FINAL PRE

Hongfa Cheng

Institute of Statistics & Big Data
Renmin University of China

- After the data cleaning in our proposal which can be seen in WIKI, the prepared dataset is as follows:

```
head(D)
```

```
##      date status      age    sex
## 1 0 days      1 (40,50]   male
## 2 7 days      1  (0,40]   male
## 3 7 days      1 (40,50]   male
## 4 7 days      1  (0,40]   male
## 5 8 days      1  (0,40] female
## 6 8 days      1 (40,50] female
```

Figure: Prepared Data

- Visualization of cumulative (confirmed number) function of different group people.

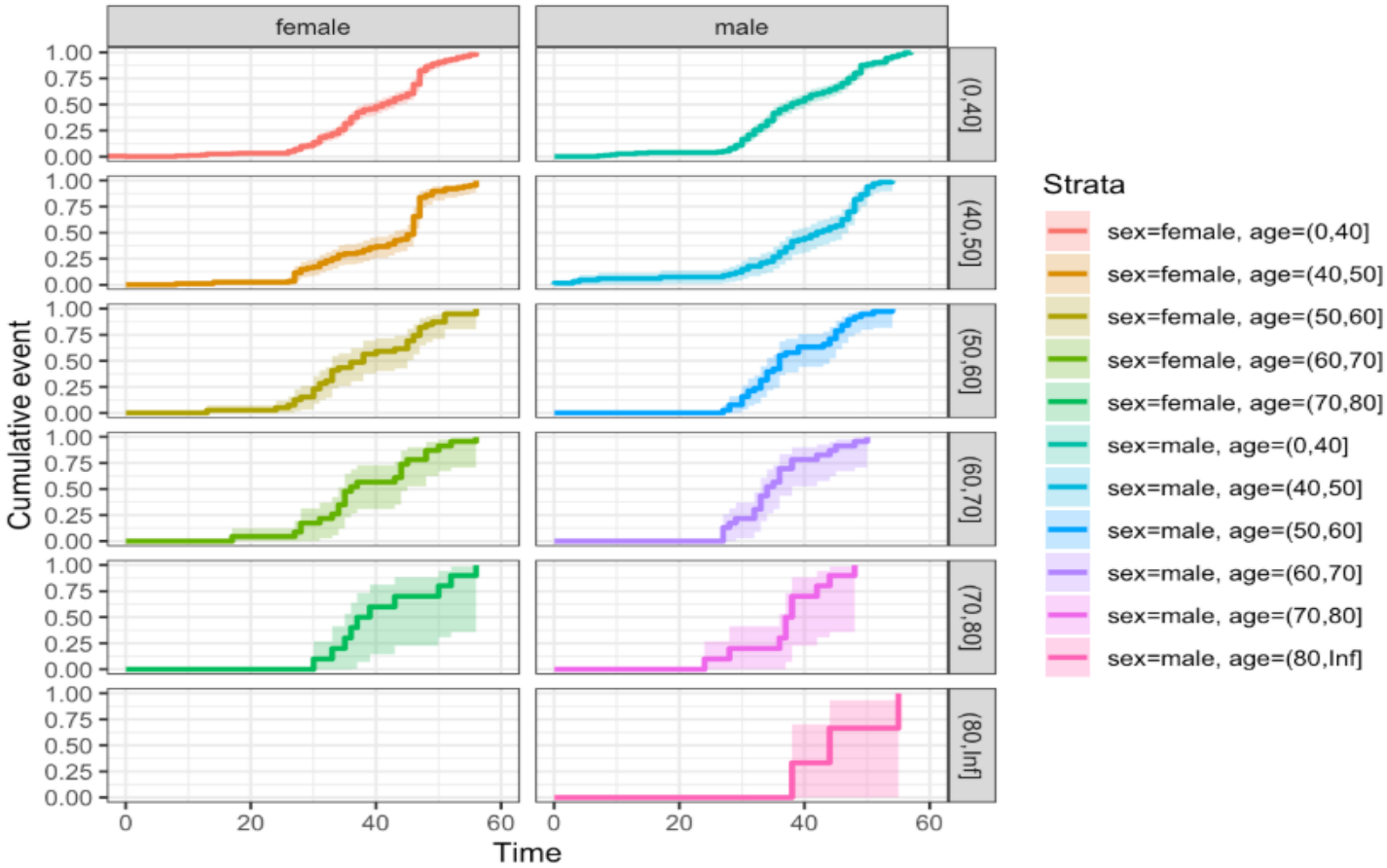


Figure: Cumulative Function 1

- Visualization of cumulative (death number) function of different group people.

