STAT 578 Final Project: Framingham Heart Study

Allyson Meyers, Christopher Young, Shen Qu

June 11, 2020

Appendices

Table 1: Original Framingham Longitudinal Data (continued below)

RANDID	SEX	TOTCHOL	AGE	SYSBP	DIABP	CURSMOKE	CIGPDAY	BMI	DIABETES	BPMEDS	HEARTRTE	GLUCOSE
2448	1	195	39	106	70	0	0	26.97	0	0	80	77
2448	1	209	52	121	66	0	0	NA	0	0	69	92
6238	2	250	46	121	81	0	0	28.73	0	0	95	76
6238	2	260	52	105	69.5	0	0	29.43	0	0	80	86
6238	2	237	58	108	66	0	0	28.5	0	0	80	71
9428	1	245	48	127.5	80	1	20	25.34	0	0	75	70

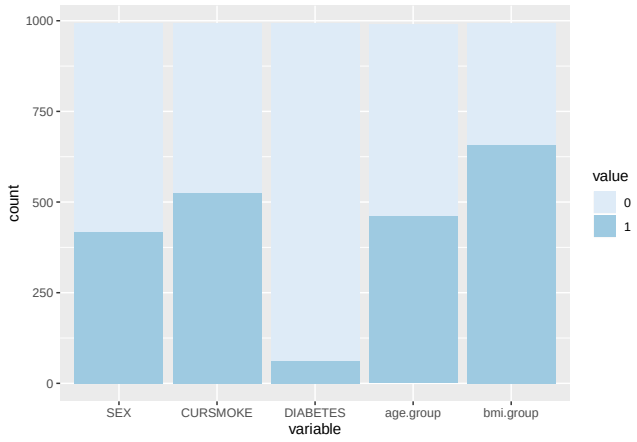
Table 2: Table continues below

educ	PREVCHD	PREVAP	PREVMI	PREVSTRK	PREVHYP	TIME	PERIOD	HDLC	LDLC	DEATH	ANGINA	HOSPMI
4	0	0	0	0	0	0	1	NA	NA	0	0	1
4	0	0	0	0	0	4628	3	31	178	0	0	1
2	0	0	0	0	0	0	1	NA	NA	0	0	0
2	0	0	0	0	0	2156	2	NA	NA	0	0	0
2	0	0	0	0	0	4344	3	54	141	0	0	0
1	0	0	0	0	0	0	1	NA	NA	0	0	0

MI_FCHD	ANYCHD	STROKE	CVD	HYPERTEN	TIMEAP	TIMEMI	TIMEMIFC	TIMECHD	TIMESTRK	TIMECVD	TIMEDTH	TIMEHYP
1	1	0	1	0	8766	6438	6438	6438	8766	6438	8766	8766
1	1	0	1	0	8766	6438	6438	6438	8766	6438	8766	8766
0	0	0	0	0	8766	8766	8766	8766	8766	8766	8766	8766
0	0	0	0	0	8766	8766	8766	8766	8766	8766	8766	8766
0	0	0	0	0	8766	8766	8766	8766	8766	8766	8766	8766
0	0	0	0	0	8766	8766	8766	8766	8766	8766	8766	8766

Table 4: The characteristics data are provided in the dataset

Variable	Description	Units	Range or count
RANDID	Unique identification number for each participant		2448-9999312
SEX	Participant sex	1=Men 2=Women	n=5022 n=6605
PERIOD	Examination Cycle	1=Period 1;2=Period 2; 3=Period 3	n=4434 n=3930 n=3263
TIME	Number of days since baseline exam		0-4854
AGE	Age at exam (years)		32-81
SYSBP	Systolic Blood Pressure (mean of last two of three measurements) (mmHg)		83.5-295
DIABP	Diastolic Blood Pressure (mean of last two of three measurements) (mmHg)		30-150
BPMEDS	Use of Anti-hypertensive medication at exam	0=Not currently used 1=Current Use	n=10090 n=944
CURSMOKE	Current cigarette smoking at exam	0=Not current smoker 1=Current smoker	n=6598 n=5029
CIGPDAY	Number of cigarettes smoked each day	0=Not current smoker 1-90 cigarettes per day	CIGPDAY
TOTCHOL	Serum Total Cholesterol (mg/dL)		107-696
HDLC	High Density Lipoprotein Cholesterol (mg/dL)	available for period 3 only	10-189
LDLC	Low Density Lipoprotein Cholesterol (mg/dL)	available for period 3 only	20-565
BMI	Body Mass Index, weight in kilograms/height meters squared		14.43-56.8
GLUCOSE	Casual serum glucose (mg/dL)		39-478
DIABETES	Diabetic according to criteria of first exam treated or first exam with casual glucose of 200 mg/dL or more	0=Not a diabetic 1=Diabetic	n=11097 n=530



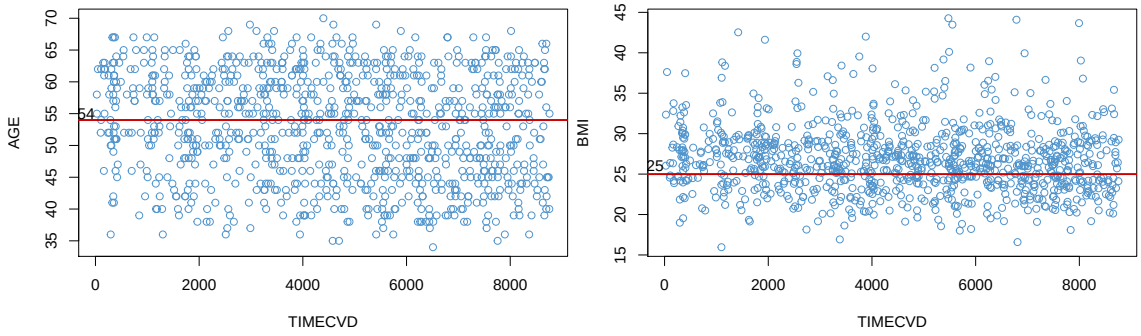
	0	1
SEX	575	417
CURSMOKE	468	524
DIABETES	930	62
age.group	529	463
bmi.group	335	657

Analysis of AGE and BMI

Based on the sample distribution and relevant definition, we choose 54 years old and 25 BMI as the values to classify.

```
## [1] "median of age = 54"
```

```
## [1] "median of BMI = 25.48"
```



- Status data

Table 6: The risk factor data

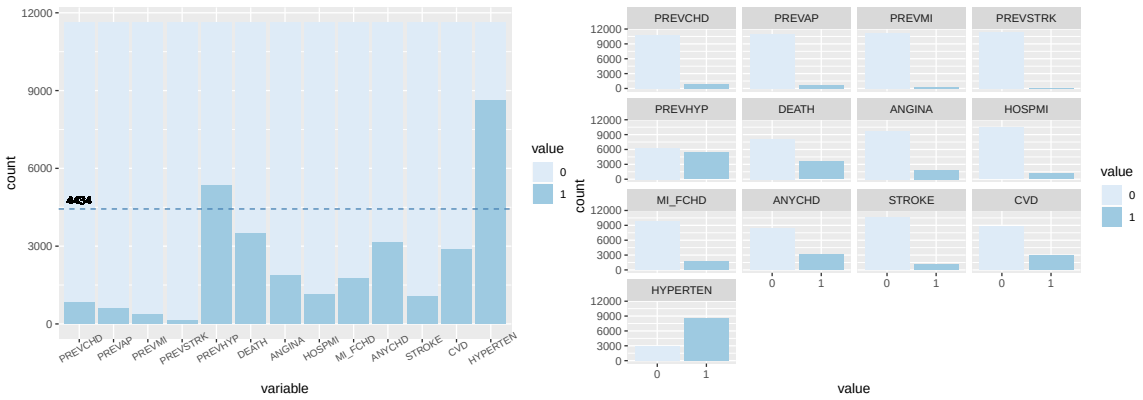
Variable	Description	Units	Range or count
HEARTRTE	Heart rate (Ventricular rate) in beats/min		37-220
PREVAP	Prevalent Angina Pectoris at exam	0=Free of disease 1=Prevalent disease	n=11000 n=627
PREVCHD	Prevalent Coronary Heart Disease defined as pre-existing Angina Pectoris, Myocardial Infarction (hospitalized, silent or unrecognized), or Coronary Insufficiency (unstable angina)	0=Free of disease 1=Prevalent disease	n=10785 n=842
PREVMI	Prevalent Myocardial Infarction	0=Free of disease 1=Prevalent disease	n=11253 n=374

Variable	Description	Units	Range or count
PREVSTRK	Prevalent Stroke	0=Free of disease 1=Prevalent disease	n=11475 n=152
PREVHYP	Prevalent Hypertensive. Subject was defined as hypertensive if treated or if second exam at which mean systolic was >=140 mmHg or mean Diastolic >=90 mmHg	0=Free of disease 1=Prevalent disease	n=6283 n=534

- Event selection

Table 7: The event data

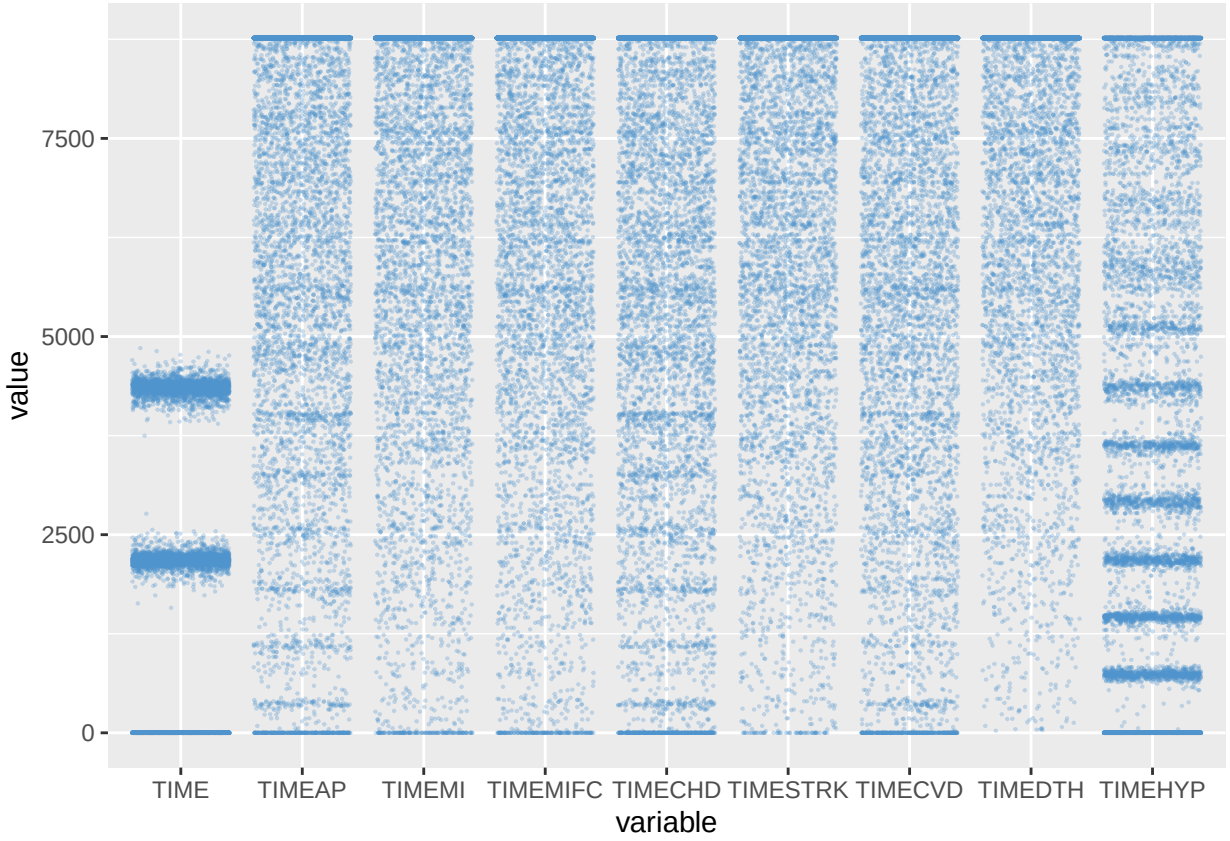
Variable	Description
ANGINA	Angina Pectoris
HOSPMI	Hospitalized Myocardial Infarction
MI_FCHD	Hospitalized Myocardial Infarction or Fatal Coronary Heart Disease
ANYCHD	Angina Pectoris, Myocardial infarction (Hospitalized and silent or unrecognized), Coronary Insufficiency (Unstable Angina), or Fatal Coronary Heart Disease
STROKE	Atherothrombotic infarction, Cerebral Embolism, Intracerebral Hemorrhage, or Subarachnoid Hemorrhage or Fatal Cerebrovascular Disease
CVD	Myocardial infarction (Hospitalized and silent or unrecognized), Fatal Coronary Heart Disease, Atherothrombotic infarction, Cerebral Embolism, Intracerebral Hemorrhage, or Subarachnoid Hemorrhage or Fatal Cerebrovascular Disease
HYPERTEN	Hypertensive. Defined as the first exam treated for high blood pressure or second exam in which either Systolic is 140 mmHg or Diastolic 90mmHg
DEATH	Death from any cause

