Assignment 2 - STA2201H Applied Statistics II

Luis Correia - Student No. 1006508566

February 07th 2020

Question 1 - Gompertz

Gompertz hazards are of the form

$$\lambda(t) = \alpha e^{\beta t} \tag{1}$$

for t $\in [0, \infty)$ with $\alpha, \beta > 0$. It is named after Benjamin Gompertz, who suggested a similar form to capture a 'law of human mortality' in 1825.

This question uses the `ON_mortality.RDS` file in the data folder of the class repo. This file contains hazard rates (`hx`) and density of deaths (`dx`) by age and year for Ontario. Note that in this case, the survival times we are interested in are age.

(a) (5 points) Show that S(t) = $\exp\left[-\frac{\alpha}{\beta}\left(e^{\beta t}\right)\right]$ and $f(t) = \alpha \exp\left[\beta t - \frac{\alpha}{\beta}\left(e^{\beta t} - 1\right)\right]$.

{*Solution.*}

We know that the *hazzard function* can be written as function of p.d.f. of T (where T is the time until event) and its *survival function* as follows:

$$\implies \lambda\left(t\right) = \frac{f\left(t\right)}{S\left(t\right)} \tag{2}$$

From this known relation between $\lambda\left(t\right)$, $f\left(t\right)$ and $S\left(t\right)$, developing (2) we have:

$$\begin{aligned}\lambda\left(t\right) &= -\frac{d}{dt}\left[1 - F\left(t\right)\right]\frac{1}{S\left(t\right)}\\ &= -\frac{dS\left(t\right)}{dt}\frac{1}{S\left(t\right)}\end{aligned}$$

$$\implies \lambda\left(t\right) = -\frac{d}{dt}\ln\Big[S\left(t\right)\Big] \tag{3}$$

Integrating both sides in the interval $[0, t]$ we have:

$$\begin{aligned}\int_0^t \lambda\left(s\right)ds &= -\ln\Big[S\left(t\right)\Big]\\ \implies \exp\Big[-\int_0^t \lambda\left(s\right)ds\Big] &= S\left(t\right)\end{aligned}$$

Considering the Gompertz hazzard function as defined in (1) and applying on left side of equation above, we have:

$$\begin{aligned} S(t) &= \exp\Big[-\int_0^t \alpha e^{\beta s} ds\Big] \\ &= \exp\Big[-\frac{\alpha}{\beta} e^{\beta s}\Big|_0^t\Big] \\ &= \exp\Big[-\frac{\alpha}{\beta}\left(e^{\beta t}-1\right)\Big] \end{aligned}$$

$$\implies S(t) = \exp\Big[-\frac{\alpha}{\beta}\left(e^{\beta t}-1\right)\Big]. \tag{4}$$

Using the relation in (2) we have:

$$\begin{aligned} f(t) &= S(t)\,\lambda(t) \\ &= \exp\Big[-\frac{\alpha}{\beta}\left(e^{\beta t}-1\right)\Big]\alpha e^{\beta t} \\ &= \alpha \exp\Big[-\frac{\alpha}{\beta}\left(e^{\beta t}-1\right)+\beta t\Big] \end{aligned}$$

$$\implies f(t) = \alpha \exp\Big[\beta t - \frac{\alpha}{\beta}\left(e^{\beta t}-1\right)\Big]. \tag{5}$$

(b) (5 points) Find an expression in terms of α and β for the modal time to death.

{*Solution.*}

Modal Time to Death is the age of most deaths occurs. In this sense we need to find the *absolute maximum* of p.d.f of T.

Deriving the $f(t)$ obtained in equation (5) and making it equals to *zero*, we have:

$$\begin{aligned} f'(t) = \frac{df}{dt} &= \frac{d}{dt}\alpha \exp\Big[\beta t - \frac{\alpha}{\beta}\left(e^{\beta t}-1\right)\Big] \\ &= \alpha \exp\Big[\beta t - \frac{\alpha}{\beta}\left(e^{\beta t}-1\right)\Big]\left(\beta - \frac{\alpha}{\beta}e^{\beta t}\beta\right) = 0 \end{aligned}$$

From our hypotesis of *Gompertz Hazzard*, $\alpha > 0$ and $\beta > 0$, so this equality holds only if:

$$\begin{aligned} \beta - \alpha e^{\beta t} &= 0 \\ \beta &= \alpha e^{\beta t} \\ \frac{\beta}{\alpha} &= e^{\beta t} \\ \ln\frac{\beta}{\alpha} &= \beta t \end{aligned}$$

Then *'modal time to death'* (let say ψ) is given by:

$$\implies \psi = \frac{1}{\beta}\ln\left(\frac{\beta}{\alpha}\right). \tag{6}$$

(c) (10 points) Restrict the dataset to just look at ages between 40 and 100. (Note: the age column is a character, so you will first have to change it to be a numeric value). For the years 1961 and 2011, estimate α and β using `lm()` (with the appropriate transformation). Interpret you results. What do the estimates of α and β for the two years tell you about the difference in mortality conditions in the two years?

```
## # A tibble: 6 x 4
##    year age        hx      dx
##   <dbl> <chr>   <dbl>   <dbl>
## 1  1921 0     0.114   0.106
## 2  1921 1     0.0137  0.0122
## 3  1921 2     0.00631 0.00554
## 4  1921 3     0.00464 0.00405
## 5  1921 4     0.00447 0.00389
## 6  1921 5     0.00367 0.00318
```