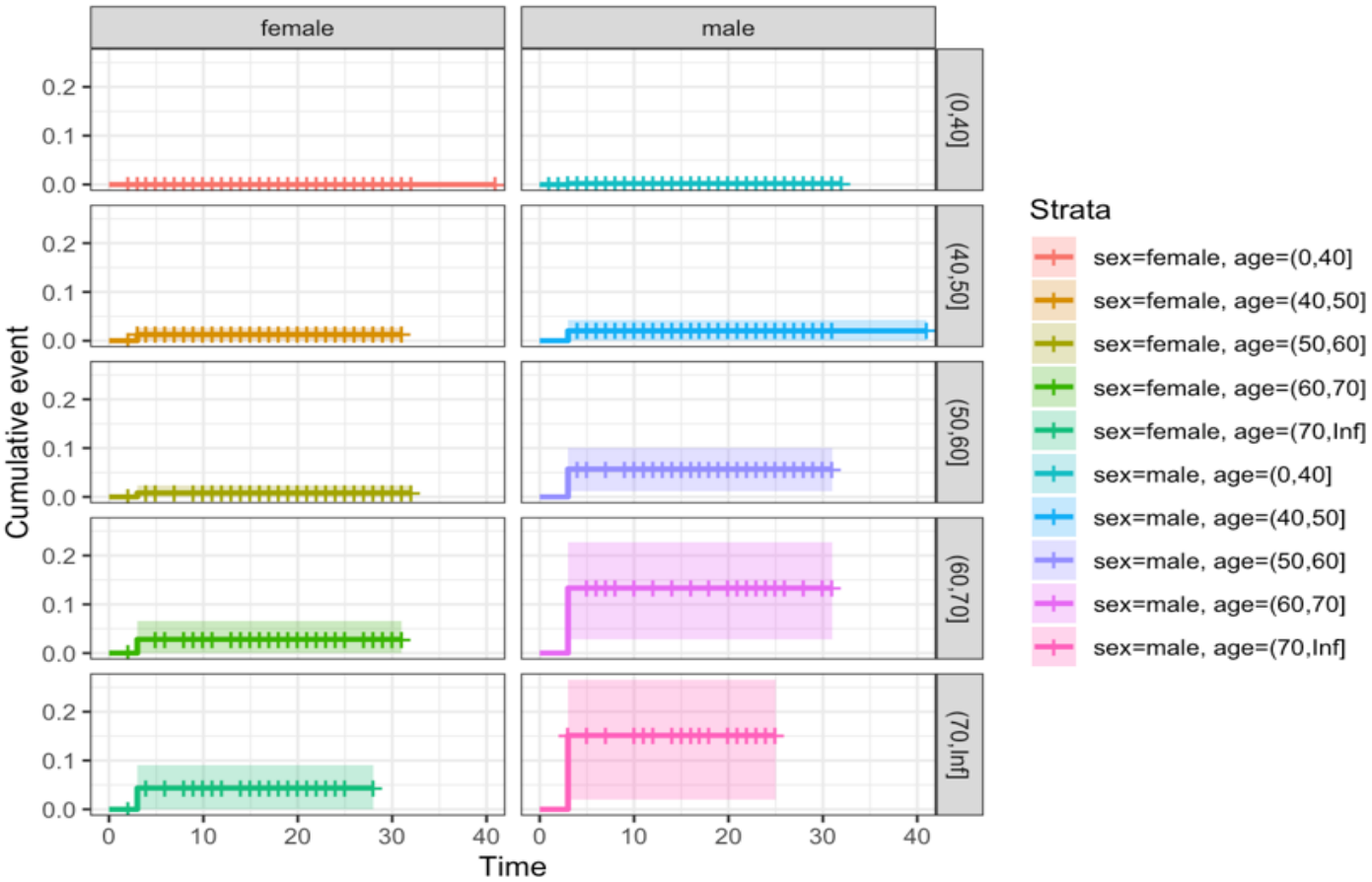


Figure: Cumulative Function 2

- We obtain our KMs for different groups of people.

