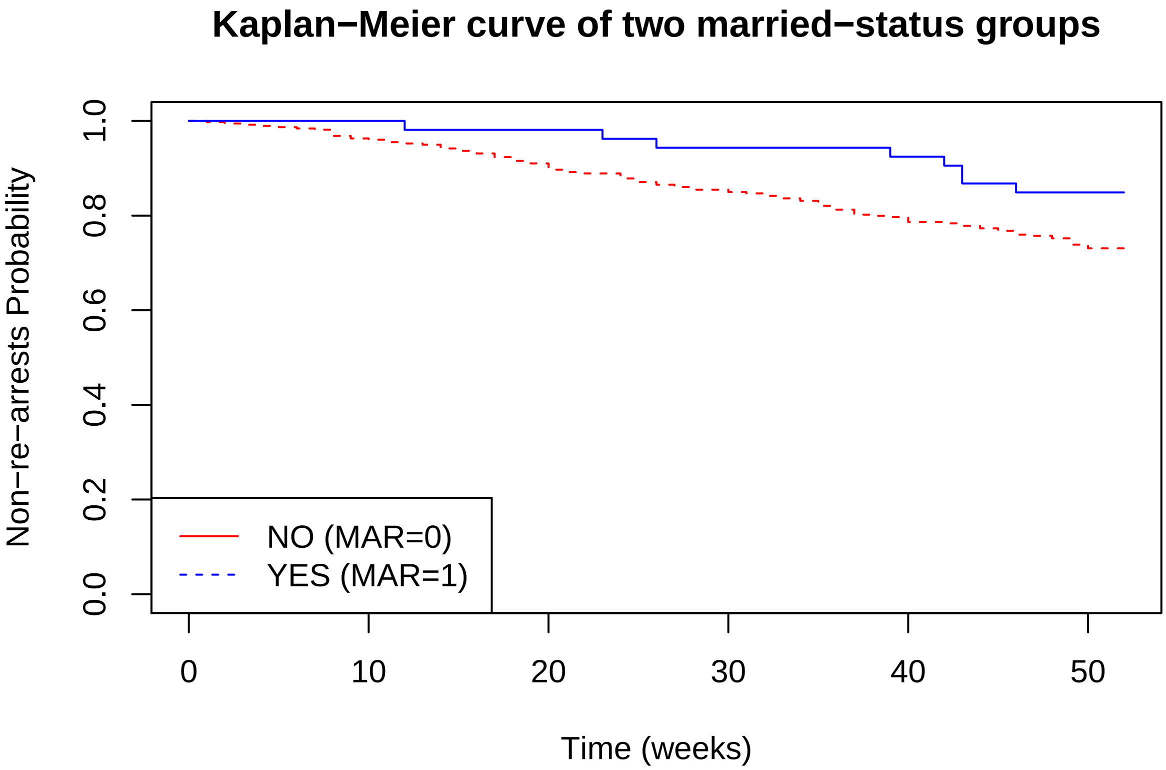
```
fit <- survfit(Surv(WEEK,ARREST)~RACE, data = df,conf.type = "none")
plot(fit, xlab="Time (weeks)", ylab="Non-re-arrests Probability",
   conf.int=FALSE, col=c("red", "blue"), lty=c(2,1))
legend("bottomleft", c("Not African American (RACE=0)", "African American(RACE=1) "), col=c("red", "blu
title("Kaplan-Meier curve of two race groups")
```



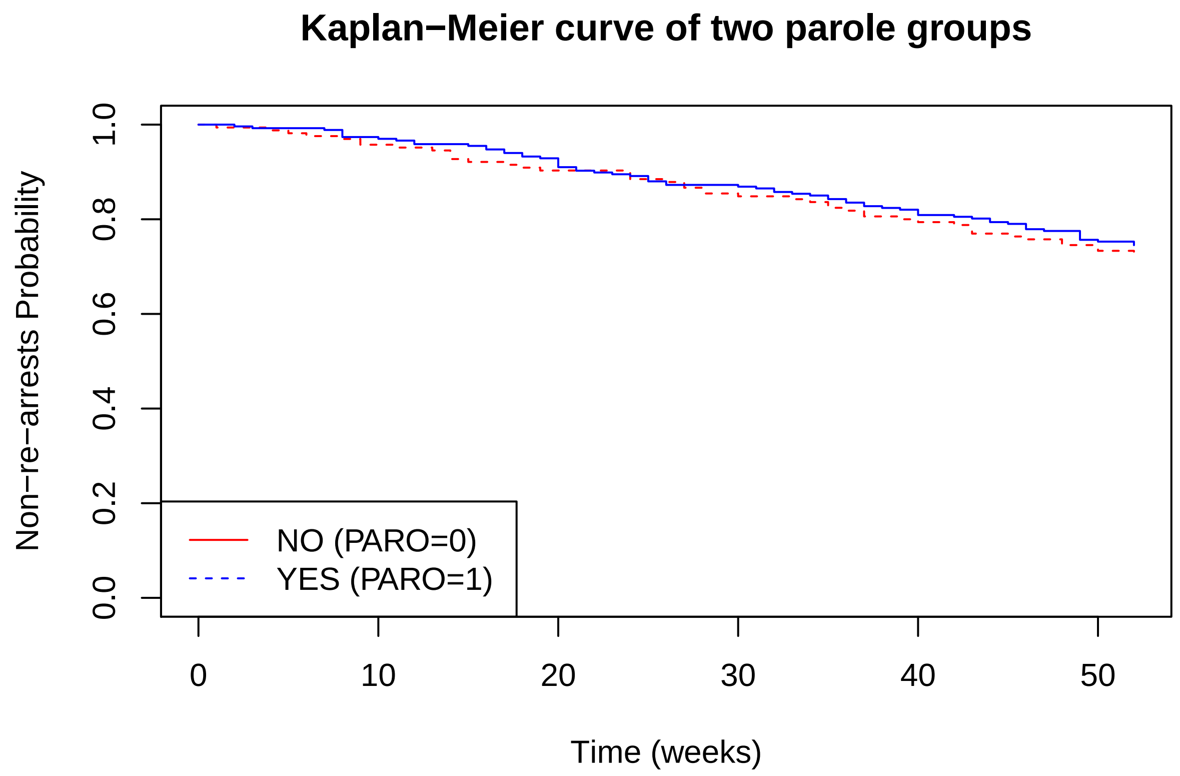
```
fit <- survfit(Surv(WEEK,ARREST)~WEXP, data = df,conf.type = "none")
plot(fit, xlab="Time (weeks)", ylab="Non-re-arrests Probability",
   conf.int=FALSE, col=c("red", "blue"), lty=c(2,1))
legend("bottomleft", c("No (WEXP=0)", "YES (WEXP=1) "), col=c("red", "blue"), lty=c(1,2))
title("Kaplan-Meier curve of two work experience groups")
```



```
fit <- survfit(Surv(WEEK,ARREST)~MAR, data = df,conf.type = "none")
plot(fit, xlab="Time (weeks)", ylab="Non-re-arrests Probability",
   conf.int=FALSE, col=c("red", "blue"), lty=c(2,1))
legend("bottomleft", c("NO (MAR=0)", "YES (MAR=1) "), col=c("red", "blue"), lty=c(1,2))
title("Kaplan-Meier curve of two married-status groups")
```



```r
fit <- survfit(Surv(WEEK,ARREST)~PARO, data = df,conf.type = "none")
plot(fit, xlab="Time (weeks)", ylab="Non-re-arrests Probability",
   conf.int=FALSE, col=c("red", "blue"), lty=c(2,1))
legend("bottomleft", c("NO (PARO=0)", "YES (PARO=1) "), col=c("red", "blue"), lty=c(1,2))
title("Kaplan-Meier curve of two parole groups")
```