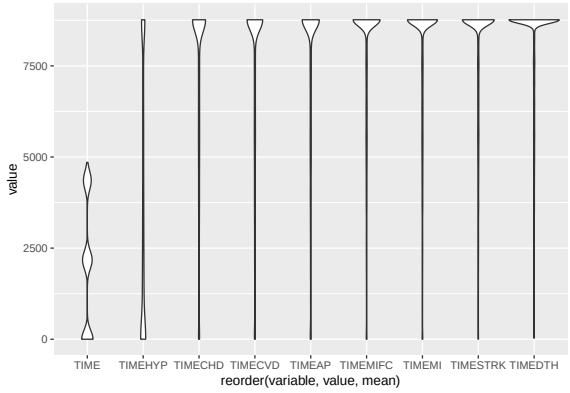


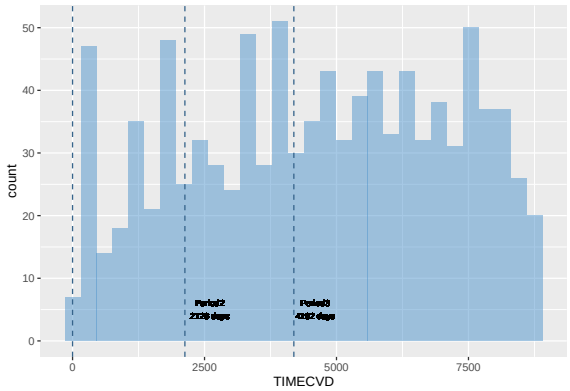
- Event time

Table 8: The event time data

Variable	Description
TIMEAP	Number of days from Baseline exam to first Angina during the followup or Number of days from Baseline to censor date. Censor date may be end of followup, death or last known contact date if subject is lost to followup
TIMEMI	Defined as above for the first HOSPMI event during followup
TIMEMIFC	Defined as above for the first MI_FCHD event during followup
TIMECHD	Defined as above for the first ANYCHD event during followup
TIMESTRK	Defined as above for the first STROKE event during followup
TIMECVD	Defined as above for the first CVD event during followup
TIMEHYP	Defined as above for the first HYPERTEN event during followup
TIMEDTH	Number of days from Baseline exam to death if occurring during followup or Number of days from Baseline to censor date. Censor date may be end of followup, or last known contact date if subject is lost to followup



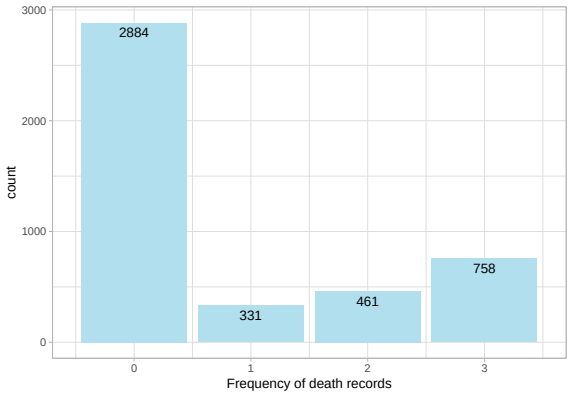




- The censored observations

In this data, 'death' represents non-censoring. Both of the number of "DEATH=1" and "TIMEDTH<8766" are 3527. The number of records of alive is 8100. However, there are more than one 'DEATH' records for a patient. The numbers of death records by period are

```
##        DEATH
## PERIOD    0    1
##      1 2884 1550
##      2 2728 1202
##      3 2488  775
```



Frequency of death records for a patient by period

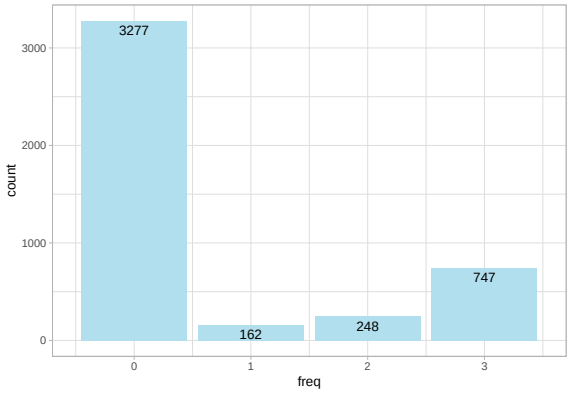
Once in per1	twice in per1	twice in per2	twice in per3	3 times in per1	3 times in per2	3 times in per3
331	461	444	17	758	758	758

The real number of death is $331 + 461 + 758 = 1550$. We can confirm the number of death records by $331 + 461 * 2 + 758 * 3 = 3527$

- Investigate CVD event

The number of records of CVD is 2899. The numbers of CVD records by period are

```
##        CVD
## PERIOD    0    1
##      1 3277 1157
##      2 2953  977
##      3 2498  765
```



Frequency of CVD records for a patient by period

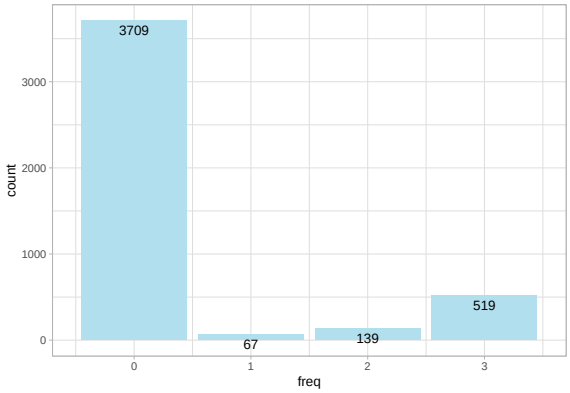
Once in per1	twice in per1	twice in per2	twice in per3	3 times in per1	3 times in per2	3 times in per3
162	248	230	18	747	747	747

The real number of CVD is $162 + 248 + 747 - 161 = 996$ (There are 161 observation have zero CVD time). We can confirm the number of death records by $162 + 248 * 2 + 747 * 3 = 2899$

- Investigate ANGINA event

The number of records of ANGINA is 1902. The numbers of ANGINA records by period are

```
##       ANGINA
## PERIOD    0    1
##      1 3709  725
##      2 3281  649
##      3 2735  528
```



Once in per1	twice in per1	twice in per2	twice in per3	3 times in per1	3 times in per2	3 times in per3
67	139	130	9	519	519	519

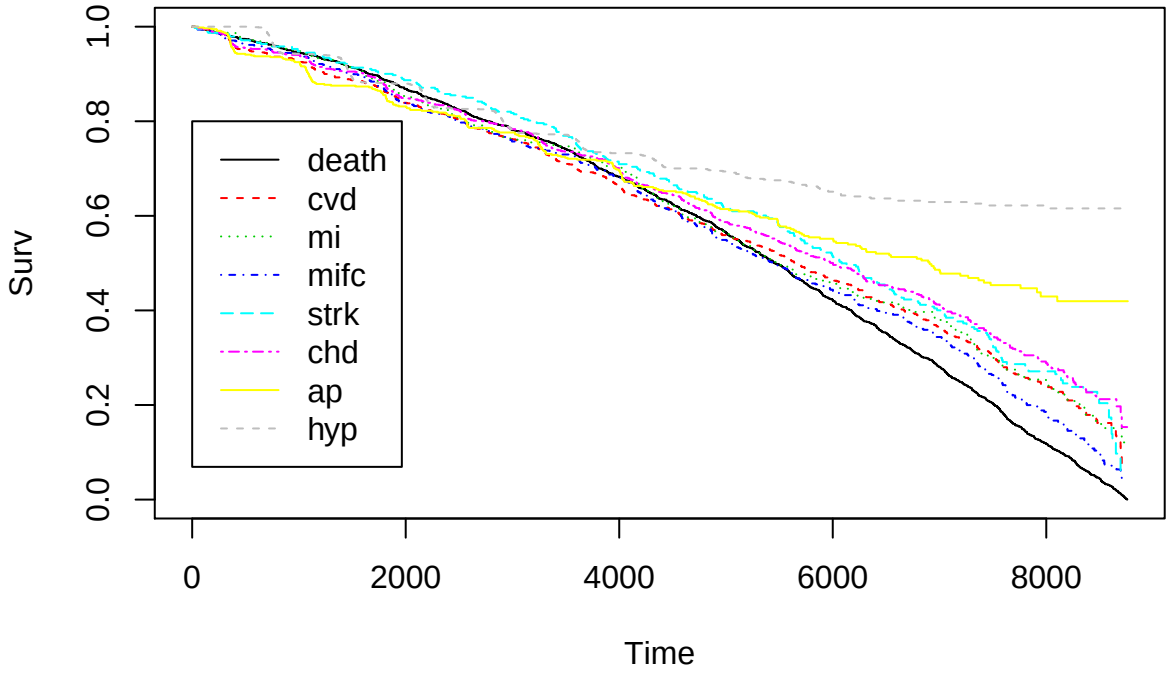
The real number of CVD is $67 + 139 + 519 = 725$. We can confirm the number of ANGINA records by $67 + 139 * 2 + 519 * 3 = 1902$

The calculation show that both CVD an ANGINA events contain the whole event number in the first period. So, we can filter the data by period 1 to get the unique observations for each patient.

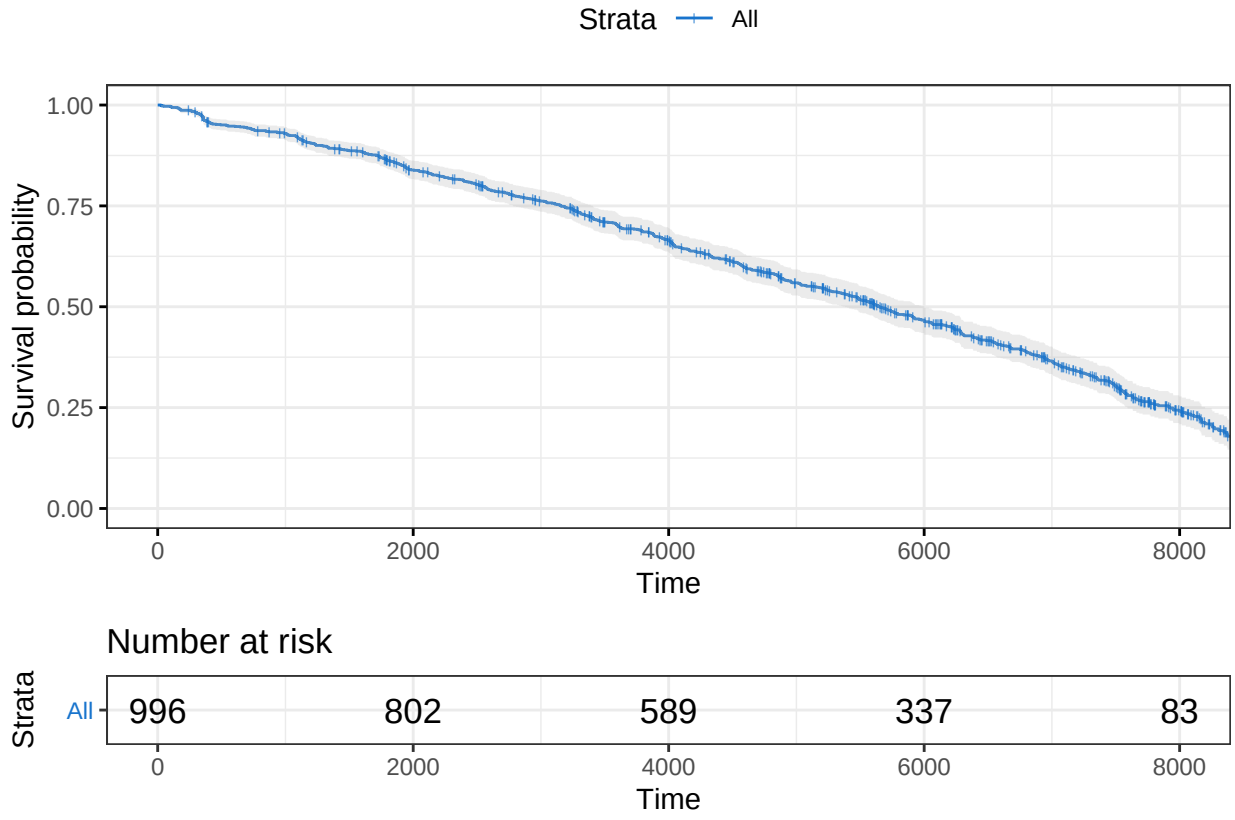
The total number of CVD is 1157. There are 161 patient who dead at once have zero TIMECVD value. There are 996 patients have a positive TIMECVD value.