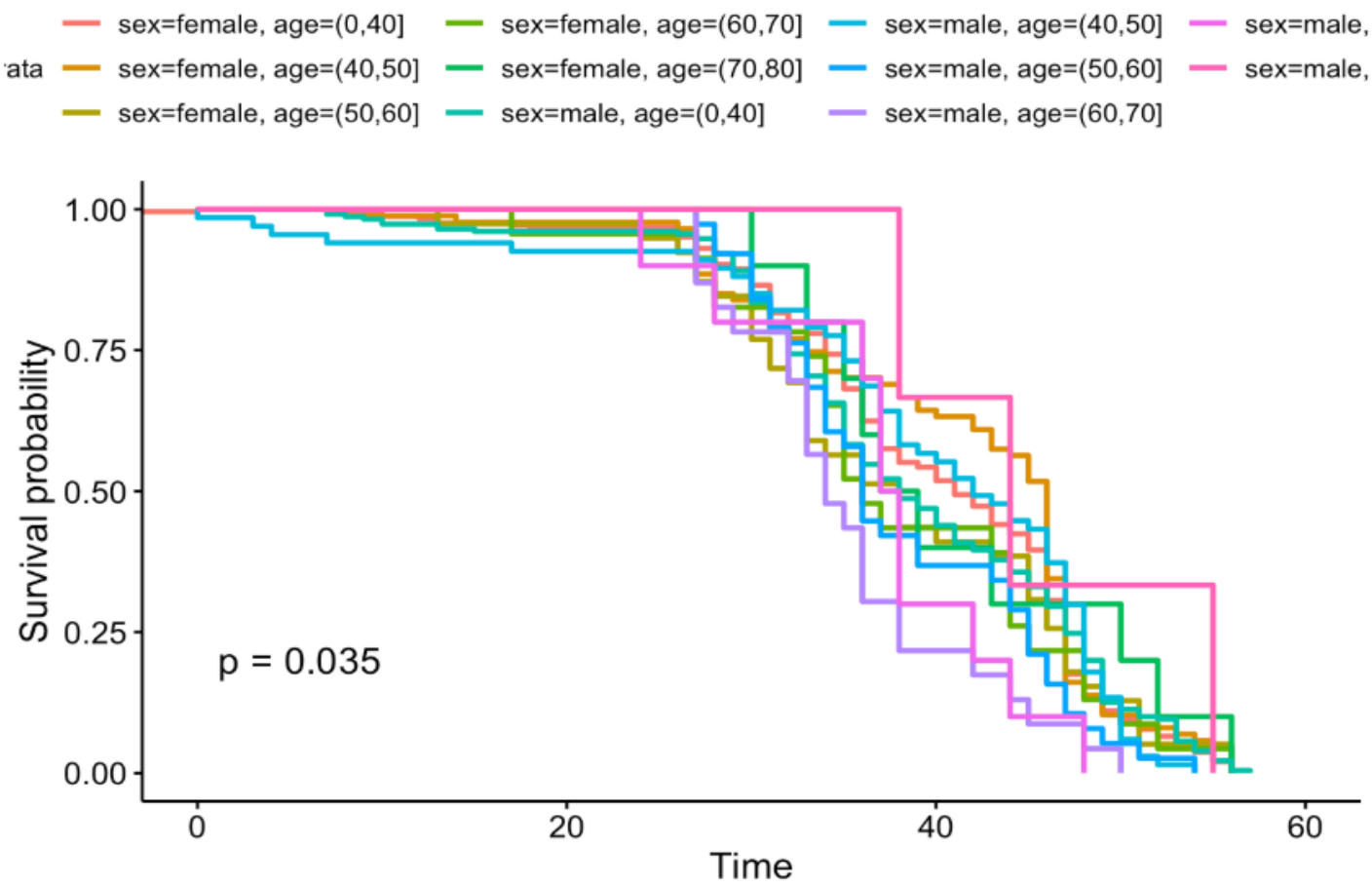


Figure: KMs of Different Groups

log-rank test

- We do log-rank test for different groups of people as above.

```
D = D[complete.cases(D),]

fit = survfit(Surv(date, status) ~ sex + age, data = D)
summary(fit)$table
```

```
##                          records n.max n.start events   *rmean *se(rmean)
## sex=female, age=(0,40]       245   245     245    245 39.98367  0.6015381
## sex=female, age=(40,50]       87    87      87     87 41.16092  1.0240694
## sex=female, age=(50,60]       39    39      39     39 38.30769  1.5546811
## sex=female, age=(60,70]       23    23      23     23 38.43478  1.9158340
## sex=female, age=(70,80]       10    10      10     10 41.10000  2.6398864
## sex=male, age=(0,40]         230   230     230    230 39.04348  0.6573643
## sex=male, age=(40,50]         67    67      67     67 39.43284  1.4362869
## sex=male, age=(50,60]         38    38      38     38 38.21053  1.1710721
## sex=male, age=(60,70]         23    23      23     23 35.52174  1.3384944
## sex=male, age=(70,80]         10    10      10     10 37.20000  2.1156559
## sex=male, age=(80,Inf]         3     3       3      3 45.66667  4.0642980
##                          median 0.95LCL 0.95UCL
## sex=female, age=(0,40]     41.0      38      44
## sex=female, age=(40,50]    46.0      43      46
## sex=female, age=(50,60]    38.0      33      45
## sex=female, age=(60,70]    36.0      34      45
## sex=female, age=(70,80]    38.0      35      NA
## sex=male, age=(0,40]       38.0      36      41
## sex=male, age=(40,50]      42.0      38      47
## sex=male, age=(50,60]      36.0      34      44
## sex=male, age=(60,70]      34.0      33      38
## sex=male, age=(70,80]      37.5      36      NA
## sex=male, age=(80,Inf]     44.0      38      NA
```

```
ggsurvplot(fit, data = D, pval = TRUE)
```

