STA2201H Winter 2021 Assignment 1

Due: 5pm, 4 February 2021

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

How to hand in: Submit files via Quercus

1 Overdispersion

Suppose that the conditional distribution of outcome Y given an unobserved variable θ is Poisson, with a mean and variance θ, so

$$Y|\theta \sim \text{Poisson}(\mu\theta)$$

a) Assume $E(\theta) = 1$ and $Var(\theta) = \sigma^2$. Using the laws of total expectation and total variance, show $E(Y) = \mu$ and $Var(Y) = \mu(1 + \mu\sigma^2)$.
b) Assume θ is Gamma distributed with α and β as shape and scale parameters, respectively. Show the unconditional distribution of Y is Negative Binomial.
c) In order for $E(Y) = \mu$ and $Var(Y) = \mu(1 + \mu\sigma^2)$, what must α and β equal?

2 Opioid mortality in the US

The following questions relate to the `opioids` dataset, which you can find in the `data` folder of the repo. It's an RDS file, which you can read in using `read_rds` from the `tidyverse`. There is also a `opioids_codebook.txt` file which explains each of the variables in the dataset.

The data contains deaths due to opioids by US from 2008 to 2017. In addition, there are population counts and a few other variables of interest. The goal is to explore trends and patterns in opioid deaths over time and across geography. The outcome of interest is `deaths`.

Please make sure to clearly explain any findings or observations you make, rather than just handing in code and output. You will be assessed not only on the code but also on how you communicate your findings with a combination of writing and analysis.

a) Perform some exploratory data analysis (EDA) using this dataset, and briefly summarize in words, tables and charts your main observations. You may use whatever tools or packages you wish. You may want to explore the `geofacet` package, which plots US state facets in the correct geographic orientation.

b) Run a Poisson regression using `deaths` as the outcome and `tot_pop` as the offset. (remember to `log` the offset). Include the `state` variable as a factor and change the reference category to be Illinois. Investigate which variables to include, justifying based on your EDA in part a). Interpret your findings, including visualizations where appropriate. Include an analysis of which states, after accounting for other variables in the model, have the highest opioid mortality.

c) What's an issue with using population as an offset, given the limited information available in this dataset?

d) Rerun your Poisson regression using `expected_deaths` as an offset. How does this change the interpretation of your coefficients?

e) Investigate whether overdispersion is an issue in your current model.

f) If overdispersion is an issue, rerun your analysis using negative binomial regression. Does this change the significance of your explanatory variables? Do a Likelihood Ratio Test to see which is the preferred model.

g) Summarize your findings, giving the key insights into trends in opioid mortality over time and across states, and any factors that may be associated with these changes. What other variables may be of interest to investigate in future?

3 Gompertz

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) Find an expression in terms of α and β for the modal age at death.
b) For every year, estimate α, β and the mode age at death.
c) Create plots 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.

4 Infant mortality

In this part we will be looking at the infant mortality data set. This is in the `data` folder called `infant.RDS`.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.

The goal is to investigate differences in ages at death by race of mother and prematurity (from extremely preterm to full-term).

a) 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) 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. 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).

c) 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) 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) 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) 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. Note that you may want to investigate interaction terms. 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) Fit a piecewise hazards model to the whole population (i.e. just have `interval` as a covariate) and calculate the survival curve. Compare to the KM estimate from b) by plotting the two curves on the one graph. The fit should be fairly reasonable, so if it's not there could be an issue in your part f) model.