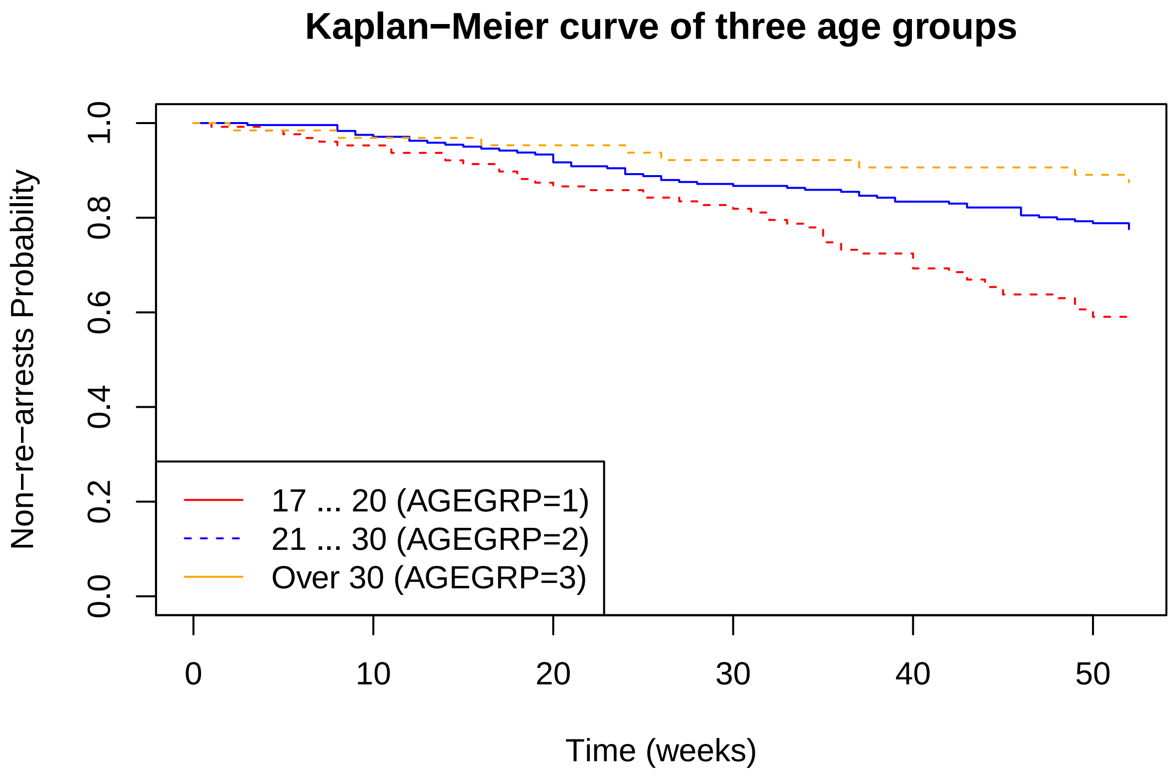


```r
fit <- survfit(Surv(WEEK,ARREST)~AGEGRP, data = df,conf.type = "none")
plot(fit, xlab="Time (weeks)", ylab="Non-re-arrests Probability",
   conf.int=FALSE, col=c("red", "blue","orange"), lty=c(2,1))
legend("bottomleft", c("17 – 20 (AGEGRP=1)", "21 – 30 (AGEGRP=2)","Over 30 (AGEGRP=3)"), col=c("red", "
```

