Project 3
Computational Numerical Statistics
DM, FCT-UNL

NOVA SCHOOL OF SCIENCE & TECHNOLOGY

Deadline: 12/01/22

The writing of the project report, either in LaTeX or R Markdown, should detail all the theoretical aspects of the methods involved in this project and provide all the necessary graphical displays essential for the understanding of the particular cases. **From the three groups/topics below you should choose only two.**

The Exponential Family

1. Let $X \sim Weibull(\alpha, \beta)$, with β known and α unknown, which has p.d.f.

$$f(x;\alpha,\beta) = \frac{\alpha}{\beta}\Big(\frac{x}{\alpha}\Big)^{\beta-1} e^{-(\frac{x}{\alpha})^{\beta}}, \qquad \alpha,\beta>0, \qquad x \in \mathbb{R}^{+}.$$

   (a) Show that this distribution belongs to the exponential family.

   (b) Clearly identify the canonical link and the sufficient statistic. Do you already have the canonical form? If not, write it down.

   (c) Use the canonical form to
      i. compute $E(X^{\beta})$ and $V(X^{\beta})$
      ii. write the score function $S_n(\alpha)$ and see if it is possible to analytically derive the maximum likelihood estimator of α, α_{MLE}
      iii. compute the Fisher Information $I_n(\alpha)$
      iv. report the asymptotic variance of the maximum likelihood estimator α_{MLE}

2. Say some distribution depending on unknown parameters $(\alpha,\beta) \in \mathbb{R}^{+} \times \mathbb{R}^{+}$ has p.d.f. such that

$$f(x;\alpha,\beta) = \exp\Big\{ \sum_{i=1}^{2} \eta_i(\alpha,\beta) T_i(x) - A(\alpha,\beta) + c(x) \Big\}$$

   with

   - $\eta(\alpha,\beta) = (\eta_1(\alpha,\beta), \eta_2(\alpha,\beta)) = (\alpha, -\beta)$
   - $(T_1(x), T_2(x)) = (\log x, x)$
   - $A(\alpha,\beta) = -\alpha \log\beta + \log\Gamma(\alpha)$
   - $c(x) = -\log x$

   Use the canonical form of the p.d.f. to compute $E(X)$, $V(X)$ and $I_n(\alpha,\beta)$. Report the asymptotic variance of the maximum likelihood estimator $(\alpha_{MLE}, \beta_{MLE})$.

Generalized Linear Models

The class of generalized linear models also comprises the Poisson distribution, with the log link function being the mathematically convenient option as it allows the linear predictor to span the entire real line. The Poisson regression method is often employed for the statistical analysis of data that involve counts of events occurring within a certain amount of time. Such is the case, for example, of epidemiological studies that require the calculation of rates, typically rates of death or incidence rates of a chronic or acute disease. Here, the parameter of interest is usually the expected counts per unit of observed time, i.e., the rate at which events occur.

1. The following data refers to the number of new AIDs cases (y) each year (x) in Belgium, from 1981 to 1993 (Venables & Ripley, *Modern Applied Statistics with S*)

   ```
   12 14 33 50 67 74 123 141 165 204 253 246 240
   ```

   The question here is whether these data support evidence that the increase in the rate of new case generation is slowing down.

   You may use here the `R` built-in `glm()` function.

   (a) Fit a Poisson regression model to the data (Y, x), i.e., to $Y \sim x$. Report the model fit residual deviance and AIC, plot the residual plots for the fitted model and comment on model fit adequacy.

   (b) Fit a Poisson regression model again for the relationship $Y \sim x + x^2$. Report the model fit residual deviance and AIC, plot the residual plots for the fitted model and comment on model fit adequacy when compared to the previous model fit.

   (c) One now wishes to compare the previous two models by means of an analysis of variance table. Describe model selection via the ANOVA table. Use the `R` built-in `anova()` function to compare both models. Which model better fits the data?

   (d) Provide model summary, confidence intervals for the fixed parameters and model interpretation for the selected model. Plot the data. Predict 100 values from the fitted model (use the `R` built-in `predict()` function). Plot the data versus the fitted line. Add confidence bands for the fitted line at $\pm 2se$. Make an attempt at answering the main question.

2. Fully describe and implement the IRWLS for Poisson regression such that your implementation returns the same summary output table as the `R` `glm()` function. Use your implementation to fit the Poisson model to the previous dataset considering $Y \sim x$.

Bayesian Inference and Computation

The `janka` data frame, from the `R` package `SemiPar`, has 36 observations on Australian timber samples, which refer to measurements of the density (predictor variable) and hardness (response variable) of the timber.

1. Describe in detail, and fit from scratch, a Bayesian linear model to the `janka` data using the Gibbs sampler. Report point estimates, credible intervals and whatever more you feel is important regarding the analysis of these data.

2. Go to

   `https://cran.r-project.org/web/packages/bayestestR/vignettes/bayestestR.html`

   and install the `R` package `bayestestR`. Use the `stan_glm()` to fit the Bayesian linear model to the `janka` data and compare these results with the results from 1.. You can use the `rjags` library instead or other that you feel is more convenient for you.

3. Compare the Bayesian results (point estimates and credible intervals) with the results from the classical analysis performed in week 5, slides $12 - 14$, part II.