Non-parameter: Kaplan-Meier Method

- Comparing the 8 kinds of events by Kaplan-Meier (K-M) estimate



- Overall K-M curve of CVD patients

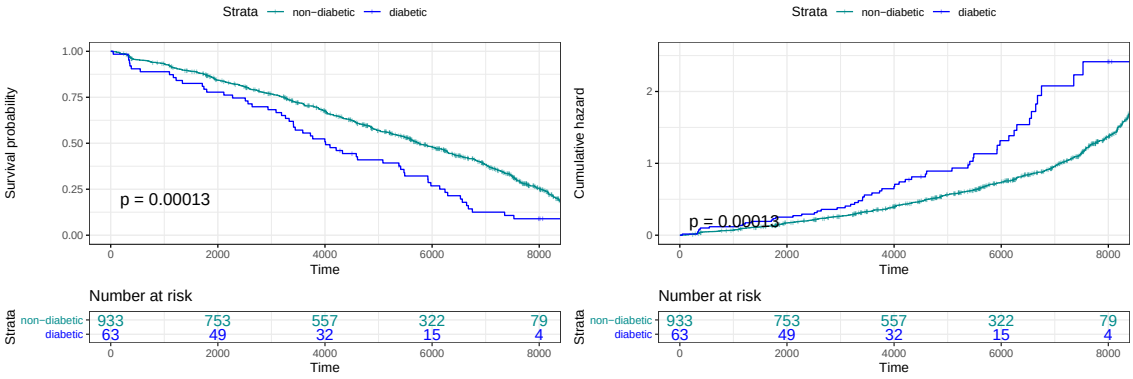


The mean and median survival times.

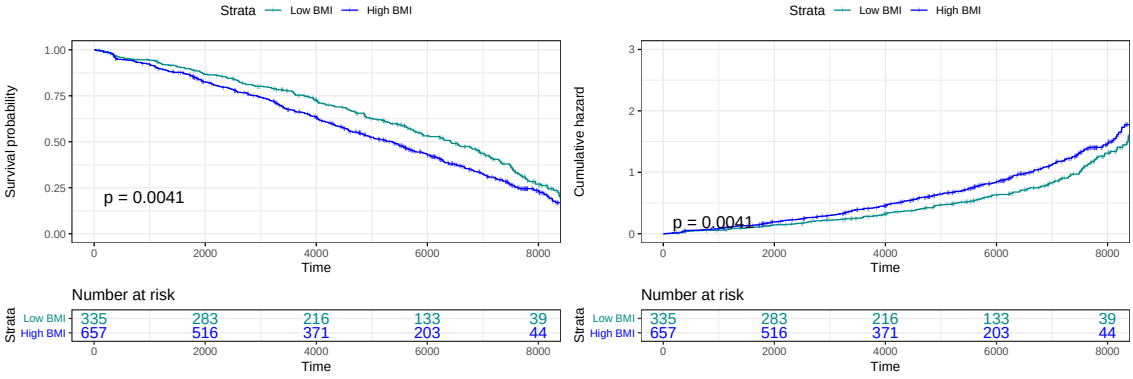
```
## Call: survfit(formula = Surv(TIMECVD, DEATH) ~ 1, data = cvd, type = "kaplan-meier")
##
```

```
##          n     events      *rmean *se(rmean)     median    0.95LCL    0.95UCL
##      996.0      635.0     5339.2       89.9     5645.0     5380.0     6001.0
##     * restricted mean with upper limit =  8758
```

- the K-M survival curve and cumulative hazard for a specific group.




```
## Call:
## survdiff(formula = Surv(TIMECVD, DEATH) ~ DIABETES, data = cvd)
##
##                N Observed Expected (O-E)^2/E (O-E)^2/V
## DIABETES=0 933      580    601.6     0.772      14.7
## DIABETES=1  63       55     33.4    13.887      14.7
##
##  Chisq= 14.7  on 1 degrees of freedom, p= 0.0001
```




```
## Call:
## survdiff(formula = Surv(TIMECVD, DEATH) ~ bmi.group, data = cvd)
##
## n=992, 4 observations deleted due to missingness.
##
##                N Observed Expected (O-E)^2/E (O-E)^2/V
## bmi.group=0 335      207      242      5.04      8.22
## bmi.group=1 657      425      390      3.12      8.22
##
##  Chisq= 8.2  on 1 degrees of freedom, p= 0.004
```

- Stratify by two factors.

```
##         cvd.CURSMOKE
## cvd.SEX   0   1
##       0 222 357
##       1 247 170
```

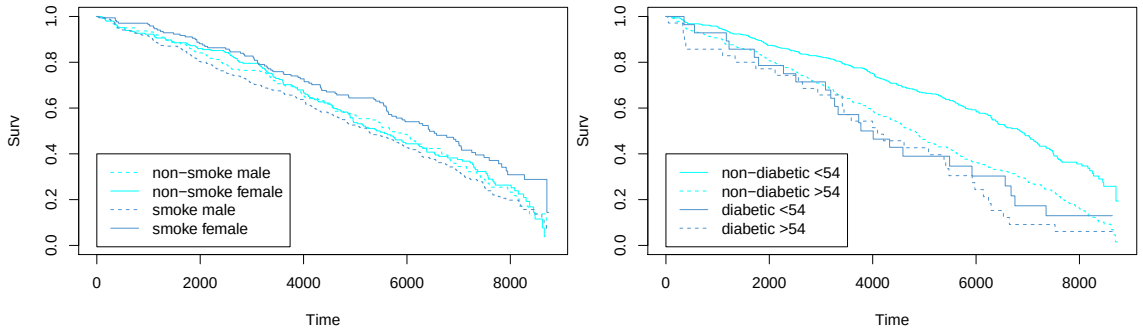
```
## Call:
## survdiff(formula = Surv(TIMECVD, DEATH) ~ CURSMOKE + strata(SEX),
##     data = cvd)
##
##              N Observed Expected (O-E)^2/E (O-E)^2/V
## CURSMOKE=0 469      302      293     0.263     0.501
## CURSMOKE=1 527      333      342     0.225     0.501
##
##  Chisq= 0.5  on 1 degrees of freedom, p= 0.5
```

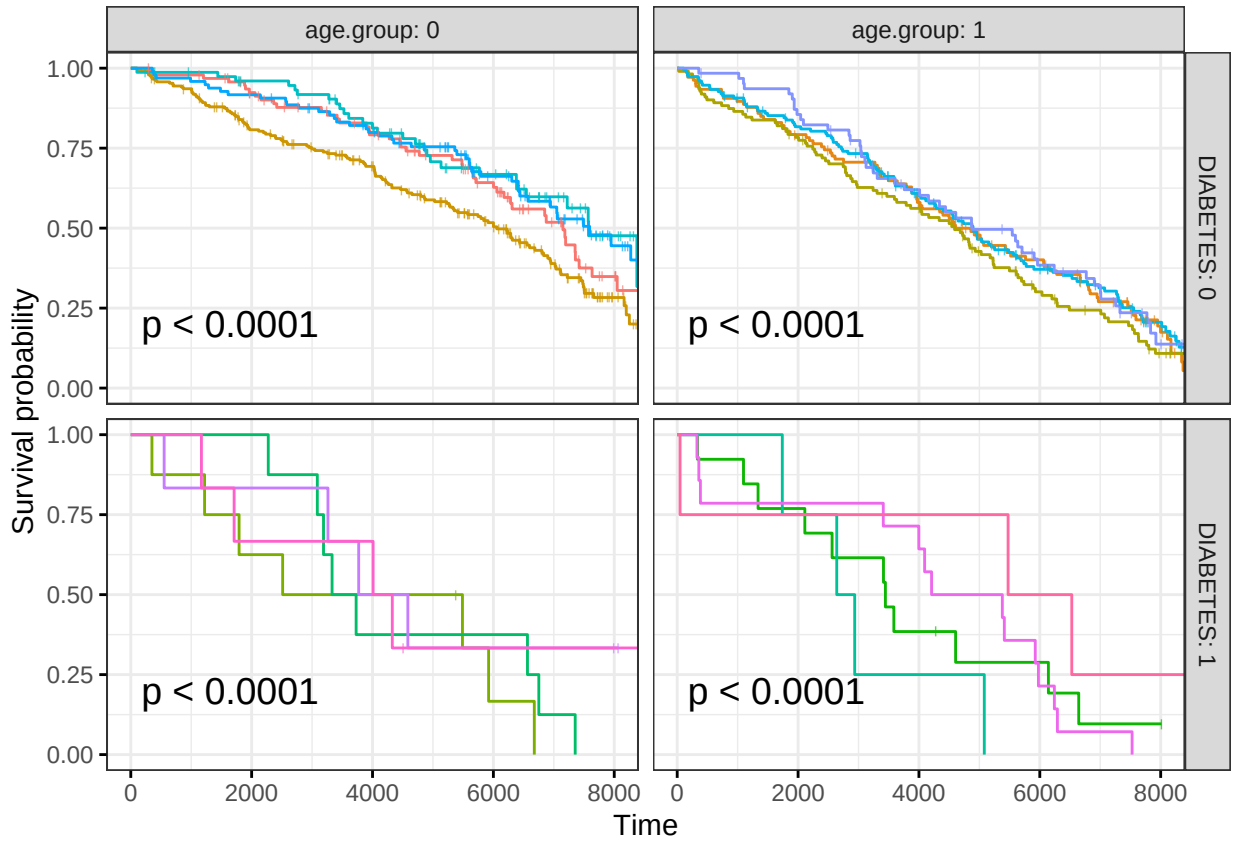
```
## Call: survfit(formula = Surv(TIMECVD, DEATH) ~ CURSMOKE + strata(SEX),
##     data = cvd)
##
##                              n events *rmean *se(rmean) median 0.95LCL 0.95UCL
## CURSMOKE=0, strata(SEX)=0 222    146   5324        187   5710    5036    6644
## CURSMOKE=0, strata(SEX)=1 247    156   5363        178   5413    4902    6291
## CURSMOKE=1, strata(SEX)=0 357    237   5044        152   5243    4750    5912
## CURSMOKE=1, strata(SEX)=1 170     96   5902        211   6437    5645    7254
##     * restricted mean with upper limit =  8719
```

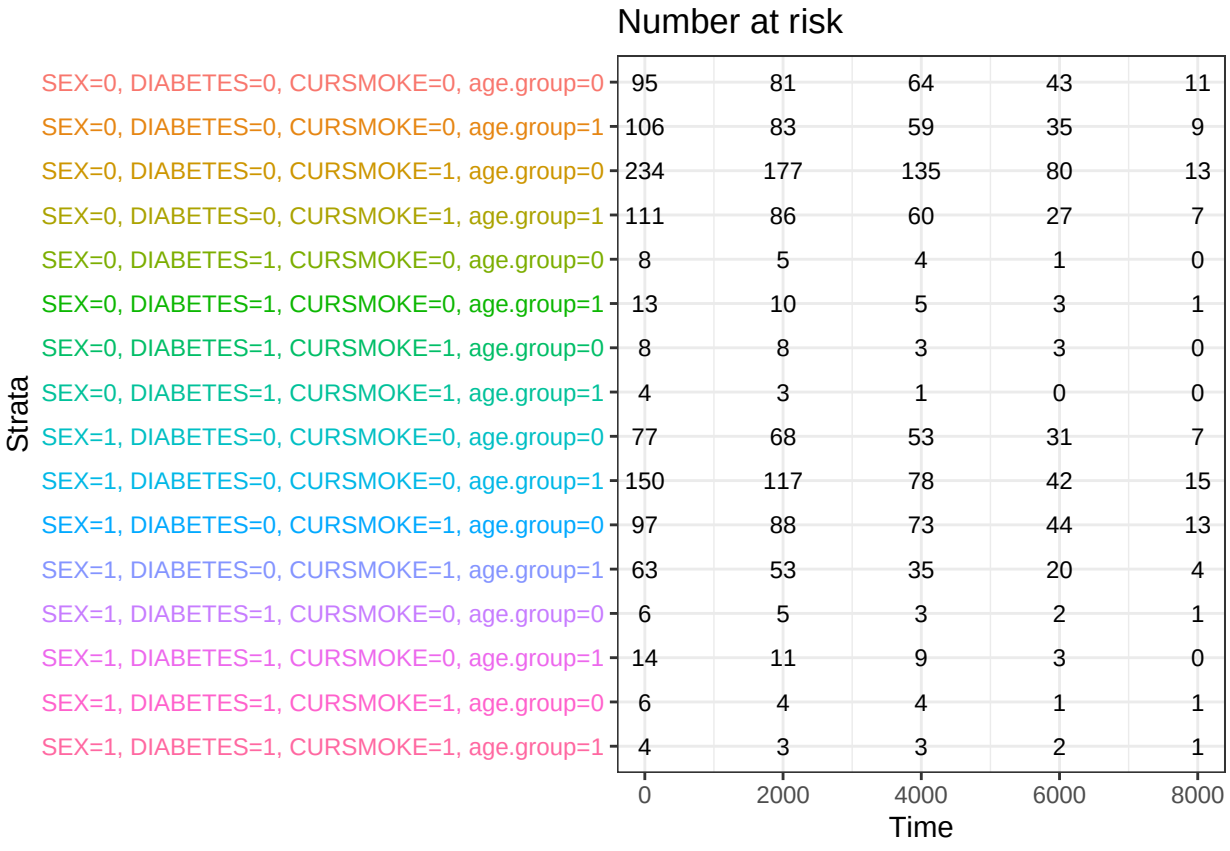
```
##              cvd.DIABETES
## cvd.age.group   0   1
##             0 503  28
##             1 430  35
```