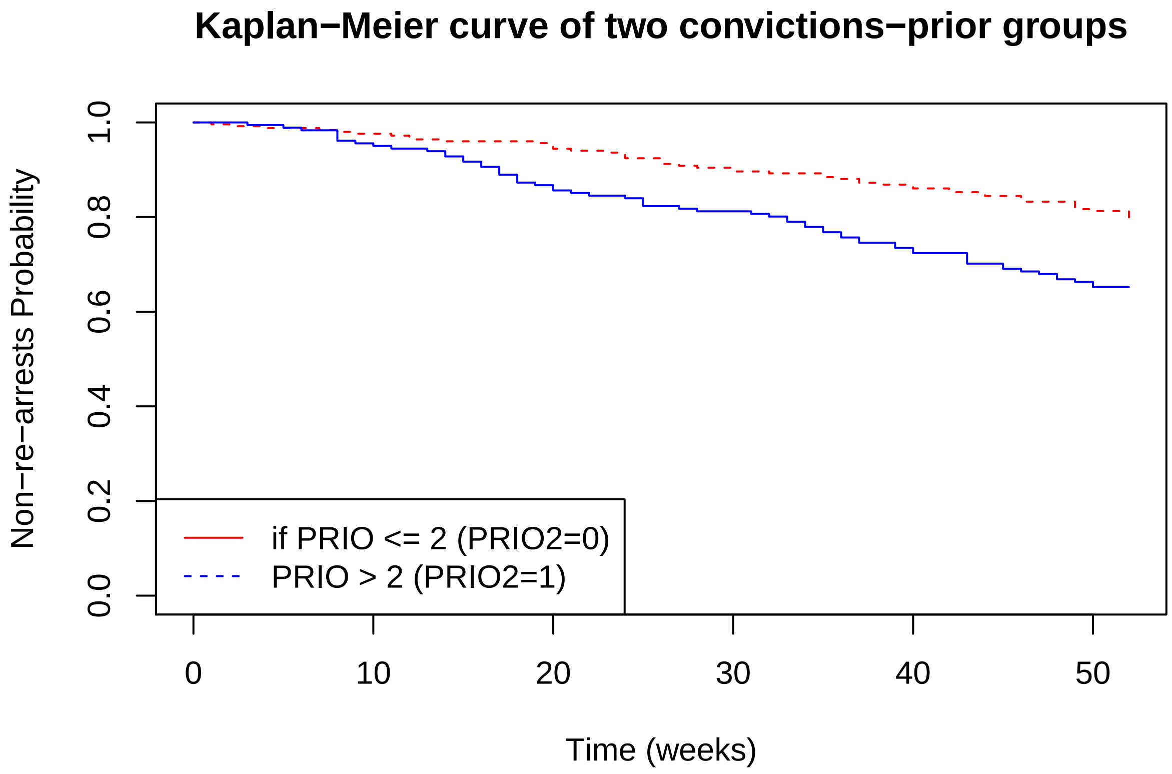
```
## Warning in strwidth(legend, units = "user", cex = cex, font = text.font):
## conversion failure on '17 – 20 (AGEGRP=1)' in 'mbcsToSbcs': dot substituted
## for <e2>

## Warning in strwidth(legend, units = "user", cex = cex, font = text.font):
## conversion failure on '17 – 20 (AGEGRP=1)' in 'mbcsToSbcs': dot substituted
## for <80>

## Warning in strwidth(legend, units = "user", cex = cex, font = text.font):
## conversion failure on '17 – 20 (AGEGRP=1)' in 'mbcsToSbcs': dot substituted
## for <93>

## Warning in strwidth(legend, units = "user", cex = cex, font = text.font):
## conversion failure on '21 – 30 (AGEGRP=2)' in 'mbcsToSbcs': dot substituted
## for <e2>

## Warning in strwidth(legend, units = "user", cex = cex, font = text.font):
## conversion failure on '21 – 30 (AGEGRP=2)' in 'mbcsToSbcs': dot substituted
## for <80>

## Warning in strwidth(legend, units = "user", cex = cex, font = text.font):
```

```
## conversion failure on '21 - 30 (AGEGRP=2)' in 'mbcsToSbcs': dot substituted
## for <93>

## Warning in text.default(x, y, ...): conversion failure on '17 - 20
## (AGEGRP=1)' in 'mbcsToSbcs': dot substituted for <e2>

## Warning in text.default(x, y, ...): conversion failure on '17 - 20
## (AGEGRP=1)' in 'mbcsToSbcs': dot substituted for <80>

## Warning in text.default(x, y, ...): conversion failure on '17 - 20
## (AGEGRP=1)' in 'mbcsToSbcs': dot substituted for <93>

## Warning in text.default(x, y, ...): conversion failure on '21 - 30
## (AGEGRP=2)' in 'mbcsToSbcs': dot substituted for <e2>

## Warning in text.default(x, y, ...): conversion failure on '21 - 30
## (AGEGRP=2)' in 'mbcsToSbcs': dot substituted for <80>

## Warning in text.default(x, y, ...): conversion failure on '21 - 30
## (AGEGRP=2)' in 'mbcsToSbcs': dot substituted for <93>
```

```
title("Kaplan-Meier curve of three age groups")
```



```
fit <- survfit(Surv(WEEK,ARREST)~PRIO2, data = df,conf.type = "none")
plot(fit, xlab="Time (weeks)", ylab="Non-re-arrests Probability",
   conf.int=FALSE, col=c("red", "blue"), lty=c(2,1))
legend("bottomleft", c("if PRIO <= 2 (PRIO2=0)", "PRIO > 2 (PRIO2=1) "), col=c("red", "blue"), lty=c(1,
title("Kaplan-Meier curve of two convictions-prior groups")
```



b. Create a table summarizing the log-rank test results for the 7 variables.

Please see the attached hand-written paper which I summarized based on the results generated by the following code.

```
survdiff(Surv(WEEK,ARREST) ~ FIN, data=df)
survdiff(Surv(WEEK,ARREST) ~ RACE, data=df)
survdiff(Surv(WEEK,ARREST) ~ WEXP, data=df)
survdiff(Surv(WEEK,ARREST) ~ MAR, data=df)
survdiff(Surv(WEEK,ARREST) ~ PARO, data=df)
survdiff(Surv(WEEK,ARREST) ~ AGEGRP, data=df)
survdiff(Surv(WEEK,ARREST) ~ PRIO2, data=df)
```

Describe the relationships that you see in the context of the current study and give your conclusions based on the log-rank test for each predictor.

We want to interpret both interpretation of logrank test and the interpretation of KM curves. Note that I have included the former in the hand-written table. Now I'm going to interpret KM curves.
- For the variable FIN, we see those receiving financial assistance seemed to have higher probability of not being re-arrested.
- For the variable RACE, we see those non-African American seemed to have higher probability of not being re-arrested.
- For the variable WEXP, we see those inmate had full-time work experience before incarceration seemed to have higher probability of not being re-arrested then inmate did not have full-time work experience before incarceration.
- For the variable MAR, we see those inmate married at time of release seemed to have higher probability of not being re-arrested than inmate who did not.
- For the variable PAROLE, we see those inmate released on parole seemed to have same probability of re-arrested comparing to inmate who was not being on parole.

- For the variable AGEGRP, we see those inmate at release when he/she was over 30 seemed to have higher probability of not being re-arrested than inmate was at 21-30, and it followed that inmate was at 21-30 which has the lowest probability of not being re-arrested among all the three age groups.
- For the variable PRIO, we see those inmate who had less or equal than 2 times of convictions prior to incarceration seemed to have higher probability of re-arrested comparing to inmate who had more than 2 times of convictions prior to incarceration.

If you observe a significant association between AGEGRP and time to re-arrest, conduct pairwise tests to determine which age groups are significantly different from one another. Use a Bonferroni adjustment for all pairwise tests and report the adjusted value of (i.e.,alpha*) used when assessing the significance of the raw p-values.

We do observe a significant difference.

```r
lrbon <- function(formula, data){
  logrank = survdiff(formula, data=data)
  ngroups = length(logrank$n)
  rawpvals = matrix(NA, ngroups, ngroups)
  adjpvals = matrix(NA, ngroups, ngroups)
  rownames(rawpvals) = dimnames(logrank$n)$groups
  colnames(rawpvals) = dimnames(logrank$n)$groups
  O = logrank$obs
  E = logrank$exp
  U = matrix((O-E), nrow=length(O))
  V = logrank$var
  for (i in 1:ngroups) {
    for (j in (1:ngroups)[-i]) {
      X2ij = (U[i] - U[j])^2/(V[i,i]+V[j,j]-2*V[i,j])
      rawpvals[i,j] <- 1-pchisq(X2ij,1)
      adjpvals[i,j] <- min(1,rawpvals[i,j]*ngroups*(ngroups-1)/2)
    }
  }
list("Raw p-values", rawpvals, "Bonferroni-adjusted p-values", adjpvals)
}

lrbon(formula=Surv(WEEK,ARREST) ~ AGEGRP, data=df)
```

```
## [[1]]
## [1] "Raw p-values"
##
## [[2]]
##              AGEGRP=1     AGEGRP=2     AGEGRP=3
## AGEGRP=1           NA 0.0004910547 4.986644e-06
## AGEGRP=2 4.910547e-04           NA 9.264032e-01
## AGEGRP=3 4.986644e-06 0.9264032235           NA
##
## [[3]]
## [1] "Bonferroni-adjusted p-values"
##
## [[4]]
##              [,1]        [,2]         [,3]
## [1,]           NA 0.001473164 1.495993e-05
## [2,] 1.473164e-03          NA 1.000000e+00
## [3,] 1.495993e-05 1.000000000           NA
```

The hypothesis is still the same as we listed in the Table 1 for AGEGRP.