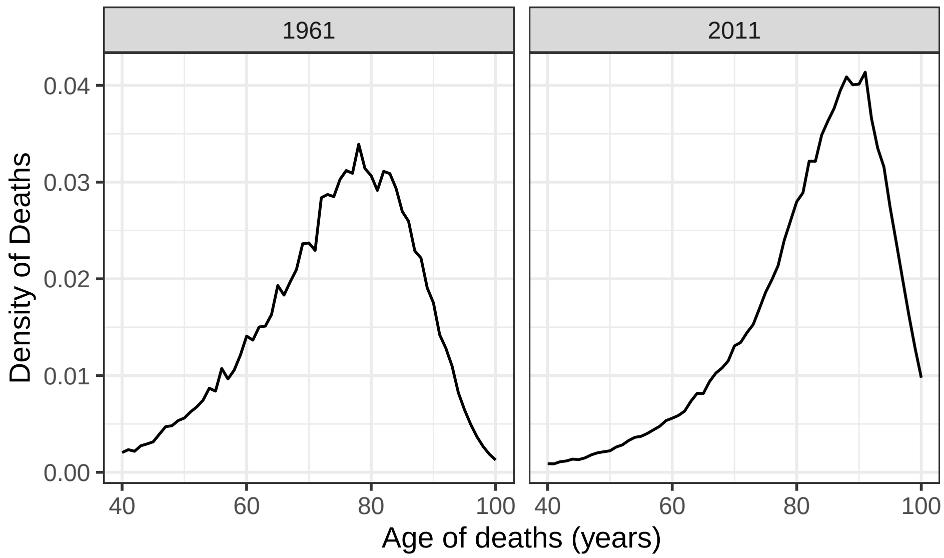

Figure 1: Density of deaths in Ontario From ages 40 - 100 years

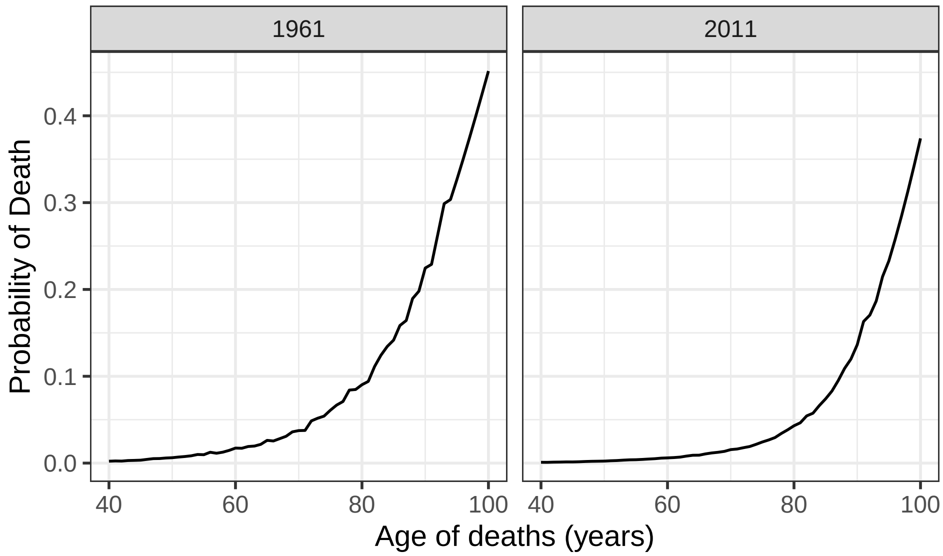

Figure 2: Probability of death in Ontario From ages 40 - 100 years

Table 1: Observed Modal of Death in 1961 and 2011

year	age	hx	dx
1961	78	0.08411	0.03392
2011	91	0.16307	0.04136

Initial assessment over the data brings to our attention some aspects that can be further investigated in relation to apparent different patterns of deaths on years 1961 and 2011. We can enumerate some of them, as follows:

- The density of deaths in year 1961 seems to be centered around 78 years while in 2011 it shifts to 91 years-old. This suggests health and life conditions improvements from one year in relation to the other;
- The variance also seems to have changed from 1961 to 2011, with deaths concentrated around the mean on year 2011 than in 1961, which presents a more dispersed pattern;
- The probability of death in 2011 is reduced when compared to 1961.

```
##
## Call:
## lm(formula = log(hx) ~ age, data = dtf1961)
##
## Residuals:
##       Min        1Q    Median        3Q       Max
## -0.264354 -0.031628  0.009561  0.041952  0.170266
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept) -9.5504575  0.0429962  -222.1   <2e-16 ***
## age          0.0892529  0.0005957   149.8   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
##
## Residual standard error: 0.08191 on 59 degrees of freedom
## Multiple R-squared:  0.9974, Adjusted R-squared:  0.9973
## F-statistic: 2.245e+04 on 1 and 59 DF,  p-value: < 2.2e-16
```

```
##
## Call:
## lm(formula = log(hx) ~ age, data = dtf2011)
##
## Residuals:
##       Min        1Q    Median        3Q       Max
## -0.177553 -0.082544  0.007177  0.089554  0.155594
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept) -11.123876   0.048863  -227.7   <2e-16 ***
## age           0.100601   0.000677   148.6   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.09309 on 59 degrees of freedom
## Multiple R-squared:  0.9973, Adjusted R-squared:  0.9973
## F-statistic: 2.208e+04 on 1 and 59 DF,  p-value: < 2.2e-16
```

The intercept $ln\left(\alpha\right)$ can be seen as a addictive factor influencing the value of $ln\left(\mu_{ProbabilityofDeath}\right)$ in the model, idependent of other covariates. On the other side, the β influences the *Probability of Death* according with the person's age. As it is positive in both model adjustments, we can conclude the probability increases with age.

The estimates of α and β for the two years shows a difference in mortality as a function of `age` between 1961 and 2011: the latter has decreased significantly when compared with 1961 as we can see in Table 2.

The behaviour shows that this improvement is more observable on early ages (i.e., 40-60 than 80-100) suggesting that healthcare in yonger people may reflect less probability of death and, as consequence, a longer life.

Besides, the difference of parameters estimates in 1961 and 2011 might suggest that mortality's association to other factors than just age is more present on 1961 than 2011, possibly reflecting improvements on health care and advances of medicine.