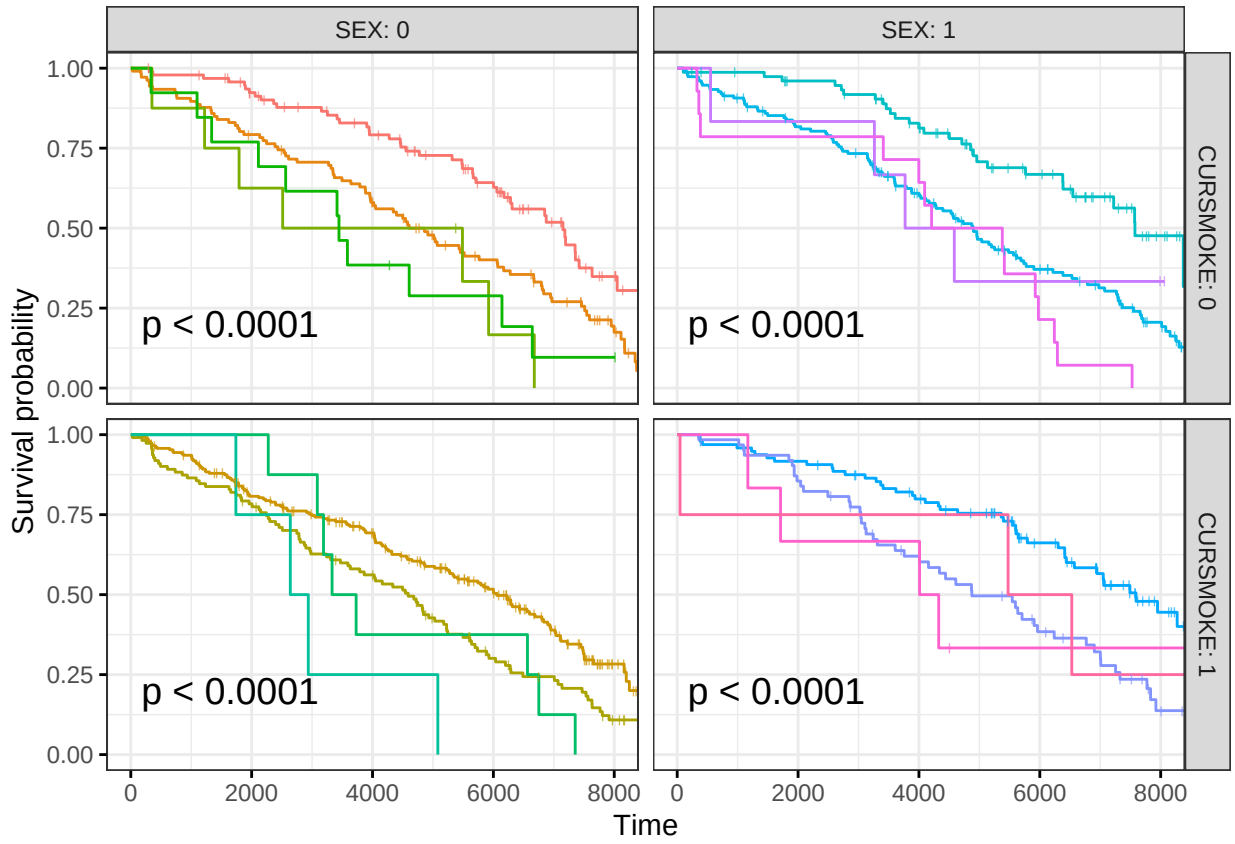
```
## Call:
## survdiff(formula = Surv(TIMECVD, DEATH) ~ DIABETES + strata(age.group),
##     data = cvd)
##
##              N Observed Expected (O-E)^2/E (O-E)^2/V
## DIABETES=0 933      580    599.3     0.623      11.2
## DIABETES=1  63       55     35.7    10.474      11.2
##
##  Chisq= 11.2  on 1 degrees of freedom, p= 0.0008
```

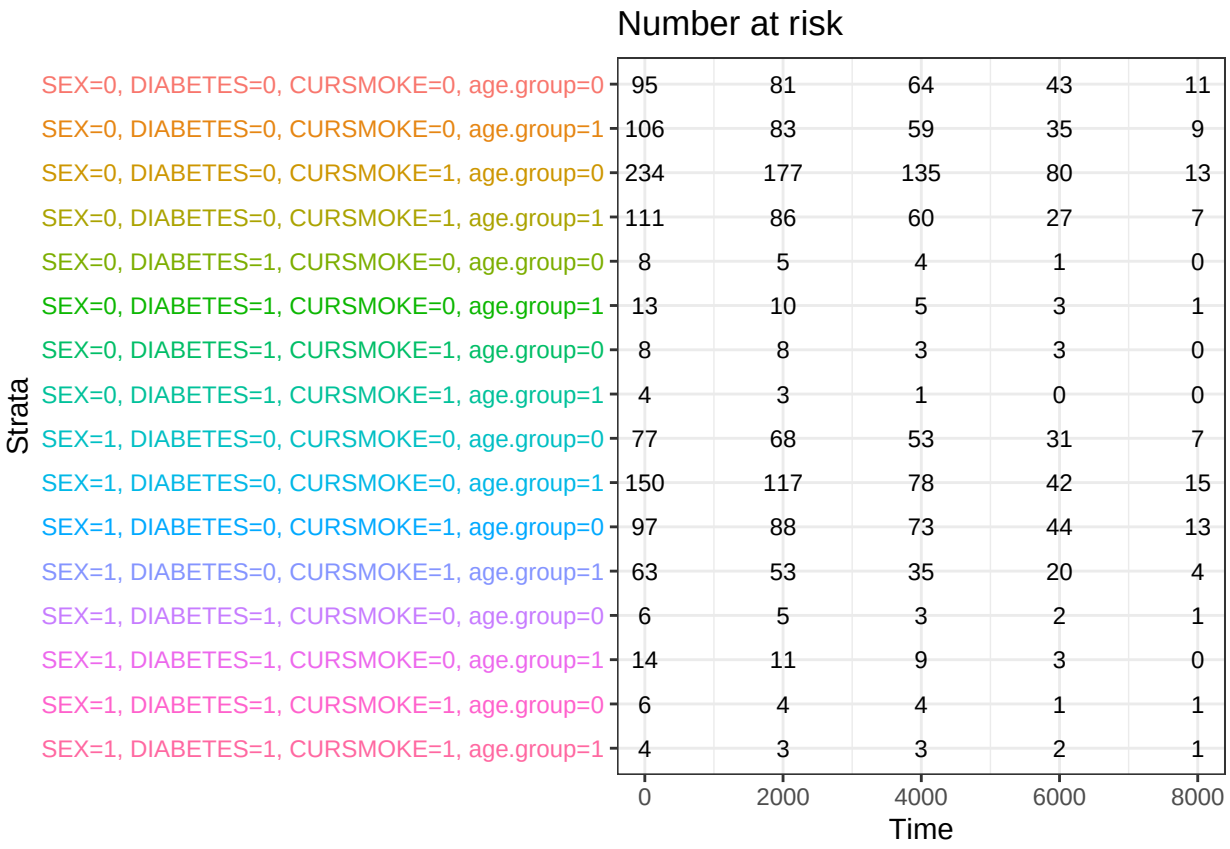
```
## Call: survfit(formula = Surv(TIMECVD, DEATH) ~ DIABETES + strata(age.group),
##     data = cvd)
##
##                                    n events *rmean *se(rmean) median 0.95LCL
## DIABETES=0, strata(age.group)=0 503    248   6016        127   6849    6380
## DIABETES=0, strata(age.group)=1 430    332   4784        132   4822    4397
## DIABETES=1, strata(age.group)=0  28     23   4427        486   3890    3189
## DIABETES=1, strata(age.group)=1  35     32   4086        412   4094    3410
##                                  0.95UCL
## DIABETES=0, strata(age.group)=0    7195
## DIABETES=0, strata(age.group)=1    5157
## DIABETES=1, strata(age.group)=0    6677
## DIABETES=1, strata(age.group)=1    5925
##     * restricted mean with upper limit =  8712
```




age.group: 0
age.group: 1
DIABETES: 0
DIABETES: 1
Survival probability
1.00
0.75
0.50
0.25
0.00
p < 0.0001
0
2000
4000
6000
8000
Time

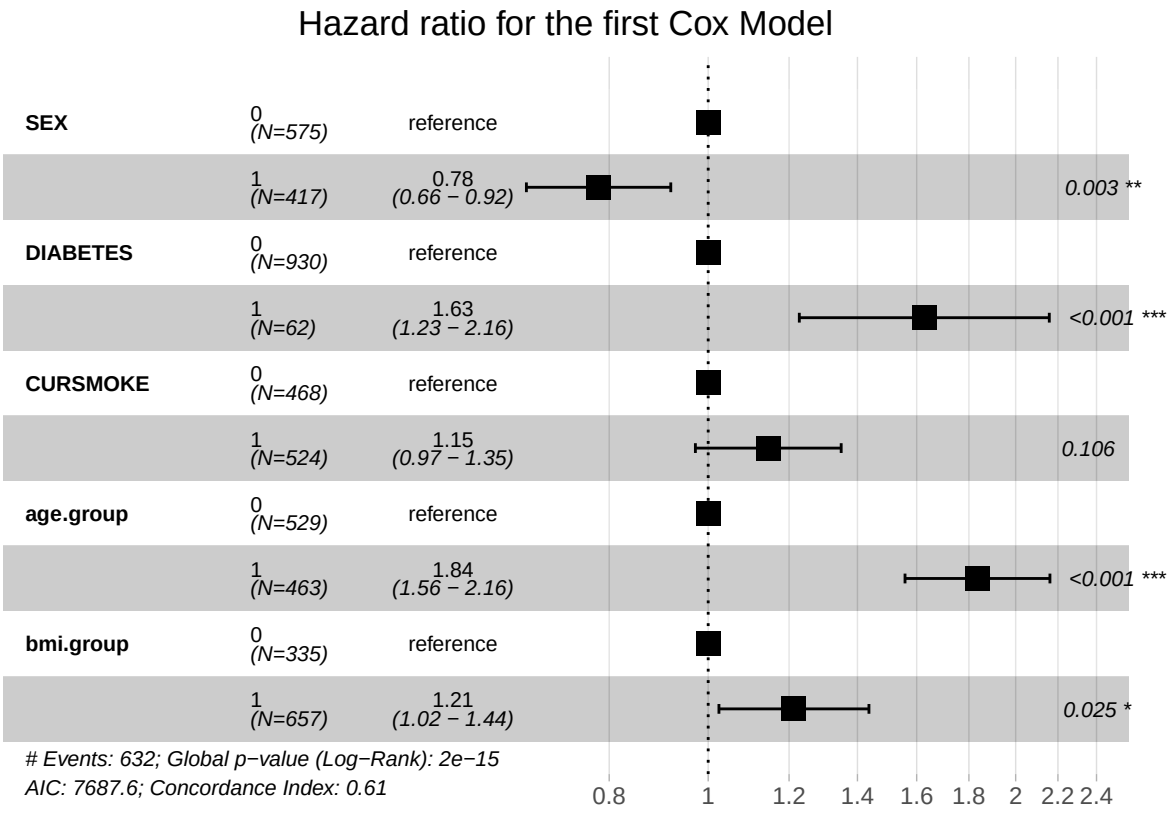

Number at risk
Strata
SEX=0, DIABETES=0, CURSMOKE=0, age.group=0 95 81 64 43 11
SEX=0, DIABETES=0, CURSMOKE=0, age.group=1 106 83 59 35 9
SEX=0, DIABETES=0, CURSMOKE=1, age.group=0 234 177 135 80 13
SEX=0, DIABETES=0, CURSMOKE=1, age.group=1 111 86 60 27 7
SEX=0, DIABETES=1, CURSMOKE=0, age.group=0 8 5 4 1 0
SEX=0, DIABETES=1, CURSMOKE=0, age.group=1 13 10 5 3 1
SEX=0, DIABETES=1, CURSMOKE=1, age.group=0 8 8 3 3 0
SEX=0, DIABETES=1, CURSMOKE=1, age.group=1 4 3 1 0 0
SEX=1, DIABETES=0, CURSMOKE=0, age.group=0 77 68 53 31 7
SEX=1, DIABETES=0, CURSMOKE=0, age.group=1 150 117 78 42 15
SEX=1, DIABETES=0, CURSMOKE=1, age.group=0 97 88 73 44 13
SEX=1, DIABETES=0, CURSMOKE=1, age.group=1 63 53 35 20 4
SEX=1, DIABETES=1, CURSMOKE=0, age.group=0 6 5 3 2 1
SEX=1, DIABETES=1, CURSMOKE=0, age.group=1 14 11 9 3 0
SEX=1, DIABETES=1, CURSMOKE=1, age.group=0 6 4 4 1 1
SEX=1, DIABETES=1, CURSMOKE=1, age.group=1 4 3 3 2 1
0
2000
4000
6000
8000
Time