- The p-value is smaller than 0.05, which means that there is differences in these groups.

Cox Proportional Hazards Model

- The above mentioned methods Kaplan-Meier curves and logrank tests are examples of univariate analysis. They describe the survival according to one factor under investigation, but ignore the impact of any others.
- Additionally, Kaplan-Meier curves and logrank tests are useful only when the predictor variable is categorical (e.g.: treatment A vs treatment B; males vs females). They don't work easily for quantitative predictors such as gene expression, weight, or age.
- An alternative method is the Cox proportional hazards regression analysis, which works for both quantitative predictor variables and for categorical variables. Furthermore, the Cox regression model extends survival analysis methods to assess simultaneously the effect of several risk factors on survival time.

Cox Proportional Hazards Model

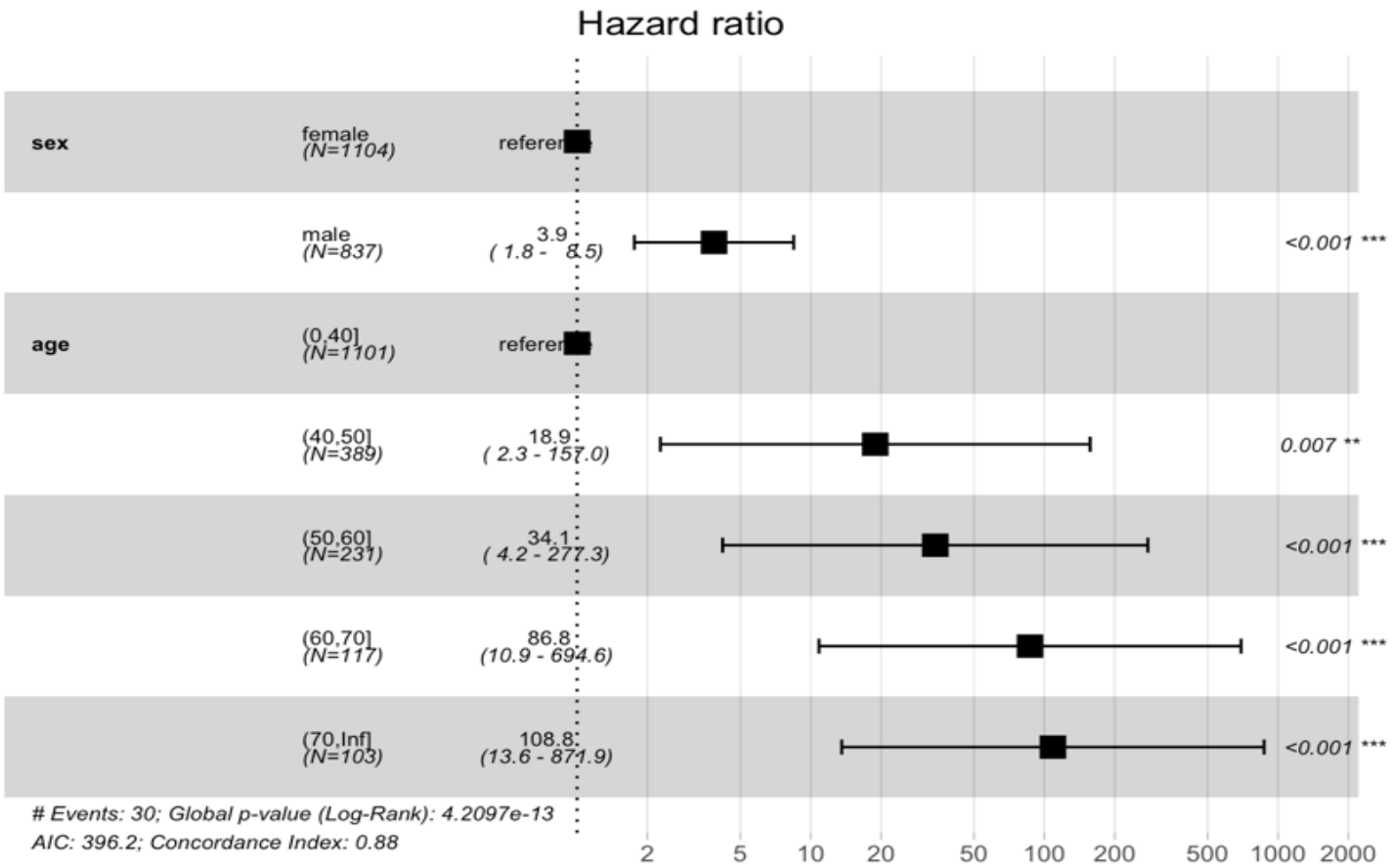
- I also fitted a Cox proportional hazards model with age groups and gender as covariates. The summary table of this model is as follows:

```
## Call:
## coxph(formula = Surv(time, state) ~ sex + age, data = D)
##
##   n= 1941, number of events= 30
##
##                 coef exp(coef) se(coef)     z Pr(>|z|)
## sexmale       1.3494    3.8552   0.4005 3.370 0.000753 ***
## age(40,50]    2.9384   18.8864   1.0804 2.720 0.006532 **
## age(50,60]    3.5297   34.1153   1.0690 3.302 0.000961 ***
## age(60,70]    4.4638   86.8145   1.0610 4.207 2.59e-05 ***
## age(70,Inf]   4.6892  108.7691   1.0620 4.416 1.01e-05 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
##             exp(coef) exp(-coef) lower .95 upper .95
## sexmale         3.855   0.259391     1.759     8.451
## age(40,50]     18.886   0.052948     2.273   156.957
## age(50,60]     34.115   0.029312     4.197   277.285
## age(60,70]     86.815   0.011519    10.851   694.599
## age(70,Inf]   108.769   0.009194    13.569   871.867
##
## Concordance= 0.878  (se = 0.025 )
## Likelihood ratio test= 67.05  on 5 df,   p=4e-13
## Wald test            = 38.31  on 5 df,   p=3e-07
## Score (logrank) test = 80.24  on 5 df,   p=7e-16
```