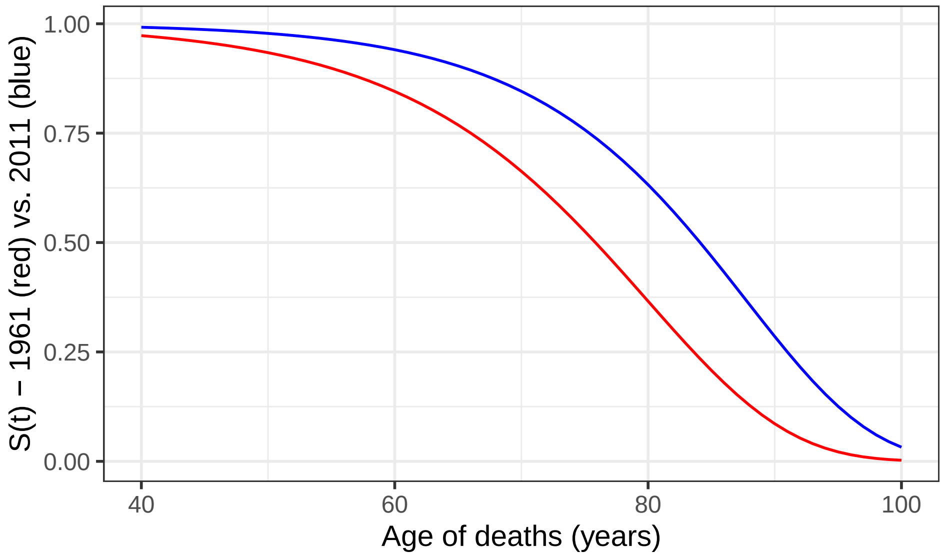

Figure 3: Estimated Survival Functions

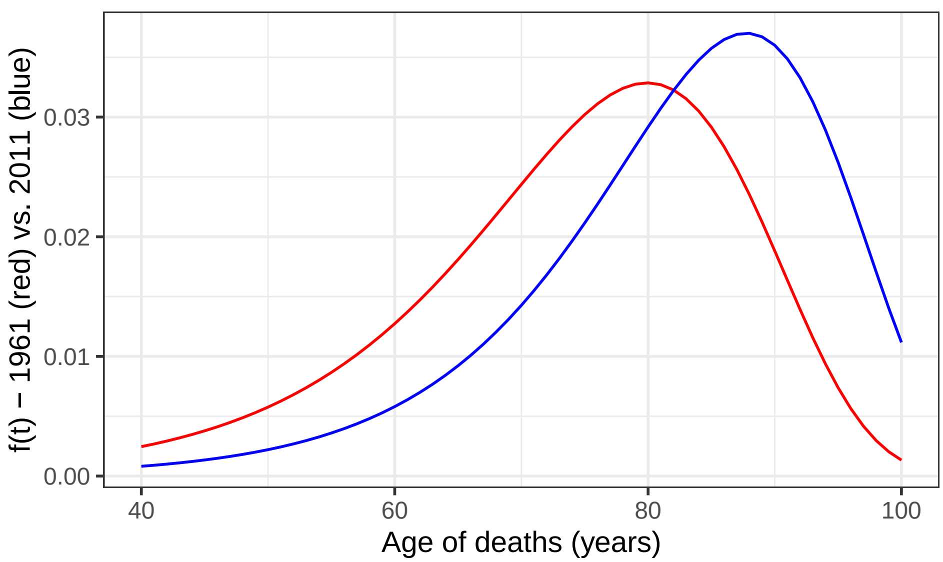

Figure 4: Estimated Survival Functions

Clearly the combined effect of estimated α and β suggests that improvements on probability of survival as the age grows, and for the reduction of probability of death as well, when we compare the years of 1961 and 2011.

Table 2: Relative Probability of death in 2011 in relation to 1961

	age	rt
1	40	-0.6735537
11	50	-0.6343240
21	60	-0.5903800
31	70	-0.5411551
41	80	-0.4860148
51	90	-0.4242482
61	100	-0.3550589

One can verify, for instance, the probability of death for a 50-year-old person has decreased around 63,4% from 1961 to 2011.

(d) (5 points) Plot the observed and estimated hazards from c) for both years on the log scale. How appropriate do you think the assumption of Gompertz hazards is for these data?

The Gompertz Hazzard describes very well the behaviour of probability of deaths in this model. This can be seen from the R^2 of linear model adjusted and from the graph of estimated and observed hazzards.

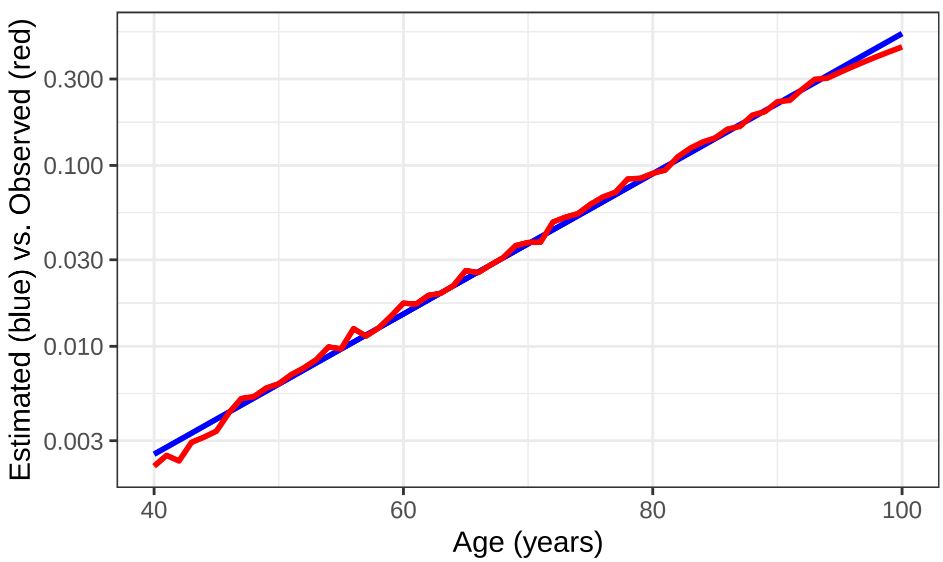

Figure 5: Hazards Rates in Ontario in 1961 using Gompertz Hazzard

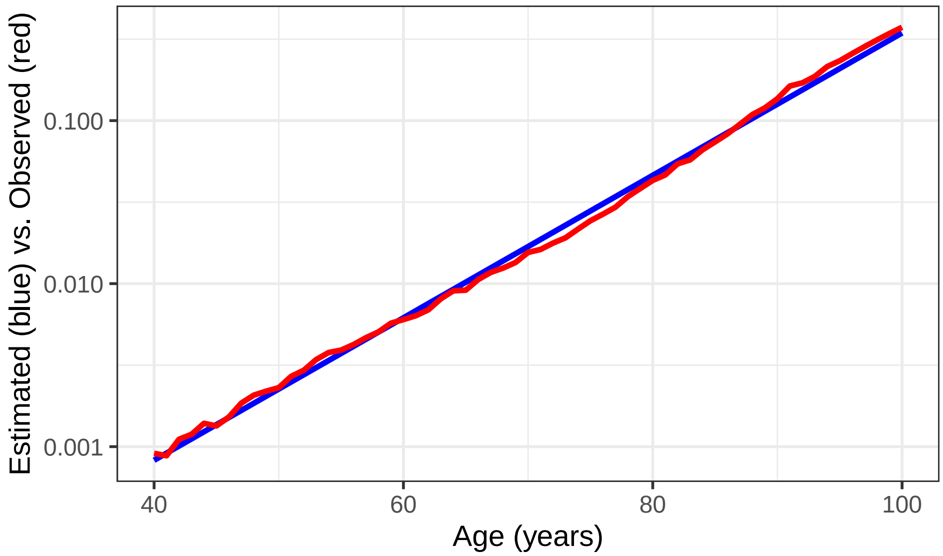

Figure 6: Hazards Rates in Ontario in 2011 using Gompertz Hazzard

Comparing both adjustments in a log-scale we can visually identify that probability of death decreased from 1961 to 2011.