- Log-rank test p = 1e-05
- group number P = 3 with c = P(P − 1)/2 = 3 possible pairwise comparisons
- $\alpha^* = 0.05/c = 0.05/3 = 0.0167$ (On lecture 3 slides page 61, we can see that we use raw p value compare with α^*)
- There is a significant difference between age group 1 (17 – 20) and 2 (21 – 30) ($p = 4.910547\text{e-}04 < \alpha^* = 0.0167$); and between age group 1 (17 – 20) and age group 3 (Over 30) ($p = 4.986644\text{e-}06 < \alpha^* = 0.0167$).
- We cannot conclude there is a significant difference between age age group 2 and 3 (21 – 30, and over 30) because their raw p-value is 0.9264032235 which is greater than α^*.

c. Conduct a test for trend for AGEGRP using weights (1, 2, 3). We hypothesize that the hazard of re-arrest decreases with age.

```r
lrtrend <- function(formula, weights, data){
  logrank = survdiff(formula, data=data)
  df = length(logrank$n) - 1
  O = logrank$obs
  E = logrank$exp
  U = matrix((O-E), nrow=length(O))
  V = logrank$var
  w = weights
  z = matrix(w, nrow=length(w))
  Xtrend = (t(z)%*%U)/sqrt(t(z)%*% V %*% z)

  cat("\nLog-rank Test: Chi^2(", df, " df) = ", logrank$chisq, ", p = ", 1-pchisq(logrank$chisq, df), se
  cat("\n Test Trend: Xtrend ~ N(0,1) = ", Xtrend, ", 2-sided p = ", 2*(1-pnorm(abs(Xtrend))),
      ", lower p = ", pnorm(Xtrend),
      ", upper p = ", 1-pnorm(Xtrend), sep="" )
}
lrtrend(formula=Surv(WEEK,ARREST) ~ AGEGRP, weights=c(1, 2, 3), data=df)
```

```
##
## Log-rank Test: Chi^2(2 df) = 22.29816, p = 1.43885e-05
##  Test Trend: Xtrend ~ N(0,1) = -4.565349, 2-sided p = 4.986644e-06, lower p = 2.493322e-06, upper p =
```

- Since we hypothesize that hazard of re-arrest decreases with age, we are doing an lower-tail test - Hypothesis: $H_0 : S_1(t) = S_2(t) = S_3(t)$ for $t \leq 52$ vs. $H_1 : S_1(t) < S_2(t) < S_3(t)$ for $t \leq 52$

- Significant level: two-sided $\alpha = 0.05$

- Test statistic: $X_{trend} = 22.29816$
- At $\alpha = 0.05$, reject H0 in favor of 1-sided upper-tailed H1 if z > $z_{1.05} = z_{.95} = 1.645$

- Since $X_{trend} = 22.29816 > 1.645$, reject H0 with p = 2.493322e-06 (we look at lower p value because decrease hazards)
- There is evidence to suggest age is negatively associated with hazard of re-arrest. The reason why negatively associated here because of decrease.
- The results extend the previous conclusion that the survival curves are not equal across all three groups.

