COMH 7247 – BAYESIAN METHODS IN HEALTH RESEARCH *ASSIGNMENT*

Upload on Ulwazi by 8:00am on of the 26th of April 2021.

Note: This will constitute 60% towards your final module mark.

1. **a.** The Gamma distribution with mean μ and variance μ^2/α has density function

$$f(y) = \frac{\alpha^{\alpha}}{\mu^{\alpha}\Gamma(\alpha)} y^{\alpha-1} e^{-\frac{y\alpha}{\mu}} \qquad (y > 0)$$

(i) Show that this may be written in the form of an exponential family i.e generalized linear model (GLM).

(ii) Use the properties of exponential families to confirm that the mean and variance of the distribution are μ and μ^2/α.

b. For each of the three distributions (Normal, Binomial and Poisson), given that the data follow a Y~.

(i) Show that Y belongs to the exponential family of distributions, specifying each component of the generalized linear model.

(ii) State the canonical link function on each of those cases.

(iii) For each of these distributions, verify the mean and variances

Note: For this question feel free to use the general formulae given in **Appendix A**.

2. The total amounts claimed each year from a portfolio of insurance policies over *n* years were $x_1, x_2, .., x_n$. The insurer believes that annual claims have a normal distribution with mean μ and variance σ^2 where μ is unknown. The prior distribution of μ is assumed to be normal with mean μ_0 and variance $\sigma^2{}_0$

(a) Derive the posterior distribution of μ

(b) Using the answer in (a), write down the Bayesian point estimate of μ under quadratic loss.

(c) Show that the answer in (b) can be expressed in the form of a credibility estimate and derive the credibility factor.

The claims experience over five years for two companies was as follows:

	Year	***1***	***2***	***3***	***4***	***5***
Company A	Amount	421	417	438	456	463
Company B	Amount	343	335	356	366	380

(d) Determine the Bayes credibility estimate of the premiums the insurer should charge for each company based on the modelling assumptions of part (a), a profit loading of 25% and the following parameters:

	Company A	***Company B***
μ_0	400	300
σ^2	500	350
$\sigma^2{}_0$	800	600

(e) Calculate the two posterior means and the two posterior variances using any Bayesian software of your choice using the data shown on (d) above.

3. The number, *X*, of claims on a given medical aid policy over one year has probability distribution given by

$$P(X = k) = \theta^k(1-\theta) \qquad\qquad k = 0, 1, 2, \ldots$$

where θ is an unknown parameter with $0 < \theta < 1$.

Independent observations $x_1, \ldots, x_n$ are available for the number of claims in the

previous *n* years. Prior beliefs about θ are described by a distribution with density

$$f(\theta) \propto \theta^{\alpha-1}(1-\theta)^{\alpha-1.} \qquad \text{for some constant } \alpha > 0\ .$$

(a) Derive the posterior distribution of θ given the data $x_1, \ldots, x_n$.

(b) Derive the Bayesian estimate of θ under quadratic loss and show that it takes the form of a credibility estimate

$$Z\hat{\theta} + (1-Z)\mu$$

where μ is a quantity you should specify from the prior distribution of θ.

(c) Explain what happens to Z as the number of years of observed data increases.

(d) Determine the variance of the prior distribution of θ.

(e) Explain the implication for the quality of prior information of increasing the value of α. Give an interpretation of the prior distribution in the special case $\alpha = 1$

(f) Calculate the Bayesian estimate of θ under quadratic loss if $n = 3$, $x_1 = 3, x_2 = 3, x_3 = 5$ and

(i) $\alpha = 5$

(ii) $\alpha = 2$

(g) Calculate the above question using any Bayesian modelling software of your choice

4. a. The truncated exponential distribution on the interval (0, *c*) is defined by the

b. Let $\mathbf{X_1, X_2, \ldots, X_n}$ be a random form a probability distribution

$$f(x) = 3x^2; \quad 0 \leq x \leq 1$$

Use the Monte Carlo Integration approximation to estimate P $(0.25 < x \leq 0.75)$. Give your answer to an error margin of 0.001.