SEX: 0
SEX: 1
CURSMOKE: 0
CURSMOKE: 1
Survival probability
1.00
0.75
0.50
0.25
0.00
p < 0.0001
0
2000
4000
6000
8000
Time


Number at risk
Strata
SEX=0, DIABETES=0, CURSMOKE=0, age.group=0 95 81 64 43 11
SEX=0, DIABETES=0, CURSMOKE=0, age.group=1 106 83 59 35 9
SEX=0, DIABETES=0, CURSMOKE=1, age.group=0 234 177 135 80 13
SEX=0, DIABETES=0, CURSMOKE=1, age.group=1 111 86 60 27 7
SEX=0, DIABETES=1, CURSMOKE=0, age.group=0 8 5 4 1 0
SEX=0, DIABETES=1, CURSMOKE=0, age.group=1 13 10 5 3 1
SEX=0, DIABETES=1, CURSMOKE=1, age.group=0 8 8 3 3 0
SEX=0, DIABETES=1, CURSMOKE=1, age.group=1 4 3 1 0 0
SEX=1, DIABETES=0, CURSMOKE=0, age.group=0 77 68 53 31 7
SEX=1, DIABETES=0, CURSMOKE=0, age.group=1 150 117 78 42 15
SEX=1, DIABETES=0, CURSMOKE=1, age.group=0 97 88 73 44 13
SEX=1, DIABETES=0, CURSMOKE=1, age.group=1 63 53 35 20 4
SEX=1, DIABETES=1, CURSMOKE=0, age.group=0 6 5 3 2 1
SEX=1, DIABETES=1, CURSMOKE=0, age.group=1 14 11 9 3 0
SEX=1, DIABETES=1, CURSMOKE=1, age.group=0 6 4 4 1 1
SEX=1, DIABETES=1, CURSMOKE=1, age.group=1 4 3 3 2 1
0
2000
4000
6000
8000
Time

Variable Reduction: The Cox Proportional Hazards Model

```
## Call:
## coxph(formula = Surv(TIMECVD, DEATH) ~ SEX + DIABETES + CURSMOKE +
##     age.group + bmi.group, data = cvd)
##
##   n= 992, number of events= 632
##
##                coef exp(coef) se(coef)     z Pr(>|z|)
## SEX1       -0.2471    0.7811   0.0830 -2.98   0.0029 **
## DIABETES1   0.4872    1.6277   0.1438  3.39   0.0007 ***
## CURSMOKE1   0.1355    1.1451   0.0838  1.62   0.1062
## age.group1  0.6071    1.8350   0.0833  7.28  3.3e-13 ***
## bmi.group1  0.1933    1.2133   0.0863  2.24   0.0251 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
##            exp(coef) exp(-coef) lower .95 upper .95
## SEX1           0.781      1.280     0.664     0.919
## DIABETES1      1.628      0.614     1.228     2.158
## CURSMOKE1      1.145      0.873     0.972     1.350
## age.group1     1.835      0.545     1.558     2.161
## bmi.group1     1.213      0.824     1.025     1.437
##
## Concordance= 0.606  (se = 0.012 )
## Likelihood ratio test= 78.2  on 5 df,   p=2e-15
## Wald test            = 78.5  on 5 df,   p=2e-15
## Score (logrank) test = 80.3  on 5 df,   p=7e-16
```



- Simplify (reduce) the model using AIC method

```
## Stepwise Model Path
## Analysis of Deviance Table
##
```

```
## Initial Model:
## Surv(TIMECVD, DEATH) ~ SEX + DIABETES + CURSMOKE + age.group +
##     bmi.group
##
## Final Model:
## Surv(TIMECVD, DEATH) ~ SEX + DIABETES + CURSMOKE + age.group +
##     bmi.group + SEX:CURSMOKE + DIABETES:age.group
##
##
##                   Step Df Deviance Resid. Df Resid. Dev  AIC
## 1                                        627       7678 7688
## 2       + SEX:CURSMOKE  1    4.583       626       7673 7685
## 3 + DIABETES:age.group  1    2.195       625       7671 7685
```

The stepwise method evaluated all the two-way interaction effects. The reduced model retains two significant interaction effects.

```
## Call:
## coxph(formula = Surv(TIMECVD, DEATH) ~ SEX + DIABETES + CURSMOKE +
##     age.group + bmi.group + SEX:CURSMOKE + DIABETES:age.group,
##     data = cvd)
##
##   n= 992, number of events= 632
##
##                          coef exp(coef) se(coef)     z Pr(>|z|)
## SEX1                  -0.0763    0.9266   0.1158 -0.66  0.51019
## DIABETES1              0.7674    2.1542   0.2185  3.51  0.00045 ***
## CURSMOKE1              0.2739    1.3151   0.1078  2.54  0.01104 *
## age.group1             0.6379    1.8925   0.0869  7.34  2.1e-13 ***
## bmi.group1             0.1889    1.2079   0.0863  2.19  0.02864 *
## SEX1:CURSMOKE1        -0.3501    0.7046   0.1678 -2.09  0.03699 *
## DIABETES1:age.group1 -0.4319    0.6493   0.2882 -1.50  0.13395
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
##                      exp(coef) exp(-coef) lower .95 upper .95
## SEX1                     0.927      1.079     0.738     1.163
## DIABETES1                2.154      0.464     1.404     3.306
## CURSMOKE1                1.315      0.760     1.065     1.624
## age.group1               1.893      0.528     1.596     2.244
## bmi.group1               1.208      0.828     1.020     1.430
## SEX1:CURSMOKE1           0.705      1.419     0.507     0.979
## DIABETES1:age.group1     0.649      1.540     0.369     1.142
##
## Concordance= 0.607  (se = 0.012 )
## Likelihood ratio test= 85  on 7 df,   p=1e-15
## Wald test            = 82.5  on 7 df,   p=4e-15
## Score (logrank) test = 84.9  on 7 df,   p=1e-15
```

- Step II: LRT Further Reduce

```
## Call:
## coxph(formula = Surv(TIMECVD, DEATH) ~ SEX + DIABETES + CURSMOKE +
##     age.group + bmi.group + SEX:CURSMOKE, data = cvd)
##
##   n= 992, number of events= 632
##
##                     coef exp(coef) se(coef)     z Pr(>|z|)
## SEX1             -0.0753    0.9275   0.1158 -0.65  0.51563
## DIABETES1         0.5020    1.6521   0.1440  3.49  0.00049 ***
## CURSMOKE1         0.2767    1.3188   0.1079  2.57  0.01030 *
## age.group1        0.6029    1.8274   0.0835  7.22  5.2e-13 ***
## bmi.group1        0.1848    1.2030   0.0862  2.14  0.03208 *
## SEX1:CURSMOKE1 -0.3569    0.6998   0.1678 -2.13  0.03345 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
```

```
##                exp(coef) exp(-coef) lower .95 upper .95
## SEX1               0.927      1.078     0.739     1.164
## DIABETES1          1.652      0.605     1.246     2.191
## CURSMOKE1          1.319      0.758     1.068     1.629
## age.group1         1.827      0.547     1.551     2.152
## bmi.group1         1.203      0.831     1.016     1.425
## SEX1:CURSMOKE1     0.700      1.429     0.504     0.972
##
## Concordance= 0.607  (se = 0.012 )
## Likelihood ratio test= 82.8  on 6 df,   p=1e-15
## Wald test            = 82  on 6 df,   p=1e-15
## Score (logrank) test = 84  on 6 df,   p=5e-16


## [1] 0.1385
```

When we try to remove interaction between diabetes and age, the LRT tests show no evidence against the models without interaction effect of DIABETES and age.group (p-value = 0.1385).

Thus, we choose the model as $Time = SEX + CURSMOKE + DIABETES + age.group + bmi.group + SEX : CURSMOKE$

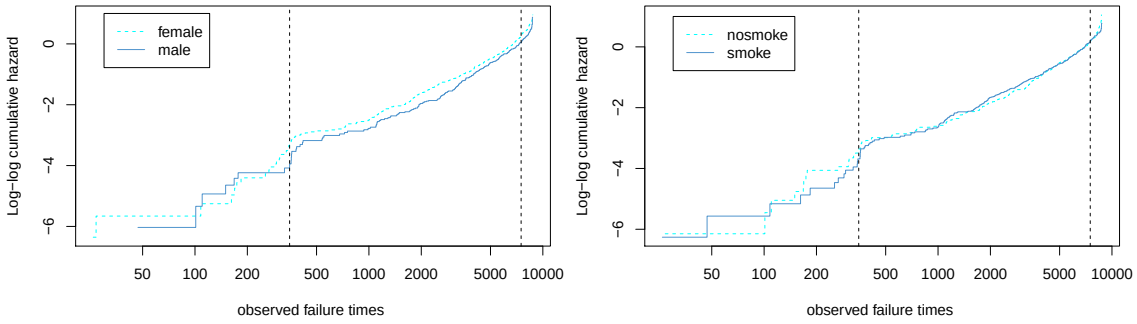
Bothstep elimination gives the same result.

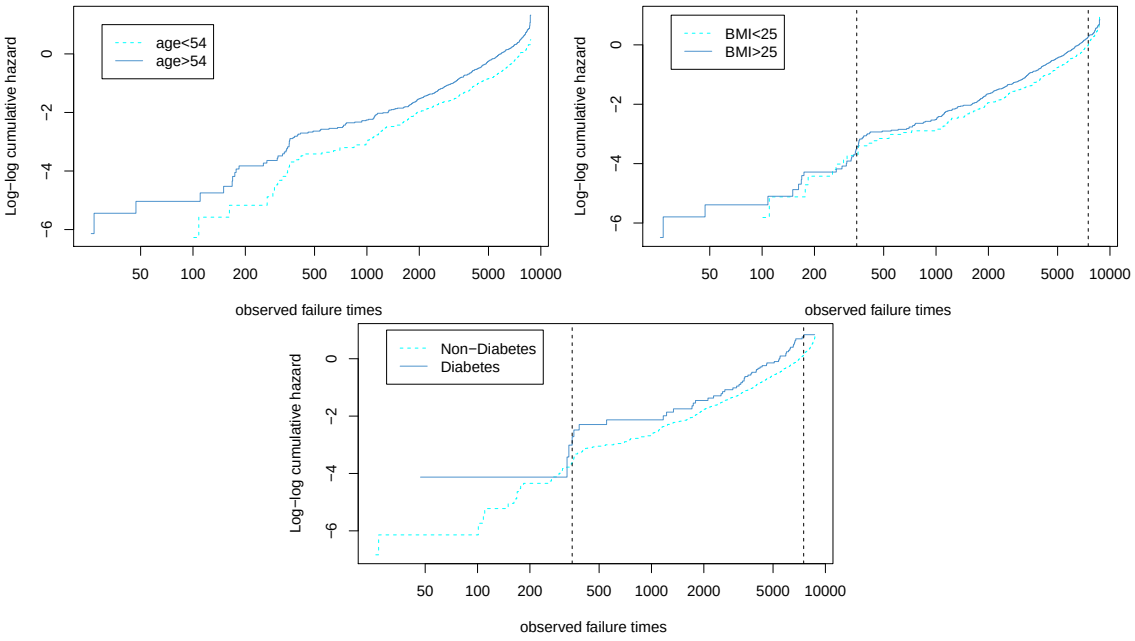
```
## Stepwise Model Path
## Analysis of Deviance Table
##
## Initial Model:
## Surv(TIMECVD, DEATH) ~ SEX + DIABETES + CURSMOKE + age.group +
##     bmi.group
##
## Final Model:
## Surv(TIMECVD, DEATH) ~ SEX + DIABETES + CURSMOKE + age.group +
##     bmi.group + SEX:CURSMOKE + DIABETES:age.group
##
##
##                    Step Df Deviance Resid. Df Resid. Dev  AIC
## 1                                         627       7678 7688
## 2       + SEX:CURSMOKE  1    4.583       626       7673 7685
## 3 + DIABETES:age.group  1    2.195       625       7671 7685
```

- Stratified Cox PH regression

If the log of cumulative hazard functions from different strata are parallel, this supports the Cox PH model without stratification with 'age.group'.