Figure: Cox Model

Cox Proportional Hazards Model

- The following figure shows that male is more dangerous than female and age is a bad factor.



Cox Proportional Hazards Model

- The martingale residuals seems that they have 0 means, indicating that overall the model fits the data reasonably well.

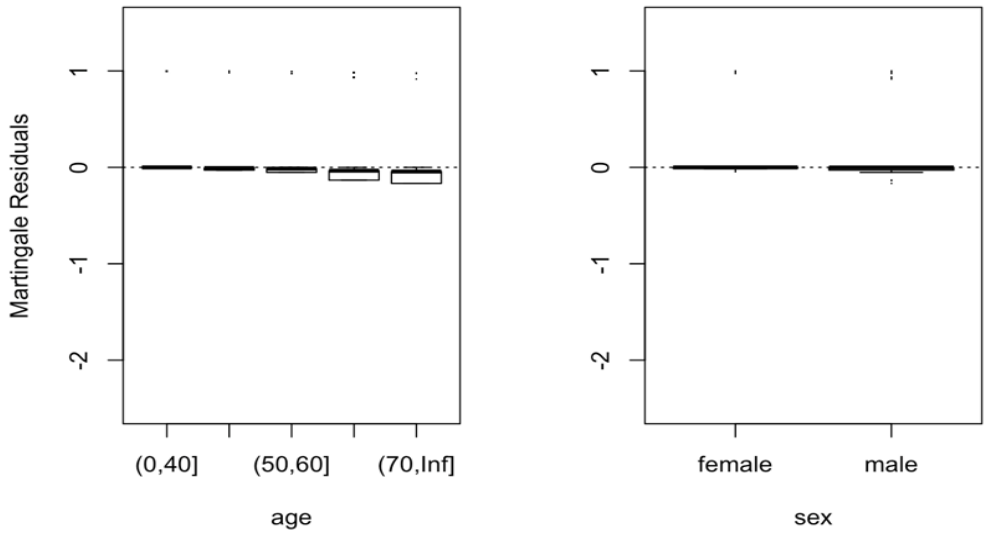


Figure: Cox-Snell residual plot

God bless our nation. Rid us of mutation.