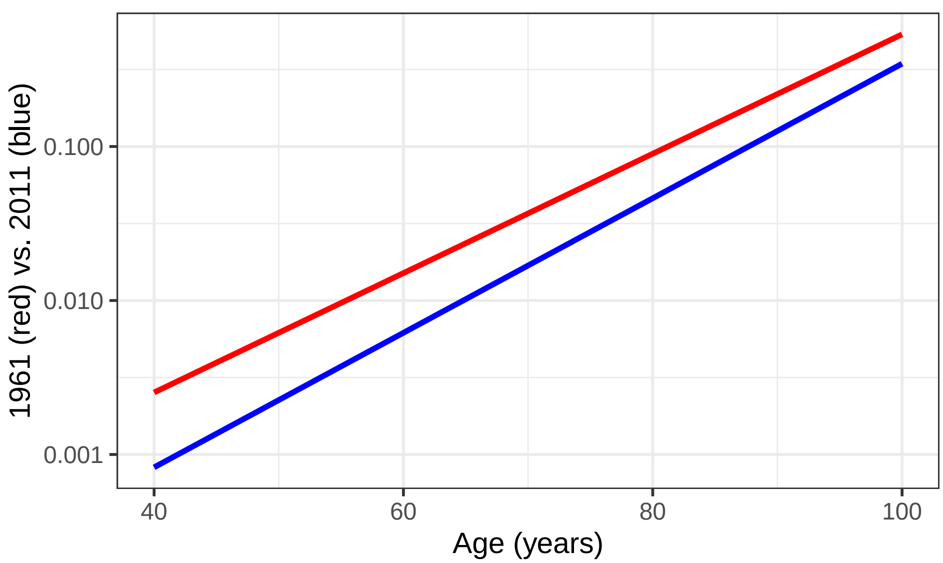

Figure 7: Comparison Hazards Rates in Ontario - 1961 vs 2011

We can observe that from 1961 to 2011 there is a crescent gap between the probability of death, that is, as the age grows, the probability of death in early 1961 is greater than in 2011. For example, at the age of 80, the probability of death in 2011 is more that $48,6\%$ less than in 1961.

(e) (5 points) Based on your estimates in c) calculate the modal age of death for 1961 and 2011. Plot the observed density of deaths and add a vertical line based on your estimated mode age at death.

Using the expression obtained in (6) we calculate the 'modal time to death' using the estimated parameters from the models adjusted in item (c).

Table 3: Modal Age to death 1961/2011

	1961	2011
Modal Age to Death	79.93211	87.74535

Reflecting this on densities from 1961 and 2011 we have:

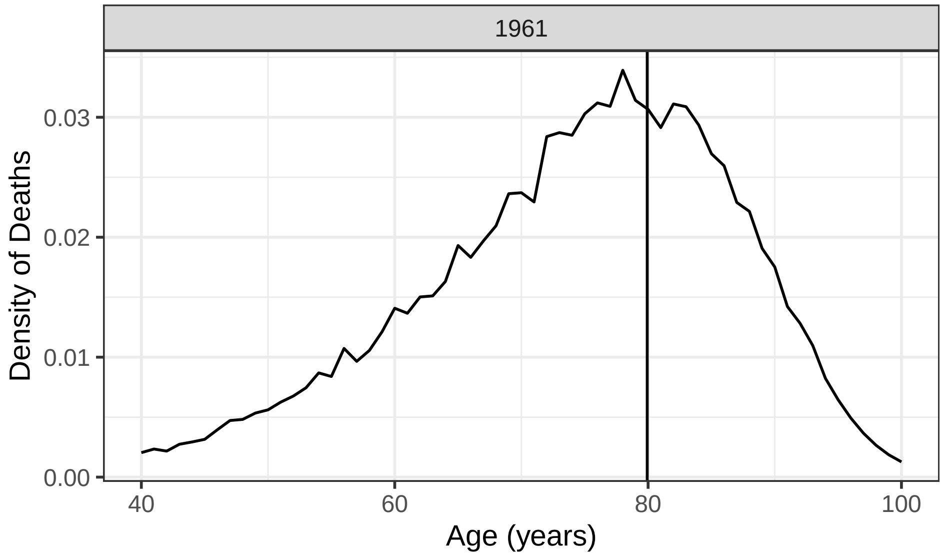

Figure 8: Density of Deaths in Ontario estimated Modal Time to Death

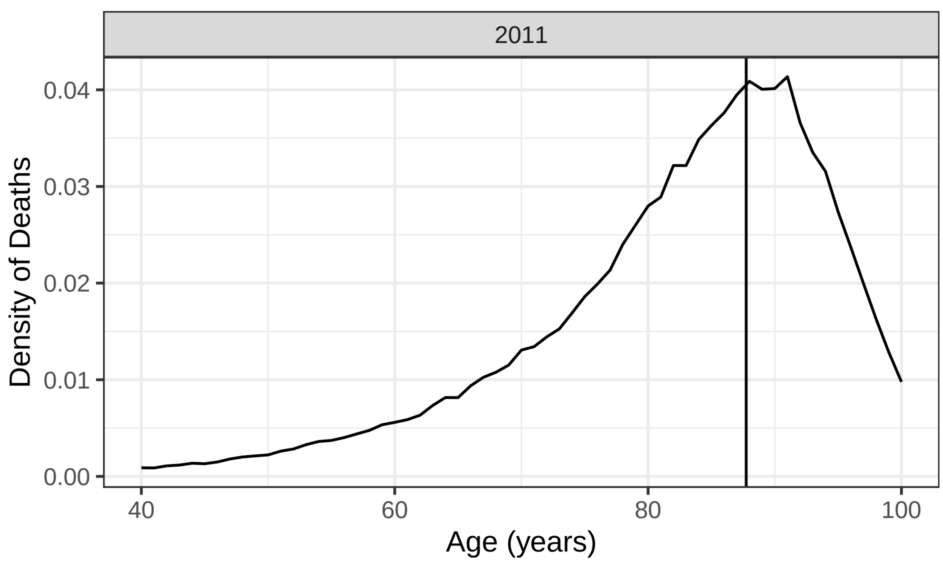

Figure 9: Density of Deaths in Ontario estimated Modal Time to Death

(f) (10 points) Repeat part d) for every year in the data set and then calculate the mode age at death for each year. Make a plot of α over time, β over time and the mode age at death over time. Write a few sentences interpreting these results in terms of how mortality has changed over time.