2. Initial Cox Proportional Hazards Model Building:

a. For each of the 7 predictor variables (use quantitative variable PRIO instead of dichotomous PRIO2 for this question), build a univariate Cox proportional hazards regression model.

```
options(show.signif.stars=FALSE)
library(survival)
# Use level "0" as the reference category for FIN, RACE, WEXP, MAR, and PARO, level "3" as the referenc
df$FIN <- factor(df$FIN, levels = c(0,1), labels = c("NO", "YES")) # LABEL FUNCTION
df$RACE <- factor(df$RACE, levels = c(0,1), labels = c("Not African American", "African American"))
df$WEXP <- factor(df$WEXP, levels = c(0,1), labels = c("NO", "YES"))
df$MAR <- factor(df$MAR, levels = c(0,1), labels = c("NO", "YES"))
df$PARO <- factor(df$PARO, levels = c(0,1), labels = c("NO", "YES"))
df$AGEGRP <- factor(df$AGEGRP, levels = c(1,2,3), labels = c("17 - 20", "21 - 30", "Over 30"))
# We treat PRIO as a continuous variable.

cox_fin <- coxph(Surv(WEEK, ARREST) ~ relevel(FIN,"NO"), data=df, ties="breslow")# SET BASELINE - here
cox_race <- coxph(Surv(WEEK, ARREST) ~ relevel(RACE,"Not African American"), data=df, ties="breslow")
cox_wexp <- coxph(Surv(WEEK, ARREST) ~ relevel(WEXP,"NO"), data=df, ties="breslow")
cox_mar <- coxph(Surv(WEEK, ARREST) ~ relevel(MAR,"NO"), data=df, ties="breslow")
cox_paro <- coxph(Surv(WEEK, ARREST) ~ relevel(PARO,"NO"), data=df, ties="breslow")
cox_agegrp <- coxph(Surv(WEEK, ARREST) ~ relevel(AGEGRP,"Over 30"), data=df, ties="breslow")
cox_prio <- coxph(Surv(WEEK, ARREST) ~ PRIO, data=df, ties="breslow")

summary(cox_fin)
```

```
## Call:
## coxph(formula = Surv(WEEK, ARREST) ~ relevel(FIN, "NO"), data = df,
##     ties = "breslow")
##
##   n= 432, number of events= 114
##
##                          coef exp(coef) se(coef)      z Pr(>|z|)
## relevel(FIN, "NO")YES -0.3686    0.6917   0.1897 -1.943    0.052
##
##                       exp(coef) exp(-coef) lower .95 upper .95
## relevel(FIN, "NO")YES    0.6917      1.446    0.4769     1.003
##
## Concordance= 0.546  (se = 0.023 )
## Likelihood ratio test= 3.83  on 1 df,   p=0.05
## Wald test            = 3.77  on 1 df,   p=0.05
## Score (logrank) test = 3.82  on 1 df,   p=0.05
```

```
summary(cox_race)
```

```
## Call:
## coxph(formula = Surv(WEEK, ARREST) ~ relevel(RACE, "Not African American"),
##     data = df, ties = "breslow")
##
##   n= 432, number of events= 114
##
##                                                         coef exp(coef)
## relevel(RACE, "Not African American")African American 0.2305    1.2593
##                                                       se(coef)     z
## relevel(RACE, "Not African American")African American   0.3052 0.755
##                                                       Pr(>|z|)
## relevel(RACE, "Not African American")African American     0.45
##
##                                                       exp(coef) exp(-coef)
## relevel(RACE, "Not African American")African American     1.259     0.7941
##                                                       lower .95 upper .95
## relevel(RACE, "Not African American")African American    0.6924      2.29
##
## Concordance= 0.514  (se = 0.014 )
## Likelihood ratio test= 0.61  on 1 df,   p=0.4
## Wald test            = 0.57  on 1 df,   p=0.5
## Score (logrank) test = 0.57  on 1 df,   p=0.4
```

```
summary(cox_wexp)
```

```
## Call:
## coxph(formula = Surv(WEEK, ARREST) ~ relevel(WEXP, "NO"), data = df,
##     ties = "breslow")
##
##   n= 432, number of events= 114
##
##                            coef exp(coef) se(coef)      z Pr(>|z|)
## relevel(WEXP, "NO")YES -0.5825    0.5585   0.1881 -3.096  0.00196
##
##                        exp(coef) exp(-coef) lower .95 upper .95
## relevel(WEXP, "NO")YES    0.5585       1.79    0.3863    0.8076
##
## Concordance= 0.577  (se = 0.023 )
## Likelihood ratio test= 9.61  on 1 df,   p=0.002
## Wald test            = 9.58  on 1 df,   p=0.002
## Score (logrank) test = 9.86  on 1 df,   p=0.002
```

```
summary(cox_mar)
```

```
## Call:
## coxph(formula = Surv(WEEK, ARREST) ~ relevel(MAR, "NO"), data = df,
##     ties = "breslow")
##
##   n= 432, number of events= 114
##
##                           coef exp(coef) se(coef)      z Pr(>|z|)
## relevel(MAR, "NO")YES -0.7106    0.4913   0.3667 -1.938   0.0526
##
##                       exp(coef) exp(-coef) lower .95 upper .95
## relevel(MAR, "NO")YES    0.4913      2.035    0.2395     1.008
##
## Concordance= 0.533  (se = 0.013 )
## Likelihood ratio test= 4.62  on 1 df,   p=0.03
## Wald test            = 3.76  on 1 df,   p=0.05
## Score (logrank) test = 3.92  on 1 df,   p=0.05
```

```
summary(cox_paro)
```

```
## Call:
## coxph(formula = Surv(WEEK, ARREST) ~ relevel(PARO, "NO"), data = df,
##     ties = "breslow")
##
##   n= 432, number of events= 114
##
##                            coef exp(coef) se(coef)      z Pr(>|z|)
## relevel(PARO, "NO")YES -0.1086    0.8970   0.1909 -0.569    0.569
##
##                         exp(coef) exp(-coef) lower .95 upper .95
## relevel(PARO, "NO")YES     0.897      1.115     0.617     1.304
##
## Concordance= 0.513  (se = 0.023 )
## Likelihood ratio test= 0.32  on 1 df,   p=0.6
## Wald test            = 0.32  on 1 df,   p=0.6
## Score (logrank) test = 0.32  on 1 df,   p=0.6
```

```
summary(cox_agegrp)
```

```
## Call:
## coxph(formula = Surv(WEEK, ARREST) ~ relevel(AGEGRP, "Over 30"),
##     data = df, ties = "breslow")
##
##   n= 432, number of events= 114
##
##                                       coef exp(coef) se(coef)     z Pr(>|z|)
## relevel(AGEGRP, "Over 30")17 - 20 1.3487    3.8524   0.3800 3.550 0.000386
## relevel(AGEGRP, "Over 30")21 - 30 0.6392    1.8950   0.3789 1.687 0.091572
##
##                                   exp(coef) exp(-coef) lower .95 upper .95
## relevel(AGEGRP, "Over 30")17 - 20     3.852     0.2596    1.8294     8.112
## relevel(AGEGRP, "Over 30")21 - 30     1.895     0.5277    0.9018     3.982
##
## Concordance= 0.61  (se = 0.024 )
## Likelihood ratio test= 21.32  on 2 df,   p=2e-05
## Wald test            = 20.54  on 2 df,   p=3e-05
## Score (logrank) test = 22.18  on 2 df,   p=2e-05
```

```
summary(cox_prio)
```

```
## Call:
## coxph(formula = Surv(WEEK, ARREST) ~ PRIO, data = df, ties = "breslow")
##
##   n= 432, number of events= 114
##
##          coef exp(coef) se(coef)     z Pr(>|z|)
## PRIO 0.10066   1.10590  0.02669 3.771 0.000162
##
##      exp(coef) exp(-coef) lower .95 upper .95
## PRIO     1.106     0.9042      1.05     1.165
##
## Concordance= 0.588  (se = 0.028 )
## Likelihood ratio test= 11.94  on 1 df,   p=5e-04
## Wald test            = 14.22  on 1 df,   p=2e-04
## Score (logrank) test = 14.56  on 1 df,   p=1e-04
```

Create a table summarizing the parameter estimate(s), standard error, Wald test statistic, p-value, estimated hazard ratio(s), and 95% confidence interval for the hazard ratio(s) from each model.

Please see the attached hand-written paper page 2 - Table 2.

Does each hazard ratio agree with what you saw in your Kaplan-Meier curves from Question 1? Explain.

- For the variable FIN, we see those receiving financial assistance seemed to have higher probability of not being re-arrested from Question 1. Here the hazard ratio is $0.6917 < 1$ which means those receiving financial assistance have lower hazard which means better survival. Therefore, it agrees with what we saw in the KM curves from Question 1.
- For the variable RACE, we see those non-African American (0) seemed to have higher probability of not being re-arrested from Question 1. Here the hazard ratio is $1.2593 > 1$, which means African American has more hazard than baseline non-African American, therefore lower survival of not being re-arrested. Therefore, it agrees with what we saw in the KM curves from Question 1.
- For the variable WEXP, we see those inmate had full-time work experience before incarceration seemed to have higher probability of not being re-arrested then inmate did not have full-time work experience before incarceration from Question 1. The hazard is $0.5585 < 1$ that means, the inmate with work experience before has lower hazard and therefore better survival. Therefore, it agrees with what we saw in the KM curves from Question 1.
- For the variable MAR, we see those inmate married at time of release seemed to have higher probability of not being re-arrested than inmate who did not from Question 1. The hazard is $0.4913 < 1$ which means inmate who marries at time of release has lower hazard and therefore better survival of not being re-arrested. Therefore, it agrees with what we saw in the KM curves from Question 1.
- For the variable PAROLE, we see those inmate released on parole seemed to have same probability of re-arrested comparing to inmate who was not being on parole from Question 1. The hazard ratio is 0.8970 close to 1. It means inmate who released on parole has about the same hazard than the inmate who did not, therefore the survival is about the same. Hence, it agrees with what we saw in the KM curves from Question 1.
- For the variable AGEGRP, we see those inmate at release when he/she was over 30 seemed to have higher probability of not being re-arrested than inmate was at 21-30, and it followed that inmate was at 21-30 which has the lowest probability of not being re-arrested among all the three age groups from Question 1. Our first hazard ratio is 3.8524 for 17-20 vs baseline >30. That means 17-20 has higher hazard than >30 group and therefore lower survival probability. The second hazard ratio is 1.8950 for 21-30 vs baseline >30. Similarly, it means 21-30 has higher hazard and therefore lower survival. It agrees with what we saw in the KM curves from Question 1, although gives less information than Q1.
- For the variable PRIO, we see those inmate who had less or equal than 2 times of convictions prior to incarceration seemed to have higher probability of re-arrested comparing to inmate who had more than 2 times of convictions prior to incarceration from Question 1. The hazard is $1.1059 > 1$. It's not comparable because here we treat PRIO as continuous variable, while categorical in the Q1.