The period from 350 to 7500 days show some parallel patterns with gender, BMI, and Diabetes.





Another form of the final model is $Time = SEX + strata(CURSMOKE) + DIABETES + age.group + bmi.group$.

```
cvd.cox4 <- coxph(Surv(TIMECVD,DEATH) ~ SEX+strata(CURSMOKE)+DIABETES+age.group+bmi.group, data=cvd)
(cvd.cox4)
```

```
## Call:
## coxph(formula = Surv(TIMECVD, DEATH) ~ SEX + strata(CURSMOKE) +
##     DIABETES + age.group + bmi.group, data = cvd)
##
##              coef exp(coef) se(coef)  z     p
## SEX1       -0.24      0.79     0.08 -3 0.004
## DIABETES1   0.51      1.66     0.14  4 4e-04
## age.group1  0.61      1.83     0.08  7 3e-13
## bmi.group1  0.19      1.21     0.09  2 0.027
##
## Likelihood ratio test=78  on 4 df, p=5e-16
## n= 992, number of events= 632
```

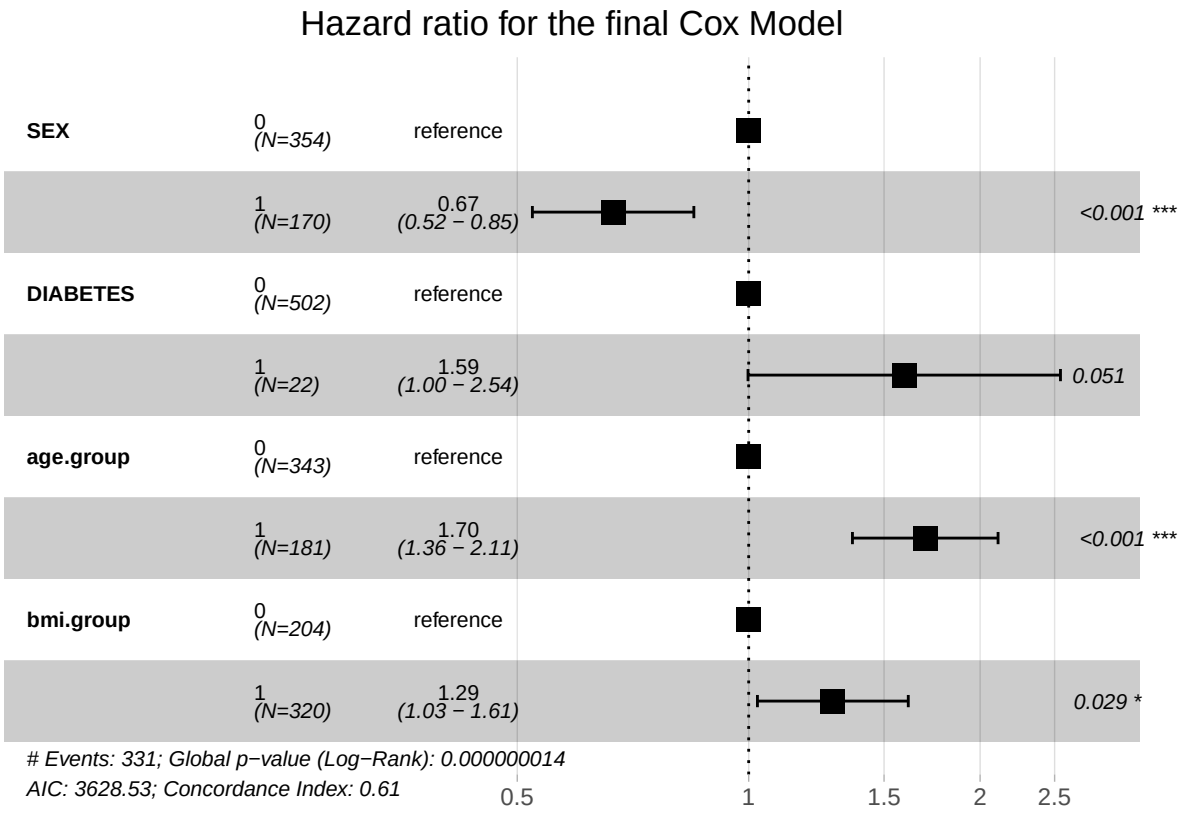
```
cvd.cox5 <- coxph(Surv(TIMECVD,DEATH) ~ SEX+DIABETES+age.group+bmi.group, data=cvd[cvd$CURSMOKE==1,])
(cvd.cox5)
```

```
## Call:
## coxph(formula = Surv(TIMECVD, DEATH) ~ SEX + DIABETES + age.group +
##     bmi.group, data = cvd[cvd$CURSMOKE == 1, ])
##
##             coef exp(coef) se(coef)  z        p
## SEX1       -0.4       0.7      0.1 -3 0.001000
## DIABETES1   0.5       1.6      0.2  2     0.05
## age.group1  0.5       1.7      0.1  5 0.000002
## bmi.group1  0.3       1.3      0.1  2     0.03
##
## Likelihood ratio test=42  on 4 df, p=0.00000001
## n= 524, number of events= 331
```

```
cvd.cox6 <- coxph(Surv(TIMECVD,DEATH) ~ DIABETES+age.group+bmi.group, data=cvd[cvd$CURSMOKE==1&cvd$SEX==1,])
(cvd.cox6)
```

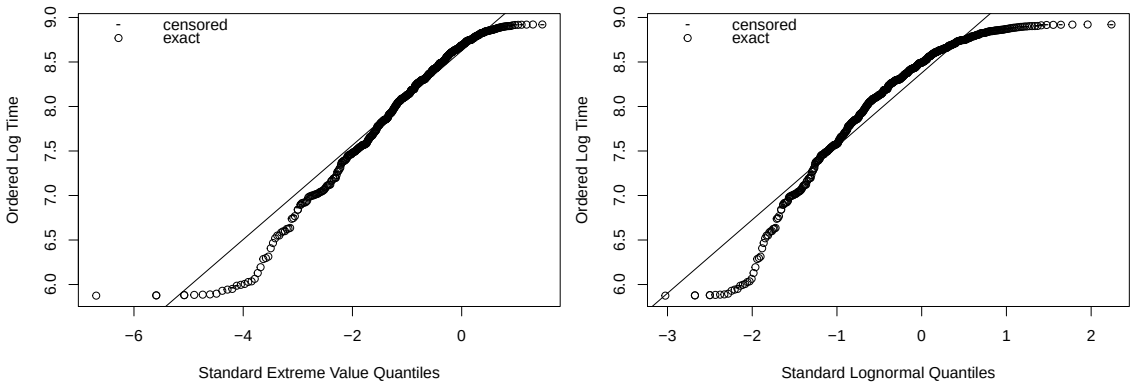
```
## Call:
```

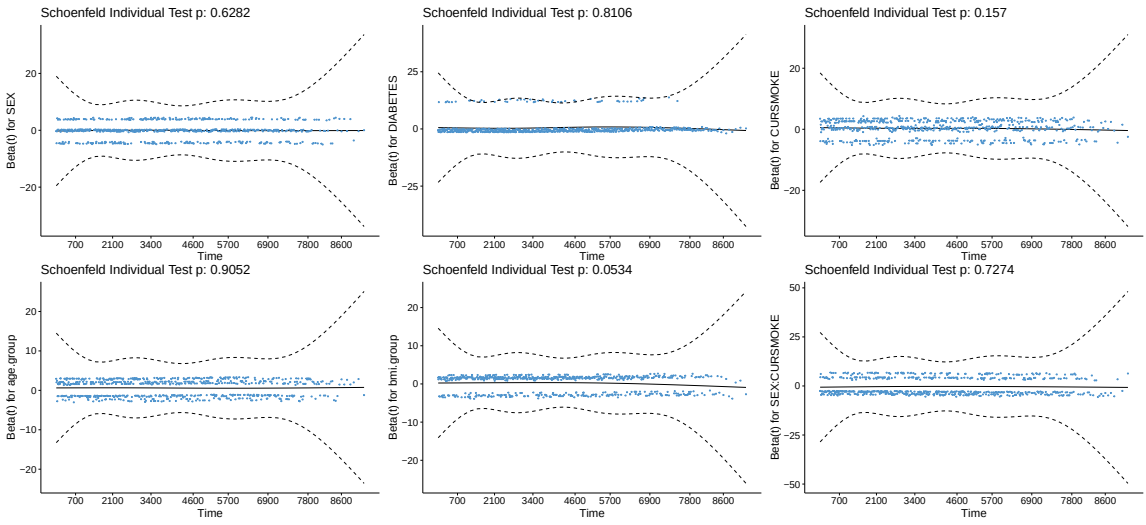
```
## coxph(formula = Surv(TIMECVD, DEATH) ~ DIABETES + age.group +
##     bmi.group, data = cvd[cvd$CURSMOKE == 1 & cvd$SEX == 1, ])
##
##              coef exp(coef) se(coef)   z     p
## DIABETES1   0.2       1.2      0.4 0.4 0.693
## age.group1  0.6       1.8      0.2 2.9 0.004
## bmi.group1  0.4       1.5      0.2 1.9 0.056
##
## Likelihood ratio test=16  on 3 df, p=0.001
## n= 170, number of events= 96
```