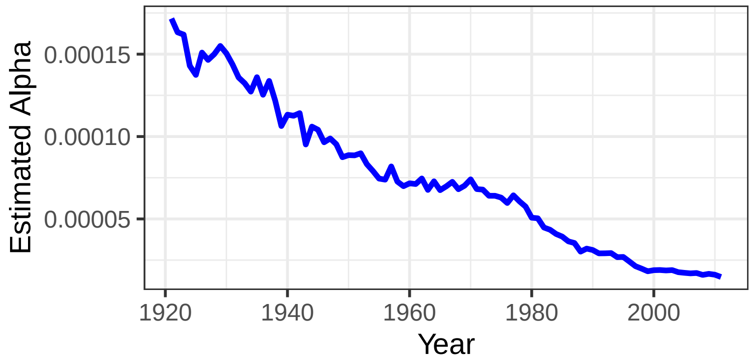

Figure 10: Alpha Estimates over Years

Figure 11: Beta Estimates over Years

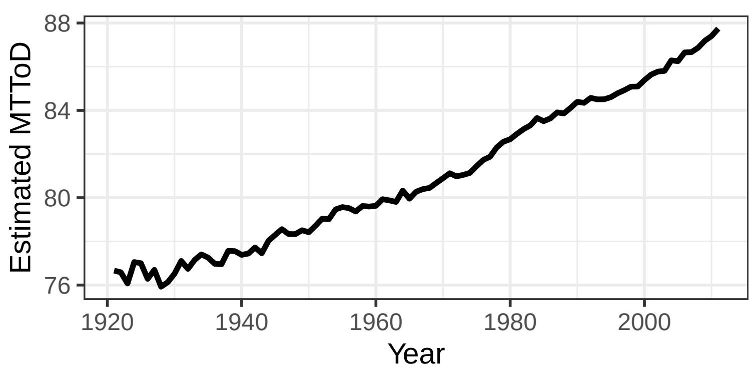

Figure 12: Modal Time To Death Estimates over Years

Comments on behaviour of estimates of α, β and *Modal Time to Death* from 1921 to 2011:

- α decreases over year, possibly reflecting improvements on health-care through the years;
- β increases meaning the association of `age` is more present in probability of death, meaning other factors might influence more the probability of death in early than more recent years;

- The final effect is that *Modal Time to Death* increases over years, indicating the populatoion from Ontario has been living longer as the years goes by, probably reflecting better quality of life and good health care.

Question 2 - Infant Mortality

In this part we will be looking at the infant mortality data set that we used in Lab 2. This is in the data folder called `infant.RDS`. Remember that this dataset contains individual-level data (i.e., every row is a death) on deaths in the first year of life for the US 2012 birth cohort. A second dataset you will be using for this question is `births.RDS`, which tabulates the total number of live births for the US 2012 birth cohort by race and prematurity. Descriptions of each variable can be found in the `infant_mortality_codebook.txt` file.

For this question we are interested in looking at the distribution of ages at death in more detail. In particular, the goal is to investigate differences in ages at death by race of mother and prematurity (from extremely preterm to full-term).

```
## # A tibble: 6 x 9
##   sex    aged race   gest ucod  cod       mom_age mom_age_group prematurity
##   <chr> <dbl> <chr> <dbl> <chr> <chr>       <dbl> <fct>         <fct>
## 1 F         0 NHW      27 P832  peri_oth       30 30            extremely prete~
## 2 M         0 NHW      36 Q913  cong_mal       32 30            later preterm
## 3 M         8 NHW      44 P360  peri_inf       25 25            full-term
## 4 F         0 NHB      21 P072  peri_comp      29 25            extremely prete~
## 5 M         8 NHB      26 P220  peri_resp      23 20            extremely prete~
## 6 M        17 NHW      39 Q249  cong_mal       34 30            full-term
```

```
## # A tibble: 6 x 3
## # Groups:   race [2]
##   race  prematurity       births
##   <chr> <fct>              <dbl>
## 1 NHB   extremely preterm   9783
## 2 NHB   very preterm       11836
## 3 NHB   later preterm      74656
## 4 NHB   full-term         486312
## 5 NHW   extremely preterm  11423
## 6 NHW   very preterm       21622
```

(a) (4 points) The infant mortality rate (IMR) is defined as the number of deaths in the first year divided by the number of live births. Calculate the IMR for the non-Hispanic black (NHB) and non-Hispanic white (NHW) populations. What is the ratio of black-to-white mortality?

{*Answer.*}

The **IMR** can be summarized in table 4, below.

Table 4: IMR by Race

Race	IMR
NHB	0.0109975
NHW	0.0049788

```
## Ratio of Black-to-White mortality is  2.208866
```

(b) (15 points) Calculate the Kaplan-Meier estimate of the survival function for each race and prematurity category (i.e. you should end up with 8 sets of survival functions). Also calculate the standard error of the estimates of the survival function, based on taking the square root of the variance formula shown in the lecture slides. Note that to calculate the survival function you will need to incorporate information from the births file, not just the deaths (otherwise it will look like everyone died). It will probably be easiest to first tabulate the number of deaths by `aged` for each group first, rather than looking at individual-level data.

{*Answer.*}

As this question is a matter of transforming and manipulate the data, we will just print part of the final data generated, including the estimated Survival function, Variance and C.I. estimates. The code can be verified in `.rmd` file.

Table 5: K-M estimates for Survival Function (illustrative - first 30 lines)