Interpret each unadjusted hazard ratio in the context of the current study

- Those received financial aid received after release group are getting re-arrest at 0.6917 the rate of those did not receive financial aid received after release group.
- Those who are African American are getting re-arrest at 1.2593 the rate of those non African American.
- Those had full-time work experience before incarceration are getting re-arrest at 0.5585 the rate of those did not have full-time work experience before incarceration.
- Those married at the time of release are getting re-arrest at 0.4913 the rate of those did not get married.
- Those released on parole are getting re-arrest at 0.8970 the rate of those did not receive on parole.

- Those who was at age 17-20 are getting re-arrest at 3.8524 the rate of those who were over 30. Those who was at age 21-30 are getting re-arrest at 1.8950 the rate of those who were over 30.

- For every additional (one-unit increase) conviction, the risk of getting re-arrest is 1.1059 times than the previous risk, which is increasing by 10%.

and assess the statistical significance of the parameter(s) in each model using the Wald test (interpret the HR even if statistical significance is not achieved).

- Model: $lnh(t,x) = lnh_0(t) + \beta_1 FIN$
  Significane level: two sided, $\alpha = 0.05$
  Use Wald Test to test: $H_0 : \beta_1 = 0 vs. H_1 : \beta_1 \neq 0$
  Test statistics: $W^2 = 3.77$
  Decision rule: At $\alpha = 0.05$, reject H_0 if $W^2 \geq \chi_{1-0.05,1df} = 3.84$
  Statistical conclusion: Since $W_2 = 3.77 < 3.84$, reject H_0, p = 0.052
  Interpretation: We fail to reject the null hypothesis and β_1 is no difference with 0. That means HR = $exp(\beta_1) = e^0 = 1$ or hazards have no difference in two financial-aid-received groups. Hazard ratio is not statistically significant.

- Model: $lnh(t,x) = lnh_0(t) + \beta_1 RACE$
  Significane level: two sided, $\alpha = 0.05$
  Use Wald Test to test: $H_0 : \beta_1 = 0 vs. H_1 : \beta_1 \neq 0$
  Test statistics: $W^2 = 0.57$
  Decision rule: At $\alpha = 0.05$, reject H_0 if $W \geq \chi_{1-0.05,1df} = 1.96$
  Statistical conclusion: Since $W_2 = 0.57 < 1.96$, fail to reject H_0, p = 0.45
  Interpretation: We fail to reject the null hypothesis and β_1 is no difference with 0. That means HR = $exp(\beta_1) = 1$ or hazards have no difference in two race groups. Therefore, hazard ratio is not statistically significant.

- Model: $lnh(t,x) = lnh_0(t) + \beta_1 WEXP$
  Significane level: two sided, $\alpha = 0.05$
  Use Wald Test to test: $H_0 : \beta_1 = 0 vs. H_1 : \beta_1 \neq 0$
  Test statistics: $W^2 = 9.58$
  Decision rule: At $\alpha = 0.05$, reject H_0 if $W^2 \geq \chi_{1-0.05,1df} = 3.84$
  Statistical conclusion: Since $W_2 = 9.58 > 3.84$, reject H_0, p = 0.00196
  Interpretation: We reject the null hypothesis and β_1 has significant difference with 0. That means HR = $exp(\beta_1) = 0.5585$ or those had full-time work experience before incarceration are getting re-arrest at 0.5585 the rate of those did not have full-time work experience before incarceration. Hazard ratio is statistically significant.

- Model: $lnh(t,x) = lnh_0(t) + \beta_1 MAR$
  Significane level: two sided, $\alpha = 0.05$
  Use Wald Test to test: $H_0 : \beta_1 = 0 vs. H_1 : \beta_1 \neq 0$
  Test statistics: $W^2 = 3.76$
  Decision rule: At $\alpha = 0.05$, reject H_0 if $W^2 \geq \chi_{1-0.05,1df} = 3.84$
  Statistical conclusion: Since $W_2 = 3.76 < 3.84$, fail to reject H_0, p = 0.0526 Interpretation: We fail to reject the null hypothesis and β_1 is no difference with 0. That means HR = $exp(\beta_1) = 1$ or hazards have no difference in two marriage groups. Hazard ratio is not statistically significant.

- Model: $lnh(t,x) = lnh_0(t) + \beta_1 PARO$
  Significane level: two sided, $\alpha = 0.05$
  Use Wald Test to test: $H_0 : \beta_1 = 0 vs. H_1 : \beta_1 \neq 0$
  Test statistics: $W^2 = 0.32$
  Decision rule: At $\alpha = 0.05$, reject H_0 if $W^2 \geq \chi_{1-0.05,1df} = 3.84$
  Statistical conclusion: Since $W_2 = 0.32 < 3.84$, fail to reject H_0, p = 0.569
  Interpretation: We fail to reject the null hypothesis and β_1 is no difference with 0. That means HR

$= exp(\beta_1) = 1$ or hazards have no difference in two on-parole groups. Hazard ratio is not statistically significant.

- Model: $lnh(t, x) = lnh_0(t) + \beta_1 AGEGRP1 + \beta_2 AGEGRP1$
We are going to do Wald test twice since it only allows to test coefficients once a time.
Significane level: two sided, $\alpha = 0.05$
Use Wald Test to test for the first age group: $H_0 : \beta_1 = 0 vs.H_1 : \beta_1 \neq 0$
Test statistics: $W^2 = 12.59$
Decision rule: At $\alpha = 0.05$, reject H_0 if $W^2 \geq \chi_{1-0.05,1df} = 3.84$
Statistical conclusion: Since $W_2 = 12.59 > 3.84$, reject H_0, p = 0.003
Interpretation: We reject the null hypothesis and β_1 has significant difference with 0. That means HR $= exp(\beta_1) = 3.8524$ or those in 17-20 age group are getting re-arrest at 3.8524 the rate of those in age group >30 . Hazard ratio is statistically significant.