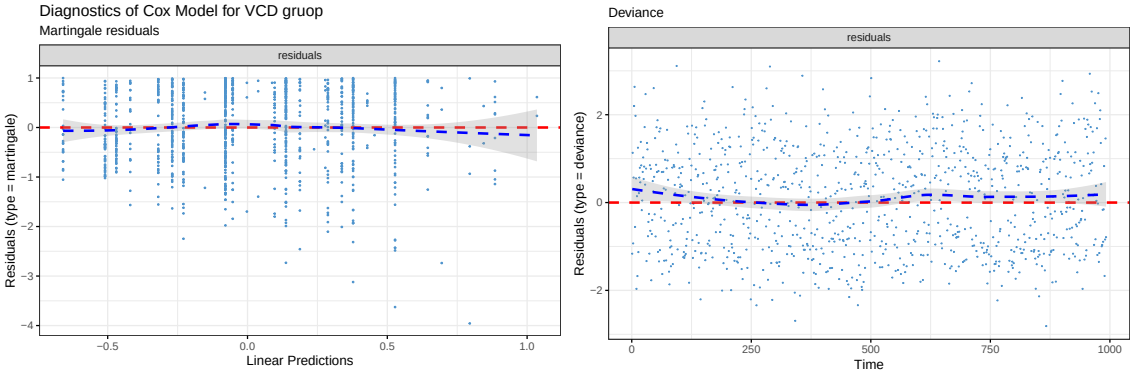


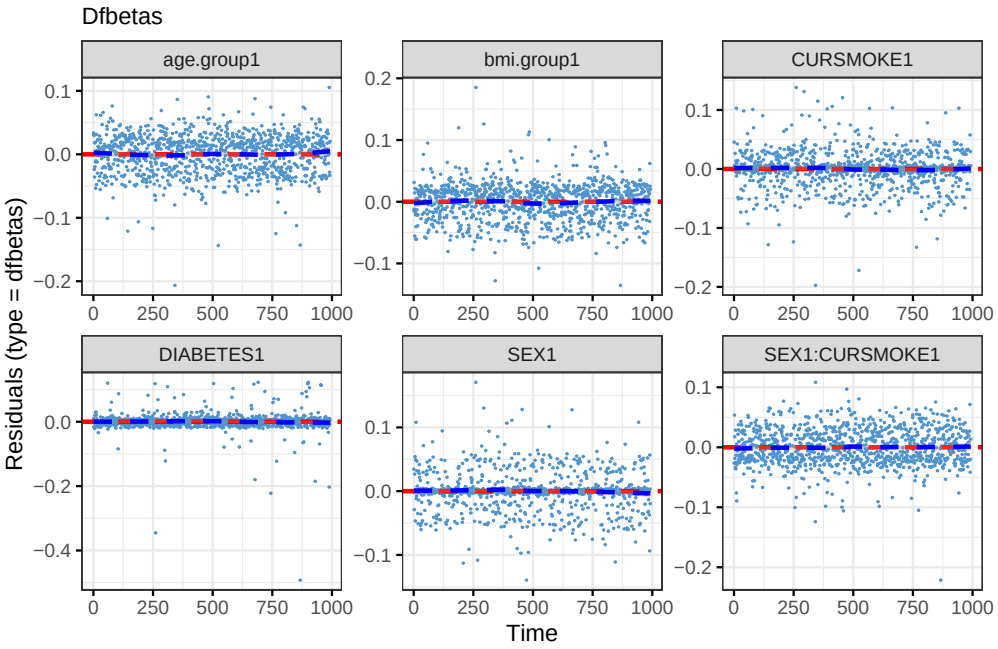
Model Checking

- QQ plot









Model Comparison

##

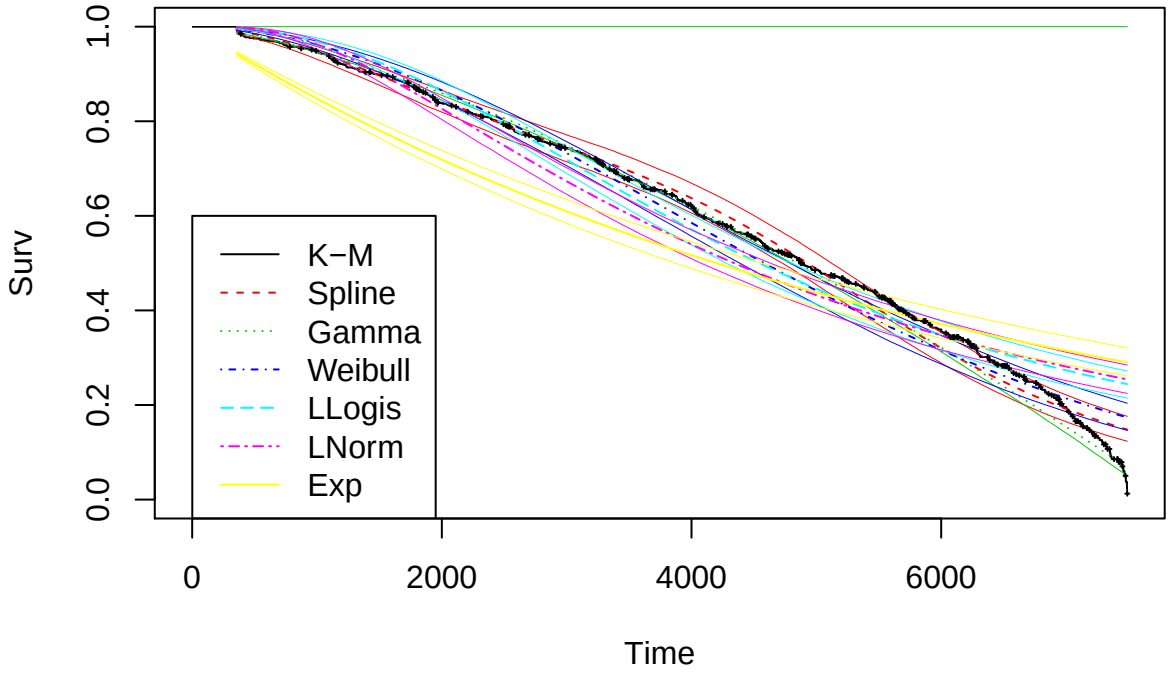
```
## Call:
## survreg(formula = Surv(TIMECVD, DEATH) ~ 1, data = cvd.int, dist = "weib")
##               Value Std. Error     z      p
## (Intercept)  8.6246     0.0226 381.1 <2e-16
## Log(scale)  -0.6353     0.0364 -17.5 <2e-16
##
## Scale= 0.53
##
## Weibull distribution
## Loglik(model)= -5221   Loglik(intercept only)= -5221
## Number of Newton-Raphson Iterations: 6
## n= 806
```

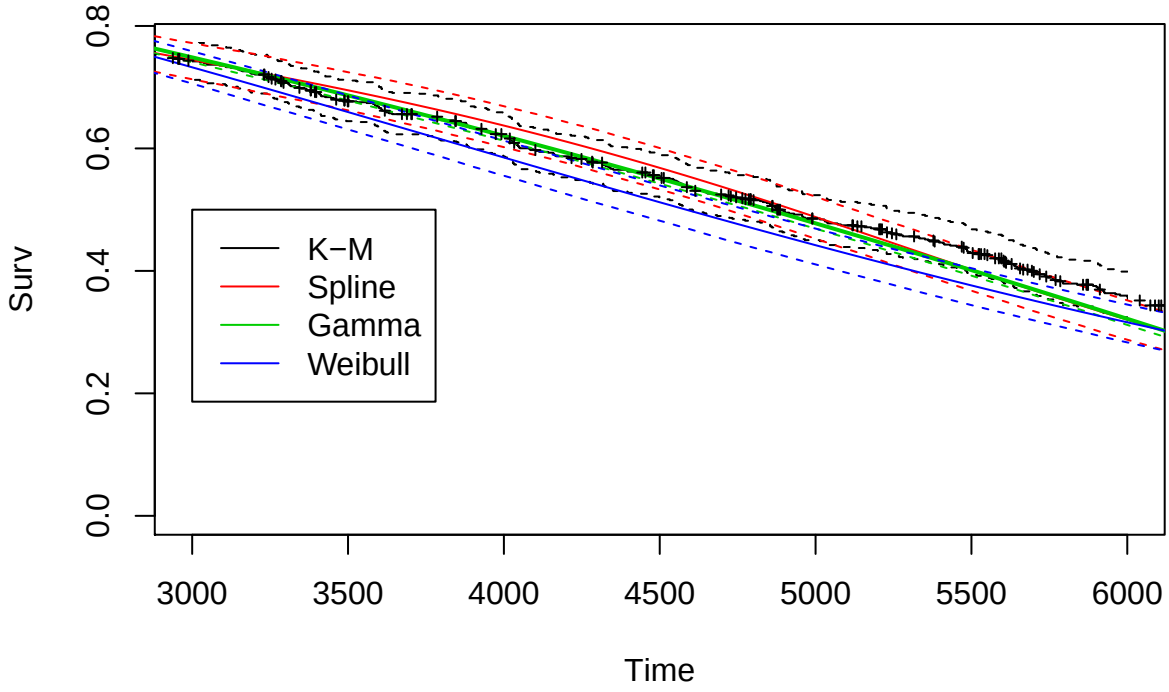
```
## Call:
## flexsurvreg(formula = Surv(TIMECVD, DEATH) ~ 1, data = cvd.int,
##     dist = "weibull")
##
## Estimates:
##        est        L95%       U95%       se
## shape     1.8876     1.7576     2.0271     0.0687
## scale  5566.9643  5325.4157  5819.4689   125.9950
##
## N = 806,  Events: 550,  Censored: 256
## Total time at risk: 3331396
## Log-likelihood = -5221, df = 2
## AIC = 10447
```

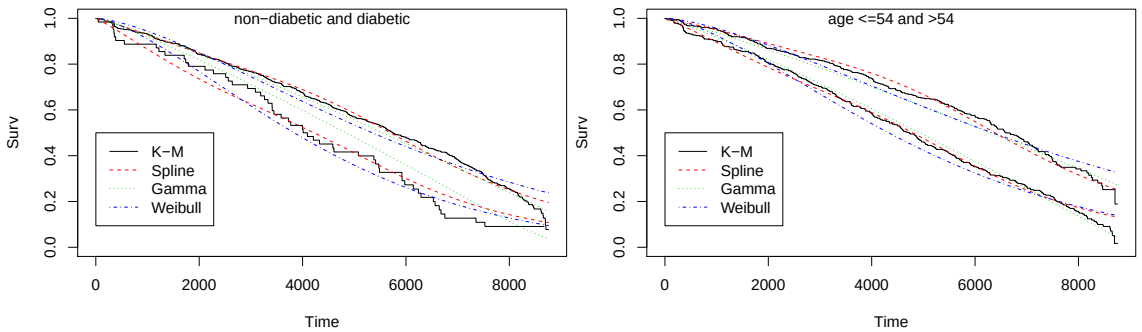
```
## Call:
## flexsurvreg(formula = Surv(TIMECVD, DEATH) ~ 1, data = cvd.int,
##     dist = "gengamma")
##
## Estimates:
##        est        L95%       U95%       se
## mu       8.94281    8.83019    9.05542    0.05746
## sigma    0.06914    0.00483    0.98879    0.09384
## Q       10.09763  -16.74527   36.94053   13.69561
##
## N = 806,  Events: 550,  Censored: 256
## Total time at risk: 3331396
## Log-likelihood = -5183, df = 3
## AIC = 10372
```

```
## Call:
## flexsurvspline(formula = Surv(TIMECVD, DEATH) ~ 1, data = cvd.int,
##     k = 2, scale = "odds")
##
## Estimates:
##         est      L95%     U95%     se
## gamma0  -15.834  -18.907  -12.762    1.568
## gamma1    1.883    1.451    2.314    0.220
## gamma2    2.032    1.460    2.603    0.292
## gamma3   -4.506   -5.626   -3.386    0.571
##
## N = 806,  Events: 550,  Censored: 256
## Total time at risk: 3331396
## Log-likelihood = -5208, df = 4
## AIC = 10425
```

- Compare these models


Surv
1.0
0.8
0.6
0.4
0.2
0.0
K−M
Spline
Gamma
Weibull
LLogis
LNorm
Exp
0
2000
4000
6000
Time


Surv
0.8
0.6
0.4
0.2
0.0
K−M
Spline
Gamma
Weibull
3000
3500
4000
4500
5000
5500
6000
Time


non−diabetic and diabetic
Surv
K−M
Spline
Gamma
Weibull
0
2000
4000
6000
8000
Time
age <=54 and >54
Surv
K−M
Spline
Gamma
Weibull
0
2000
4000
6000
8000
Time

The best fitted models are Flexible parametric model, Gamma model, and Weibull model.

Flexible parametric survival model. Flexible parametric survival regression was proposed by Royston and Parmar in 2002. It is an extension of the Weibull model, and models the log baseline cumulative hazard using restricted cubic splines.

Spline functions make the Weibull model more flexible to accommodate the survival distributions, rather than restriction to just monotonically increasing or decreasing trend.

Reasons for choosing the flexible parametric survival model

- Easier survival predictions: By harvesting information from the baseline hazard, the flexible parametric model can predict survival more easily.
- Extrapolation of survival estimates: Unlike the Cox model, the flexible parametric model allows extrapolation of survival estimates outside the study observation time. This means that it can predict survival outside the time range of the model that it was based on. However, accuracy of this prediction depends on the assumed (modelled) survival distribution in the upper tail of follow-up time.
- More flexible than other parametric models: the flexible parametric models are more flexible than other parametric models such as Weibull and exponential as it does not impose strong assumptions on the baseline hazard.
- Less sensitive to random variations due to sparse data: Since the fit from the Cox model very closely matches the data, it picks up artefacts (due by sparse data) that are specific features to the data it is based on, whilst the flexible parametric regression models the overall trend of the baseline hazard function without picking up random variations. Supplementary figure 1 (reproduced and modified from "Flexible Parametric Survival Analysis Using Stata Beyond The Cox Model": Chapter 1, page 4) shows differences in the hazard function fitted by the flexible parametric model and the Cox regression model. The curves from the Cox model were not smooth thus making them hard to interpret compared to the flexible parametric model. The latter does not pick up the sharp increase in the mortality rate post 4.5 years caused by a small number of deaths at the end of follow-up having an undue influence on the mortality rate whereas the Cox model does. Whilst this will not cause problems when calculating hazard ratios (since these sparse points will not be given much emphasis), it is an issue when calculating absolute risk as such points will lead to "wild" estimates, particularly in situations where a small number of patients remain at risk at the end of the study period.