race	prematurity	births	aged	event	tot_events	prob_death	prob_surv	surv	varSurv	upperCI	lowerCI
NHB	extremely preterm	9783	999	0	0	0.0000000	1.0000000	1.0000000	0.00e+00	1.0000000	1.0000000
NHB	very preterm	11836	999	0	0	0.0000000	1.0000000	1.0000000	0.00e+00	1.0000000	1.0000000
NHB	later preterm	74656	999	0	0	0.0000000	1.0000000	1.0000000	0.00e+00	1.0000000	1.0000000
NHB	full-term	486312	999	0	0	0.0000000	1.0000000	1.0000000	0.00e+00	1.0000000	1.0000000
NHW	extremely preterm	11423	999	0	0	0.0000000	1.0000000	1.0000000	0.00e+00	1.0000000	1.0000000
NHW	very preterm	21622	999	0	0	0.0000000	1.0000000	1.0000000	0.00e+00	1.0000000	1.0000000
NHW	later preterm	186411	999	0	0	0.0000000	1.0000000	1.0000000	0.00e+00	1.0000000	1.0000000
NHW	full-term	1912986	999	0	0	0.0000000	1.0000000	1.0000000	0.00e+00	1.0000000	1.0000000
NHB	extremely preterm	NA	0	1	2379	0.1783893	0.8216107	0.8216107	1.10e-05	0.8282410	0.8149804
NHB	very preterm	NA	0	1	118	0.0095966	0.9904034	0.9904034	8.00e-07	0.9921618	0.9886450
NHB	later preterm	NA	0	1	126	0.0016727	0.9983273	0.9983273	0.00e+00	0.9986251	0.9980296
NHB	full-term	NA	0	1	137	0.0002807	0.9997193	0.9997193	0.00e+00	0.9997672	0.9996713
NHW	extremely preterm	NA	0	1	2870	0.1840215	0.8159785	0.8159785	9.60e-06	0.8221842	0.8097727
NHW	very preterm	NA	0	1	324	0.0144237	0.9855763	0.9855763	6.00e-07	0.9871673	0.9839852
NHW	later preterm	NA	0	1	465	0.0024731	0.9975269	0.9975269	0.00e+00	0.9977560	0.9972978
NHW	full-term	NA	0	1	471	0.0002457	0.9997543	0.9997543	0.00e+00	0.9997769	0.9997317
NHB	extremely preterm	NA	1	1	143	0.0130510	0.9869490	0.8108878	1.15e-05	0.8176698	0.8041058
NHB	very preterm	NA	1	1	32	0.0026277	0.9973723	0.9878009	1.00e-06	0.9897808	0.9858210
NHB	later preterm	NA	1	1	38	0.0005053	0.9994947	0.9978229	0.00e+00	0.9981625	0.9974832
NHB	full-term	NA	1	1	28	0.0000574	0.9999426	0.9996619	0.00e+00	0.9997145	0.9996093
NHW	extremely preterm	NA	1	1	210	0.0165017	0.9834983	0.8025135	1.02e-05	0.8088890	0.7961379
NHW	very preterm	NA	1	1	49	0.0022133	0.9977867	0.9833949	7.00e-07	0.9851001	0.9816897
NHW	later preterm	NA	1	1	81	0.0004319	0.9995681	0.9970961	0.00e+00	0.9973443	0.9968479
NHW	full-term	NA	1	1	108	0.0000564	0.9999436	0.9996980	0.00e+00	0.9997231	0.9996729
NHB	extremely preterm	NA	2	1	94	0.0086924	0.9913076	0.8038392	1.18e-05	0.8107164	0.7969621
NHB	very preterm	NA	2	1	29	0.0023876	0.9976124	0.9854424	1.20e-06	0.9876027	0.9832821
NHB	later preterm	NA	2	1	20	0.0002661	0.9997339	0.9975574	0.00e+00	0.9979171	0.9971977
NHB	full-term	NA	2	1	21	0.0000430	0.9999570	0.9996189	0.00e+00	0.9996748	0.9995630
NHW	extremely preterm	NA	2	1	157	0.0125439	0.9874561	0.7924468	1.05e-05	0.7989417	0.7859519
NHW	very preterm	NA	2	1	43	0.0019466	0.9980534	0.9814807	8.00e-07	0.9832797	0.9796816

(c) (5 points) Plot your results from b), showing the estimate and +/- 2 standard errors. What the plot should look like: NHB and NHW survival curves on the one plot; one separate facet per prematurity category. Note that the survival curves are very different by prematurity category, so it might help to make the y axes different scales for each category (e.g. `facet_grid(prematurity~, scales = "free_y")`).

{*Answer.*}

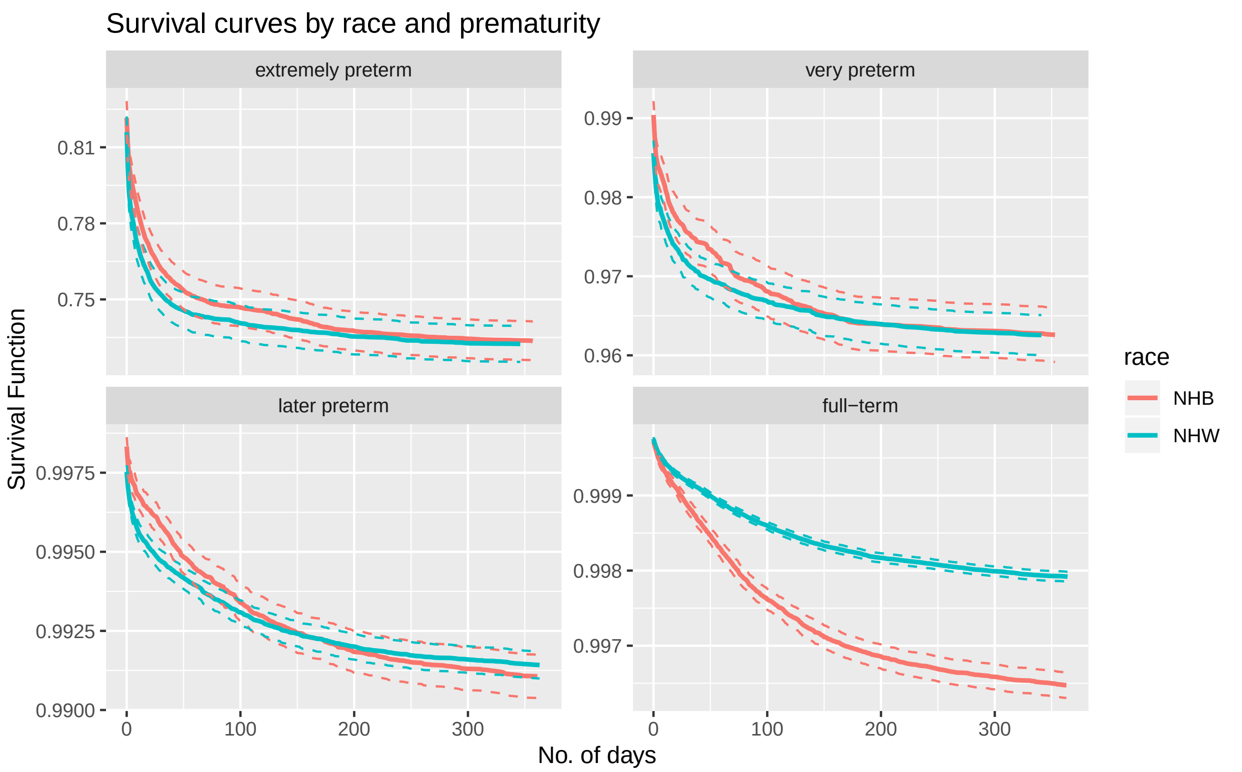

Figure 13: Kaplan-Meier Estimates of S(t) w/ CI

(d) (3 points) On first glance, your plots in c) might contradict what you expected based on a). Why is the IMR so much higher for the NHB population, even though for (most) prematurity groups, the survival curves are reasonably similar to the NHW population?