  Use Wald Test to test for the first age group: $H_0 : \beta_2 = 0 vs.H_1 : \beta_2 \neq 0$
Test statistics: $W^2 = 2.86$
Decision rule: At $\alpha = 0.05$, reject H_0 if $W^2 \geq \chi_{1-0.05,1df} = 3.84$
Statistical conclusion: Since $W_2 = 2.86 < 3.84$, reject H_0, p = 0.092
Interpretation: We fail to reject the null hypothesis that β_1 does not have significant difference with 0. That means HR $= exp(\beta_1) = 1$ or hazards have no difference between those in 21-30 age group and those in age group >30 . Hazard ratio is not statistically significant.

- Model: $lnh(t, x) = lnh_0(t) + \beta_1 PRIO$
Significane level: two sided, $\alpha = 0.05$
Use Wald Test to test: $H_0 : \beta_1 = 0 vs.H_1 : \beta_1 \neq 0$
Test statistics: $W^2 = 14.22$
Decision rule: At $\alpha = 0.05$, reject H_0 if $W^2 \geq \chi_{1-0.05,1df} = 3.84$
Statistical conclusion: Since $W_2 = 14.22 > 3.84$, reject H_0, p = 0.000162
Interpretation: We reject the null hypothesis and β_1 has significant difference with 0. That means HR $= exp(\beta_1) = 1.1059$ or For every additional (one-unit increase) conviction, the risk of getting re-arrest is 1.1059 times thanthe previous risk, which is increasing by 10%. Hazard ratio is statistically significant.

Also perform a Likelihood Ratio Test for AGEGRP to determine if overall it is an important predictor.

- full model: $\hat{h}(t, x) = \hat{h}_0(t) exp(\beta_1 AGEGRP1 + \beta_2 AGEGRP2)$, reduced model: $\hat{h}(t, x) = \hat{h}_0(t)$
- H0: $\beta_1 = \beta_2 = 0$ vs. H1: At least one $\beta_j \neq 0$

```
full_cox_agegrp <- coxph(Surv(WEEK, ARREST) ~ relevel(AGEGRP,"Over 30"), data=df, ties="breslow")
reduced_cox_agegrp <- coxph(Surv(WEEK, ARREST) ~ 1, data=df, ties="breslow")
anova(full_cox_agegrp,reduced_cox_agegrp)
```

```
## Analysis of Deviance Table
##  Cox model: response is  Surv(WEEK, ARREST)
##  Model 1: ~ relevel(AGEGRP, "Over 30")
##  Model 2: ~ 1
##    loglik  Chisq Df P(>|Chi|)
## 1 -665.02
## 2 -675.68 21.318  2 2.348e-05
```

```
-2*logLik(full_cox_agegrp)
```

```
## 'log Lik.' 1330.048 (df=2)
```

```
-2*logLik(reduced_cox_agegrp)
```

```
## 'log Lik.' 1351.367 (df=0)
```

- Significane level: two sided, $\alpha = 0.05$
- Test statistics: G = -2 * logLik(reduced_cox_agegrp) - (-2 * logLik(full_cox_agegrp))= 1351.367-1330.048 = 21.319
- Decision rule: At $\alpha = 0.05$, reject H_0 if $W^2 \geq \chi_{1-0.05,2df} = 5.99$
- Statistical conclusion: Since $W_2 = 21.319 > 5.99$, reject H_0, p = 2.348e-05
- Interpretation: Reject the null hypothesis and we conclude than agegrp is associated with response hazard and therefore it's a important predictor.

3. Refining the Model:

a. Model 1: Include all predictors found to be significant at the $\alpha = 0.10$-level (i.e., p-values < 0.10) in the univariate stage (in Question 2) in a multivariate Cox proportional hazards regression model (we will call this Model 1). [Note: For AGEGRP, you will base this decision on the Likelihood Ratio Test performed in Question 2. For all other predictors, you can use the Wald Test.]

The sigificant predictors: FIN, WEXP, MAR, PRIO, AGEGRP

What is the equation of the fitted Coxregression model?

$$ln\hat{h}(t,x) = ln\hat{h}_0(t) + \beta_1 FIN + \beta_2 WEXP + \beta_3 MAR + \beta_4 PRIO + \beta_5 AGEGRP1 + \beta_6 AGEGRP2 = ln\hat{h}_0(t) - 0.33275FIN - 0.11263WEXP - 0.46035MAR + 0.09PRIO + 1.11959AGEGPR1 + 0.55069AGEGRP2$$

Create a table summarizing the parameter estimate, standard error, Wald test statistic, p-value, estimated hazard ratio, and 95% confidence interval for the hazard ratio for all parameters in Model 1.

Please the attached hand-written table page 2 - table 3.

```r
#If not relevel - then it level alphabetically.
model1  <- coxph(Surv(WEEK, ARREST) ~  relevel(FIN,"NO") + relevel(WEXP,"NO") + relevel(MAR,"NO") + PRI
summary(model1)
```

```
## Call:
## coxph(formula = Surv(WEEK, ARREST) ~ relevel(FIN, "NO") + relevel(WEXP,
##     "NO") + relevel(MAR, "NO") + PRIO + relevel(AGEGRP, "Over 30"),
##     data = df, ties = "breslow")
##
##   n= 432, number of events= 114
##
##                                        coef exp(coef) se(coef)      z
## relevel(FIN, "NO")YES              -0.33275   0.71695  0.19084 -1.744
## relevel(WEXP, "NO")YES             -0.11263   0.89348  0.21172 -0.532
## relevel(MAR, "NO")YES              -0.46035   0.63107  0.37968 -1.212
## PRIO                                0.09122   1.09551  0.02862  3.187
## relevel(AGEGRP, "Over 30")17 - 20   1.11959   3.06361  0.39667  2.822
## relevel(AGEGRP, "Over 30")21 - 30   0.55069   1.73445  0.38208  1.441
##                                    Pr(>|z|)
## relevel(FIN, "NO")YES               0.08123
## relevel(WEXP, "NO")YES              0.59473
## relevel(MAR, "NO")YES               0.22534
## PRIO                                0.00144
## relevel(AGEGRP, "Over 30")17 - 20  0.00477
## relevel(AGEGRP, "Over 30")21 - 30  0.14950
##
##                                    exp(coef) exp(-coef) lower .95 upper .95
## relevel(FIN, "NO")YES                 0.7169     1.3948    0.4932     1.042
## relevel(WEXP, "NO")YES                0.8935     1.1192    0.5900     1.353
## relevel(MAR, "NO")YES                 0.6311     1.5846    0.2998     1.328
## PRIO                                  1.0955     0.9128    1.0357     1.159
## relevel(AGEGRP, "Over 30")17 - 20     3.0636     0.3264    1.4079     6.666
## relevel(AGEGRP, "Over 30")21 - 30     1.7344     0.5766    0.8202     3.668
##
## Concordance= 0.651  (se = 0.026 )
## Likelihood ratio test= 36.53  on 6 df,   p=2e-06
## Wald test            = 36.79  on 6 df,   p=2e-06
## Score (logrank) test = 39.34  on 6 df,   p=6e-07
```

If AGEGRP is included in Model 1, perform a Likelihood Ratio Test to determine if overall it is an important predictor in this multivariate model.