(Berhane, S., Fox, R., García-Fiñana, M., Cucchetti, A. and Johnson, P., Using prognostic and predictive clinical features to make personalised survival prediction in advanced hepatocellular carcinoma patients undergoing sorafenib treatment. British journal of cancer, p.1. 2019)

Regression models

- Fit regression model with Weibull, Gamma, and Spline.

```
## Call:
## flexsurvreg(formula = Surv(TIMECVD, DEATH) ~ SEX + strata(CURSMOKE) +
##     DIABETES + age.group + bmi.group, data = cvd.int, dist = "weibull")
##
## Estimates:
##                     data mean   est         L95%        U95%        se
## shape                      NA     1.92515     1.79324     2.06677     0.06972
## scale                      NA  6957.52729  6136.01698  7889.02412   446.02938
## SEX1                  0.41563     0.09268     0.00171     0.18365     0.04641
## strata(CURSMOKE)1     0.53350    -0.09380    -0.18652    -0.00108     0.04731
## DIABETES1             0.06452    -0.19364    -0.34804    -0.03924     0.07878
## age.group1            0.46526    -0.32422    -0.41664    -0.23181     0.04715
## bmi.group1            0.68362    -0.05901    -0.15547     0.03746     0.04922
##                     exp(est)    L95%        U95%
## shape                      NA          NA          NA
## scale                      NA          NA          NA
## SEX1                  1.09712     1.00172     1.20160
## strata(CURSMOKE)1     0.91046     0.82984     0.99892
## DIABETES1             0.82396     0.70607     0.96152
## age.group1            0.72309     0.65926     0.79310
## bmi.group1            0.94270     0.85601     1.03817
##
## N = 806,  Events: 550,  Censored: 256
## Total time at risk: 3331396
## Log-likelihood = -5191, df = 7
## AIC = 10397
```

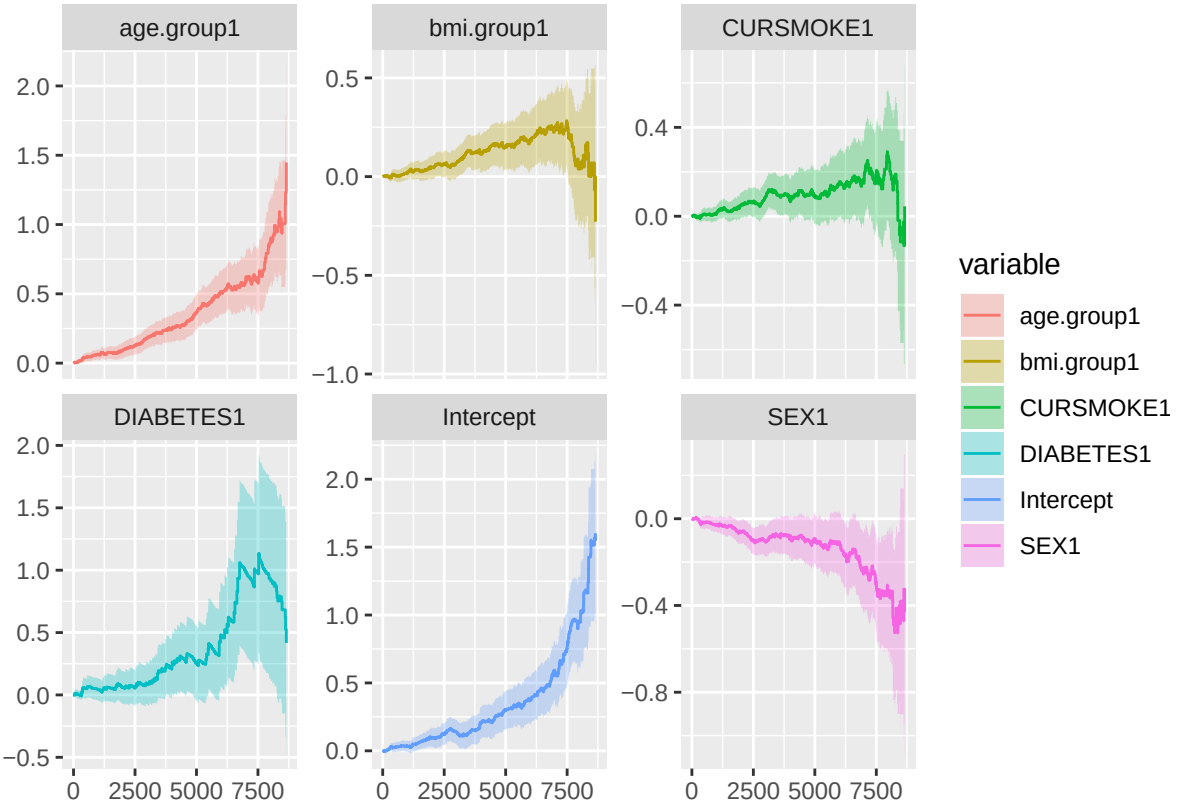
```
## Call:
## flexsurvreg(formula = Surv(TIMECVD, DEATH) ~ SEX + strata(CURSMOKE) +
##     DIABETES + age.group + bmi.group, data = cvd.int, dist = "gengamma")
##
## Estimates:
##                     data mean  est       L95%      U95%      se        exp(est)
## mu                        NA    9.04984   8.96996   9.12972   0.04076        NA
## sigma                     NA    0.06586   0.00973   0.44595   0.06427        NA
## Q                         NA   10.84247  -9.85221  31.53715  10.55871        NA
## SEX1                 0.41563    0.01233  -0.02213   0.04679   0.01758   1.01241
## strata(CURSMOKE)1    0.53350   -0.02412  -0.05150   0.00327   0.01397   0.97617
## DIABETES1            0.06452   -0.13622  -0.16648  -0.10597   0.01544   0.87265
## age.group1           0.46526   -0.14248  -0.18974  -0.09522   0.02411   0.86720
## bmi.group1           0.68362    0.01051  -0.02736   0.04837   0.01932   1.01056
##                     L95%      U95%
## mu                        NA        NA
## sigma                     NA        NA
## Q                         NA        NA
## SEX1                 0.97811   1.04790
## strata(CURSMOKE)1    0.94980   1.00327
## DIABETES1            0.84664   0.89945
## age.group1           0.82717   0.90917
## bmi.group1           0.97301   1.04956
##
## N = 806,  Events: 550,  Censored: 256
## Total time at risk: 3331396
## Log-likelihood = -5157, df = 8
## AIC = 10330
```

```
## Call:
## flexsurvspline(formula = Surv(TIMECVD, DEATH) ~ SEX + strata(CURSMOKE) +
##     DIABETES + age.group + bmi.group, data = cvd.int, k = 2,
##     scale = "odds")
##
## Estimates:
##                     data mean  est        L95%       U95%       se
## gamma0                    NA  -16.61807  -19.71443  -13.52172    1.57980
## gamma1                    NA    1.87966    1.44770    2.31161    0.22039
## gamma2                    NA    2.11619    1.52906    2.70332    0.29956
## gamma3                    NA   -4.72377   -5.87867   -3.56887    0.58924
## SEX1                 0.41563   -0.23693   -0.50556    0.03171    0.13706
## strata(CURSMOKE)1    0.53350    0.29806    0.02555    0.57057    0.13904
## DIABETES1            0.06452    0.49367    0.00849    0.97884    0.24754
## age.group1           0.46526    0.92883    0.65907    1.19858    0.13763
## bmi.group1           0.68362    0.25138   -0.02996    0.53272    0.14354
##                     exp(est)   L95%       U95%
## gamma0                    NA         NA         NA
## gamma1                    NA         NA         NA
## gamma2                    NA         NA         NA
## gamma3                    NA         NA         NA
## SEX1                 0.78905    0.60317    1.03221
## strata(CURSMOKE)1    1.34724    1.02588    1.76927
## DIABETES1            1.63831    1.00853    2.66136
## age.group1           2.53153    1.93299    3.31541
## bmi.group1           1.28579    0.97048    1.70356
##
## N = 806,  Events: 550,  Censored: 256
## Total time at risk: 3331396
## Log-likelihood = -5180, df = 9
## AIC = 10378
```

Extended Cox model

```
## Call:
## aareg(formula = Surv(TIMECVD, DEATH) ~ SEX + DIABETES + CURSMOKE +
##     age.group + bmi.group, data = cvd)
```

```
##
##   n= 992
##     603 out of 607 unique event times used
##
##                 slope      coef se(coef)     z        p
## Intercept   0.0001200  0.001100 0.000176  6.25 4.16e-10
## SEX1       -0.0000543 -0.000406 0.000142 -2.87 4.14e-03
## DIABETES1   0.0001640  0.001020 0.000372  2.76 5.83e-03
## CURSMOKE1   0.0000481  0.000273 0.000146  1.87 6.11e-02
## age.group1  0.0001500  0.001060 0.000149  7.12 1.10e-12
## bmi.group1  0.0000613  0.000326 0.000146  2.23 2.54e-02
##
## Chisq=73.72 on 5 df, p=2e-14; test weights=aalen
```



The Aalen model assumes that the cumulative hazard $H(t)$ for a subject can be expressed as $a(t) + XB(t)$, where $a(t)$ is a time-dependent intercept term, X is the vector of covariates for the subject (possibly time-dependent), and $B(t)$ is a time-dependent matrix of coefficients.

Multi-state models

The properties of recurrent event models:

- For each event type, recurrent or terminal, there exist separate event processes that might be correlated or not.
- The event-specific treatment effects related to the different event types may deviate.
- After occurrence of an event, the instantaneous baseline risk for a subsequent event, fatal or non-fatal, increases.
- The instantaneous risk for a subsequent event depends on the time when the previous event occurred.
- After occurrence of an event, the relative treatment effect for a subsequent event (in terms of the hazard ratio) may change.

The most simple analysis approach in a recurrent event setting is to count the events observed within a given time period. These counts may, for example, follow a Poisson, a quasi-Poisson or a negative binomial distribution

The Andersen-Gill model assumes independence between all observed event times irrespective whether these event times correspond to the same patient or to different patients.

The Prentice-Williams-Peterson model is a stratified Cox-based conditional model which incorporate the order of events. Based on different time scales, the gap time approach investigates the time since the last event whereas the calendar or total time scale considers the time since study entry.

Wei-Lin-Weissfeld model is an unconditional marginal model. It ignores the order of occurrence of the events. Therefore, for each subsequent event all individuals are at risk independent of a proceeding event.

(Ozga, A., Kieser, M. & Rauch, G. A systematic comparison of recurrent event models for application to composite endpoints. BMC Med Res Methodol 18, 2 (2018). https://doi.org/10.1186/s12874-017-0462-x)

```
## Cox-Aalen Model
##
## Test for Aalen terms
## Test not computed, sim=0
##
## Proportional Cox terms :
##                  Coef.      SE Robust SE D2log(L)^-1     z    P-val lower2.5%
## prop(AGE)      0.0499 0.00514   0.00524     0.00535  9.52 0.000000  0.039800
## prop(SEX)1    -0.3110 0.08910   0.08840     0.09000 -3.52 0.000434 -0.486000
## prop(GLUCOSE)  0.0049 0.00107   0.00120     0.00108  4.07 0.000046  0.002800
## prop(BMI)      0.0175 0.00938   0.00955     0.00962  1.83 0.067500 -0.000884
## prop(CIGPDAY)  0.0126 0.00323   0.00366     0.00340  3.45 0.000567  0.006270
##                upper97.5%
## prop(AGE)          0.0600
## prop(SEX)1        -0.1360
## prop(GLUCOSE)      0.0070
## prop(BMI)          0.0359
## prop(CIGPDAY)      0.0189
## Test of Proportionality
##                sup|  hat U(t) | p-value H_0
## prop(AGE)                99.00        0.894
## prop(SEX)1                9.91        0.378
## prop(GLUCOSE)           712.00        0.506
## prop(BMI)                92.00        0.292
## prop(CIGPDAY)           388.00        0.064


## Cox-Aalen Model
##
## Test for Aalen terms
## Test not computed, sim=0
##
## Proportional Cox terms :
##                   Coef.     SE Robust SE D2log(L)^-1     z    P-val lower2.5%
## prop(SEX)1       -0.247 0.0818    0.0831      0.0830 -2.97 2.94e-03   -0.4070
## prop(DIABETES)1   0.487 0.1440    0.1480      0.1440  3.28 1.03e-03    0.2050
## prop(CURSMOKE)1   0.135 0.0825    0.0864      0.0838  1.57 1.17e-01   -0.0267
## prop(age.group)1  0.607 0.0821    0.0840      0.0833  7.22 5.11e-13    0.4460
## prop(bmi.group)1  0.193 0.0863    0.0862      0.0863  2.24 2.50e-02    0.0239
##                  upper97.5%
## prop(SEX)1          -0.0867
## prop(DIABETES)1      0.7690
## prop(CURSMOKE)1      0.2970
## prop(age.group)1     0.7680
## prop(bmi.group)1     0.3620
## Test of Proportionality
##                  sup|  hat U(t) | p-value H_0
## prop(SEX)1                  12.40        0.212
## prop(DIABETES)1              4.05        0.816
## prop(CURSMOKE)1             13.90        0.130
## prop(age.group)1             5.88        0.940
## prop(bmi.group)1            10.90        0.262

##     AGE     SEX GLUCOSE     BMI CIGPDAY
##   1.096   1.106   1.020   1.019   1.165
```