{*Answer.*}

That ocurred because the proportion of *NHW births* was **3.6 times greater** than *NHB births*. Even if deaths of NHB infants are higher, the proportion of newborn NHW babies surpasses it and distorts this result. When we calculate the survival functions, we splitted thos difference between the different `prematurity` levels which provides a different perspective.

ON the other hand, we can also have differences between groups of `prematurity` that causes this misleading behaviour. For instance, for the `full-term` group, the survival curve of NHW is much higher than for group NHB.

(e) (3 points) Now consider fitting a piece-wise constant hazards model to the survival time data with cut-points at 1, 7, 14, 28, 60, 90 and 120 days. Consider a model that has race and prematurity as covariates. You could fit this model just using the deaths data, but the direction of the sign of the coefficient on race would be misleading. Why is that?

{*Answer.*}

To answer this question, we will do some data manipulation and adjust a model, like we did in Lab exercise.

```
##
## Call:
## glm(formula = event ~ offset(log(interval_length)) - 1 + interval +
##     race + prematurity, family = "poisson", data = dtinf_split)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -2.8914  -0.7379  -0.5074   0.6299   2.8400
##
## Coefficients:
##                          Estimate Std. Error  z value Pr(>|z|)
## interval-1               -0.82018    0.01574  -52.119  < 2e-16 ***
## interval1                -2.79317    0.02799  -99.790  < 2e-16 ***
## interval7                -3.24692    0.03467  -93.651  < 2e-16 ***
## interval14               -3.56662    0.03260 -109.409  < 2e-16 ***
## interval28               -3.74592    0.02881 -130.001  < 2e-16 ***
## interval60               -3.78709    0.03431 -110.387  < 2e-16 ***
## interval90               -3.78134    0.03835  -98.594  < 2e-16 ***
## interval120              -3.57049    0.02549 -140.086  < 2e-16 ***
## raceNHW                   0.11662    0.01608    7.251 4.14e-13 ***
## prematurityvery preterm  -0.63380    0.03020  -20.984  < 2e-16 ***
## prematuritylater preterm -0.93639    0.02448  -38.244  < 2e-16 ***
## prematurityfull-term     -1.23917    0.01880  -65.907  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for poisson family taken to be 1)
##
##     Null deviance: 1440805  on 56513  degrees of freedom
## Residual deviance:   51073  on 56501  degrees of freedom
## AIC: 85145
##
## Number of Fisher Scoring iterations: 6
```

After fitting the model, the coefficients of time intervals shows a decreasing influence on probability death as the time interval increases in time, this means that, as the time passes, the probability of death is expected to decrease. The same behaviour have the covariate `prematurity`, i.e., influence to death of `full-term` is less than `very preterm` and so on.

As expected, `race` plays a different behaviour influencing **positively** the death event and, in my opinion, this might to sime additional causes, for instance, differences on probability of death between intervals, with NHW and NHB alternating due to more or less frequency of death, or even because we are just considering the *death database*, not incorporating the births to it as pointed on item 2.d). In this case, as the proportion of NHB and NHW differs, this might affect the coefficients of fitted model.

(f) (20 points) Fit a piece-wise constant hazards model with cut-points as specified in e). Note given the large numbers of births/deaths, it will be much easier to run the model based on the tabulated deaths/exposures by age at death, rather than individual-level data. Include as covariates race and prematurity, and allow the hazard ratios of each to vary by interval.

Calculate the hazard of dying in the first interval (0-1 day) of extremely preterm babies born to NHB mothers. In addition, give the hazard ratios of dying for:

- extremely preterm babies to NHW mothers compared to extremeley preterm babies to NHB mothers in the first interval (0-1 days).
- full-term babies to NHB mothers compared to extremeley preterm babies to NHB mothers in the first interval (0-1 days).
- full-term babies to NHB mothers compared to extremeley preterm babies to NHB mothers in the last interval (120-365 days).
- full-term babies to NHW mothers compared to full-term babies to NHB mothers in the last interval (120-365 days).

```
##
## Call:
## glm(formula = event ~ offset(log(interval_length)) - 1 + interval +
##     race + prematurity + interval * race + interval * prematurity +
##     race * prematurity, family = "poisson", data = dtFull_split,
##     weights = n)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -74.773   -0.104   -0.059   -0.034  234.962
##
## Coefficients:
##                                       Estimate Std. Error  z value Pr(>|z|)
## interval0                             -1.48269    0.01942  -76.366  < 2e-16 ***
## interval1                             -5.42347    0.05137 -105.570  < 2e-16 ***
## interval7                             -6.00124    0.06364  -94.295  < 2e-16 ***
## interval14                            -6.60740    0.06104 -108.249  < 2e-16 ***
## interval28                            -7.49892    0.06255 -119.887  < 2e-16 ***
## interval60                            -8.74162    0.11523  -75.865  < 2e-16 ***
## interval90                            -8.98814    0.12648  -71.063  < 2e-16 ***
## interval120                          -11.00194    0.06688 -164.513  < 2e-16 ***
## raceNHW                                0.07506    0.02534    2.962  0.00305 **
## prematurityvery preterm               -2.82194    0.06297  -44.811  < 2e-16 ***
## prematuritylater preterm              -4.58204    0.05681  -80.653  < 2e-16 ***
## prematurityfull-term                  -6.45484    0.05208 -123.948  < 2e-16 ***
## interval1:raceNHW                      0.08422    0.06109    1.379  0.16803
## interval7:raceNHW                     -0.21250    0.07372   -2.883  0.00395 **
## interval14:raceNHW                    -0.27688    0.06926   -3.997 6.40e-05 ***
## interval28:raceNHW                    -0.55148    0.06232   -8.848  < 2e-16 ***
## interval60:raceNHW                    -0.59164    0.07502   -7.887 3.10e-15 ***
## interval90:raceNHW                    -0.38580    0.08483   -4.548 5.42e-06 ***
## interval120:raceNHW                   -0.41315    0.05777   -7.152 8.57e-13 ***
## interval1:prematurityvery preterm      1.06216    0.09505   11.175  < 2e-16 ***
## interval7:prematurityvery preterm      1.11686    0.12036    9.279  < 2e-16 ***
## interval14:prematurityvery preterm     1.14739    0.11708    9.800  < 2e-16 ***
## interval28:prematurityvery preterm     1.42106    0.11826   12.016  < 2e-16 ***
## interval60:prematurityvery preterm     2.28684    0.16781   13.627  < 2e-16 ***
## interval90:prematurityvery preterm     1.98443    0.19391   10.234  < 2e-16 ***
## interval120:prematurityvery preterm    1.66797    0.11606   14.372  < 2e-16 ***
## interval1:prematuritylater preterm     1.08511    0.08405   12.910  < 2e-16 ***
## interval7:prematuritylater preterm     1.17563    0.10613   11.077  < 2e-16 ***
```

```
## interval14:prematuritylater preterm    1.23599    0.10249   12.060  < 2e-16 ***
## interval28:prematuritylater preterm    1.95695    0.09419   20.777  < 2e-16 ***
## interval60:prematuritylater preterm    2.93275    0.14028   20.906  < 2e-16 ***
## interval90:prematuritylater preterm    2.83714    0.15195   18.672  < 2e-16 ***
## interval120:prematuritylater preterm   2.45756    0.09012   27.270  < 2e-16 ***
## interval1:prematurityfull-term         1.65409    0.07247   22.826  < 2e-16 ***
## interval7:prematurityfull-term         1.97168    0.08625   22.860  < 2e-16 ***
## interval14:prematurityfull-term        2.25484    0.08078   27.914  < 2e-16 ***
## interval28:prematurityfull-term        3.14635    0.07860   40.030  < 2e-16 ***
## interval60:prematurityfull-term        4.21514    0.12538   33.620  < 2e-16 ***
## interval90:prematurityfull-term        4.03640    0.13551   29.787  < 2e-16 ***
## interval120:prematurityfull-term       3.75711    0.07887   47.637  < 2e-16 ***
## raceNHW:prematurityvery preterm        0.10647    0.06414    1.660  0.09693 .
## raceNHW:prematuritylater preterm       0.13060    0.05504    2.373  0.01766 *
## raceNHW:prematurityfull-term          -0.25830    0.04630   -5.579 2.41e-08 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for poisson family taken to be 1)
##
##     Null deviance: 5426162574  on 12161  degrees of freedom
## Residual deviance:     321897  on 12118  degrees of freedom
## AIC: 356031
##
## Number of Fisher Scoring iterations: 12
```

