STA2201H Winter 2020 Assignment 2

Due: 11:59pm, 16 February 2020

What to hand in: .Rmd file and the compiled pdf.

How to hand in: Submit files via Quercus.

Please label your graphs with informative titles and axes.

1 Gompertz (40 points)

Gompertz hazards are of the form

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

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 $S(t) = \exp\left(-\frac{\alpha}{\beta}(e^{\beta t} - 1)\right)$ and $f(t) = \alpha \exp\left(\beta t - \frac{\alpha}{\beta}(e^{\beta t} - 1)\right)$.

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

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?

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?

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.

f) (10 points) Repeat part c) 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.

2 Infant mortality (60 points)

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) (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?

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.

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")`).

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?

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?

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:

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

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).