5. For a simple linear regression, $y_i = \beta_0 + \beta_1 x_i + e_i$, where our interest is to estimate $\hat{\beta}_0, \hat{\beta}_1$, and $Var(y_i) = \hat{\varphi}$ using Bayesian methods through a Gibbs sampler simulation algorithm, let:

$$y \sim N(\beta_0 + \beta_1 x, \phi)$$

$$\beta_0 \sim N(m_0, \tau_0)$$

$$\beta_1 \sim N(m_1, \tau_1)$$

$$\phi \sim InvGamma(\alpha, \beta)$$

where the distributions for $\boldsymbol{\beta_0, \beta_1}$, and $\boldsymbol{\varphi}$ are prior distributions with all hyperparameters $\boldsymbol{m_0, m_1, \tau_0, \tau_1, \alpha, \beta}$ known and the $\boldsymbol{\tau}$ s are precision parameters.

a. Write the likelihood for $\boldsymbol{y}$

b. Write the full posterior for this model ie $\boldsymbol{P(\beta_0, \beta_1, \varphi \mid y, m_0, m_1, \tau_0, \tau_1, \alpha, \beta)}$ to some proportionality constant.

c. Derive the conditionals for $\boldsymbol{\beta_0, \beta_1, \varphi}$ and write a brief Gibbs sampling algorithm that you can implement to draw samples for estimation and inference for each of these parameters.

d. Use a small data with 5 **y** values and 5 **x** values to demonstrate this on estimating the parameters $\boldsymbol{\beta_0}$ *and* $\boldsymbol{\beta_1}$ using any Bayesian software of your choice.

6. The **case fatality rate** for COVID-19 cases is a very useful indicator of how the disease contributes to the deaths experienced in a given period of time at a given place.
Use this link below to import World wide based data from the web on the last day in December 2020 (data are also provided as **bayes2021_world_covid.dta**

import delimited "https://raw.githubusercontent.com/CSSEGISandData/COVID-19/master/csse_covid_19_data/csse_covid_19_daily_reports/12-31-2020.csv", clear

The overall aim is to model the case fatality rate for African countries only based data using **all three Poisson distributions** as detailed on the tutorial “Bayesian using Stata Day 1”. So you will need to perform some data management prior to undertaking your analyses.

COMH 7247 – BAYESIAN METHODS IN HEALTH RESEARCH *ASSIGNMENT*

Upload on Ulwazi by 8:00am on of the 26th of April 2021.

Note: This will constitute 60% towards your final module mark.

a) Clean out all the data that are not part of these 54 countries in Africa based on the this link https://www.worldometers.info/geography/how-many-countries-in-africa/.

b) Also include the data from the 54 countries on the given link to the so that you end up with the data for the Population in 2020 and the sub-region merged to the cleaned data on a) above. Make sure that you names the merging variables in exactly the same way to avoid missing any of the nations.

c) Write down the posterior distribution which arises from a Poisson regression model with 1 covariate. Derive the full conditional distributions of the parameters.

d) Use the given data on the COVID-19 cases to determine whether there are differences between sub-regions on the case fatality rates of the cases.

(i) The maximum likelihood estimation (MLE) approach was initially used and the Stata commands are shown on Appendix B, shown below. Comment on what you notice from the subregions data summaries and IRR results from the Poison regression.

(ii) Can you compare the **three Poisson types (Poisson, Zero truncated Poisson and Generalized Poisson)** of models discussed on the above mentioned tutorial, and choose your best fitting model using STATA **for a no covariates** model as was done on the tutorial. Clearly state how you arrived at choosing the best model

(iii) Compare at your Poisson model (only) done in STATA with what you get from OpenBugs/WinBugs.

(iv) Now include the covariate **subregion** to answer the question asked. Use only one Bayesian based software. ***Credit will be given for good modelling practice such as use of chains, burning, thinning, inspections of parameter convergence and good use of nuisance parameters***

(v) Interpret your final selected model from the Bayesian modelling.

NB: Submit all the documentation used to arrive at the final selected model electronically attach your main answer document separate alone under Part A on Ulwazi and the rest as support documents under Part B.

Type your equations using RMarkdown or Stata Markdown on the same document as your responses to your other answers.

Appendix A: For a random variable Y from the exponential family of distributions, with natural parameter θ and scale parameter ϕ:

$$f_Y(y,\theta,\phi) = \exp\left(\frac{y\theta - b(\theta)}{a(\phi)} + \mathrm{c}(y,\phi)\right);$$

mean: $E(Y) = b'(\theta)$ variance: $\mathrm{Var}(Y) = a(\phi)b''(\theta)$

Appendix B

```
. tabstat cases_fatality, by(subregion) stats(N mean sd)

. poisson cases_fatality i.subregion,  irr baselevels
```