Calculate the hazard of dying in the first interval (0-1 day) of extremely preterm babies born to NHB mothers.

```
## Hazard: extremelly preterm babies born to NHB mothers in first interval (0-1 days) =  0.2270261
```

1) *extremely preterm babies to NHW mothers compared to extremelly preterm babies to NHB mothers in the first interval (0-1 days)*

```
## H-R: extremelly preterm babies to NHW mothers compared to extremelly preterm
## babies to NHB mothers in the first interval (0-1 days) =  5.748148
```

2) *full-term babies to NHB mothers compared to extremelly preterm babies to NHB mothers in the first interval (0-1 days)*

```
## H-R: full-term babies to NHB mothers compared to extremelly preterm babies to
## NHB mothers in the first interval (0-1 days) =  1.006928
```

3) *full-term babies to NHB mothers compared to extremelly preterm babies to NHB mothers in the last interval (120-365 days)*

```
## H-R: full-term babies to NHB mothers compared to extremelly preterm babies to
## NHB mothers in the last interval (120-365 days) =  2569148
```

4) *full-term babies to NHW mothers compared to full-term babies to NHB mothers in the last interval (120-365 days)*

```
## H-R: full-term babies to NHW mothers compared to full-term babies to NHB mothers
## in the last interval (120-365 days) =  1.058653
```

(g) (10 points) Calculate the survival curve for extremely preterm babies to NHB mothers. Compare to the KM estimate from b) by plotting the two curves on the one graph. (Note: the fit should be fairly reasonable, so if it's not there could be an issue in your part f) model).

Calculating the Survival Function for *extremelly pre-term, NHB* babies we have the following:

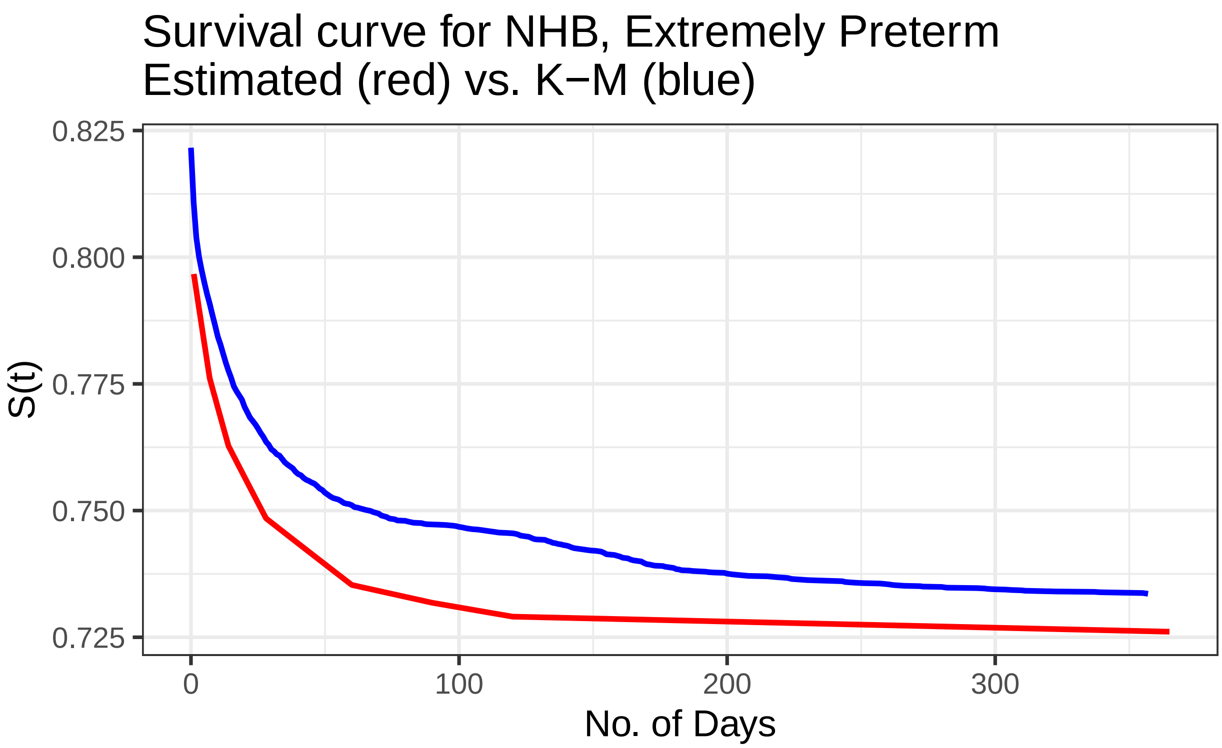

Figure 14: Kaplan-Meier Estimates of S(t) w/ CI