- H0: $\beta_5 = \beta_6 = 0$ vs. H1: At least one $\beta_j \neq 0$

```r
model1_agegrp_full <- coxph(Surv(WEEK, ARREST) ~ FIN + WEXP + MAR+ PRIO + AGEGRP, data=df, ties="breslo
model1_agegrp_reduced <- coxph(Surv(WEEK, ARREST) ~ FIN+WEXP + MAR+PRIO, data=df, ties="breslow") #mean
anova(model1_agegrp_full,model1_agegrp_reduced)
```

```
## Analysis of Deviance Table
##  Cox model: response is  Surv(WEEK, ARREST)
##  Model 1: ~ FIN + WEXP + MAR + PRIO + AGEGRP
##  Model 2: ~ FIN + WEXP + MAR + PRIO
##    loglik  Chisq Df P(>|Chi|)
## 1 -657.42
## 2 -663.53 12.213  2  0.002229
```

```r
-2*logLik(model1_agegrp_full)
```

```
## 'log Lik.' 1314.841 (df=6)
```

```r
-2*logLik(model1_agegrp_reduced)
```

```
## 'log Lik.' 1327.054 (df=4)
```

- Significane level: two sided (always), $\alpha = 0.05$
- Test statistics: G = 12.213
- Decision rule: At $\alpha = 0.05$, reject H_0 if $W^2 \geq \chi_{1-0.05,2df} = 5.99$
- Statistical conclusion: Since $W_2 = 12.213 > 5.99$, reject H_0, p = 0.002229
- Interpretation: Reject the null hypothesis and we conclude than agegrp is associated with response hazard, therefore AGEGRP is an statistically significant predictor in this multivariate model.

b. Model 2: Remove all non-statistically significant predictors (p-value > 0.05) from Model 1 all at once (we will call this Model 2). [Note: If AGEGRP was included in Model 1, you will base this decision on the Likelihood Ratio Test performed in Question 3a. For all other predictors, you can use the Wald Test.]

Nonsignificant predictor: WEXP,FIN, AGEGRP

```r
model2  <- coxph(Surv(WEEK, ARREST) ~ PRIO + relevel(AGEGRP,"Over 30"), data=df, ties="breslow")
summary(model2)
```

```
## Call:
## coxph(formula = Surv(WEEK, ARREST) ~ PRIO + relevel(AGEGRP, "Over 30"),
##     data = df, ties = "breslow")
##
##   n= 432, number of events= 114
##
##                                      coef exp(coef) se(coef)     z
## PRIO                              0.09420   1.09878  0.02751 3.424
## relevel(AGEGRP, "Over 30")17 - 20 1.29108   3.63671  0.38046 3.393
## relevel(AGEGRP, "Over 30")21 - 30 0.61308   1.84611  0.37905 1.617
##                                   Pr(>|z|)
## PRIO                              0.000618
## relevel(AGEGRP, "Over 30")17 - 20 0.000690
## relevel(AGEGRP, "Over 30")21 - 30 0.105786
##
##                                   exp(coef) exp(-coef) lower .95 upper .95
## PRIO                                  1.099     0.9101    1.0411     1.160
## relevel(AGEGRP, "Over 30")17 - 20     3.637     0.2750    1.7253     7.666
## relevel(AGEGRP, "Over 30")21 - 30     1.846     0.5417    0.8782     3.881
##
## Concordance= 0.64  (se = 0.027 )
## Likelihood ratio test= 31.36  on 3 df,   p=7e-07
## Wald test            = 32.03  on 3 df,   p=5e-07
## Score (logrank) test = 34.24  on 3 df,   p=2e-07
```

What is the equation of the final fitted Cox regression model?

$ln\hat{h}(t,x) = ln\hat{h}_0(t) + \beta_1 PRIO18 + \beta_2 AGEGRP1 + \beta_3 AGEGRP2 = ln\hat{h}_0(t) + 0.09420PRIO18 + 1.29108AGEGRP1 + 0.61308AGEGRP2$

Create a table summarizing the parameter estimate, standard error, Wald test statistic, p-value, estimated hazard ratio, and 95% confidence interval for the hazard ratio for all parameters in Model 2.

Please see the attached hand-written sheet - Table 4.

If AGEGRP is included in Model 2, perform a Likelihood Ratio Test to determine if overall it is an important predictor in this multivariate model.

```
model2_agegrp_full <- coxph(Surv(WEEK, ARREST) ~ PRIO + relevel(AGEGRP,"Over 30"), data=df, ties="bresl
model2_agegrp_reduced <- coxph(Surv(WEEK, ARREST) ~ PRIO, data=df, ties="breslow")
anova(model2_agegrp_full,model2_agegrp_reduced)
```

```
## Analysis of Deviance Table
##  Cox model: response is  Surv(WEEK, ARREST)
##  Model 1: ~ PRIO + relevel(AGEGRP, "Over 30")
##  Model 2: ~ PRIO
##    loglik Chisq Df P(>|Chi|)
## 1 -660.00
## 2 -669.71 19.42  2 6.068e-05
```

```
-2*logLik(model2_agegrp_full)
```

```
## 'log Lik.' 1320.003 (df=3)
```

```
-2*logLik(model2_agegrp_reduced)
```

```
## 'log Lik.' 1339.423 (df=1)
```

- Significane level: two sided (always), $\alpha = 0.05$
- Test statistics: G = 19.42
- Decision rule: At $\alpha = 0.05$, reject H_0 if $W^2 \geq \chi_{1-0.05,2df} = 5.99$
- Statistical conclusion: Since $W_2 = 19.42 > 5.99$, reject H_0, p = 6.068e-05
- Interpretation: Reject the null hypothesis and we conclude than AGEGRP is associated with response hazard, therefore AGEGRP is an statistically significant predictor in this multivariate model.

Interpret the adjusted hazard ratios from this final model in the context of the current study.

For PRIO, HR = 1.1. For every additional (one-unit increase) conviction, the risk of getting re-arrest is 1.10 times than the previous risk, which is increasing by 10%.
For AGEGRP1, HR = 3.64. Those who was at age 17-20 are getting re-arrest at 3.64 the rate of those who were over 30.
For AGEGRP1, HR = 1.85. Those who was at age 21-30 are getting re-arrest at 1.85 the rate of those who were over 30.