MATH 6020 – HW 2

Due back on Saturday, 13th March [1].

-1- Find the variance function, $V_Y(\lambda)$ in the following cases.

-a- $Y \sim$ Poisson(λ).

-b- $Y \sim$ Exponential(λ).

-2- Refer to the data file salmon.txt (Source: Johnson and Wichern, page 603). The salmon fishery is a valuable resource for both the US and Canada. Each country however, should ideally catch salmon that originated from its own waters. To help regulate catches, samples of fish are taken during harvest time and identified as originating from Alaskan or Canadian waters. The fish carry some information about their birthplace in the growth rings on their scales. Typically, the rings associated with freshwater growth are smaller for the Alaskan-born than for the Canadian-born salmon. The datafile is organized as follows:

- Column 1: (1 = Alaskan, 2 = Canadian).
- Column 2: Gender
- Column 3: Diameter of rings for first year freshwater fish (in hundredths of an inch).
- Column 4: Diameter of rings for first year marine fish (in hundredths of an inch).

Note: All answers to this question should be presented as part of a typed document.

-a- Suppose $Y \sim$ Bernoulli(p) represents the birthplace of the fish with $\mathrm{P}\{Y = 1\} = p$. Construct a logistic model to explain the variation in Y.

-b- Make a conditional density plot for the response variable with each of the continuous predictors and interpret what you see. Refer : `https://stat.ethz.ch/R-manual/R-patched/library/graphics/html/cdplot.html`.

-c- Write an algorithm to estimate a logistic model with the two continuous predictors. Construct a large sample 99% confidence interval for $\boldsymbol{\beta}$. Estimate the same model using the `glm` procedure in `R`. Perform a deviance analysis through a partial likelihood approach for an appropriate sequence of hypotheses. Write a short report of about 2 pages summarizing results.

-d- Make a scatter plots of residual deviances versus appropriate axes of your choice and write any observations you note.

-3- Suppose $Y_i \sim$ ind Bernoulli(p_i). Write its p.m.f.

-a- Suppose you have to estimate p_i with y_i. Write the joint p.m.f. for all n observations of Y. Call this term $L_{\text{max, reduced}}$.

-b- Suppose you have a parametric model to estimate p_i with the maximum likelihood estimators of the linear predictor. Call this $\hat{p}_i$. Write the joint p.m.f. for all n observations of Y with the estimator for p_i. Call this term $L_{\text{max, full}}$.

[1] HW Version: 2021-03-11 at 16:47

-c- Compute 2 times the log of the ratio of $L_{\text{max, reduced}}$ and $L_{\text{max, full}}$. What do you find that it equates to?

-4- $\mathbf{X}_{n\times q} = [\mathrm{X}_1, \ldots, \mathrm{X}_q]$ is a matrix of n random vectors with (population) $\text{VarCov}\{\mathbf{X}\} = \mathrm{E}\{\mathbf{X}'\mathbf{X}\} - \mathrm{E}'\{\mathbf{X}\}\mathrm{E}\{\mathbf{X}\} = \mathbf{\Sigma}_{q\times q}$. Let,

$\mathbf{x}_{n\times q} = [\mathrm{x}_1, \ldots, \mathrm{x}_q]$ be an observed sample of $\mathbf{X}$,

$\mathbf{S} = \left(\left(\frac{1}{n}\,\mathrm{x}_i'\,\mathrm{x}_j - \bar{\mathrm{x}}_i\bar{\mathrm{x}}_j\right)\right)_{i,j}$, the sample covariance matrix be the estimator of $\mathbf{\Sigma}$, and

$\mathbf{R}$ be the sample correlation matrix.

Suppose we standardize $\mathbf{x}$ as $\mathbf{x}^* = \left[\frac{\mathrm{x}_1 - \bar{\mathrm{x}}_1}{s_{\mathrm{x}_1}}, \ldots, \frac{\mathrm{x}_q - \bar{\mathrm{x}}_q}{s_{\mathrm{x}_q}}\right]$ where s_{x_i} is the sample standard deviation of the i^{th} column of $\mathbf{x}$. Show that the sample covariance matrix of $\mathbf{x}^*$ is $\mathbf{R}$.

-5- Suppose $\mathrm{X}_{q\times 1} \sim \mathrm{N}_q(\boldsymbol{\mu}, \mathbf{\Sigma})$. We know that $Y_i = (\mathrm{X} - \boldsymbol{\mu})'\,\mathbf{e}_i \sim \mathrm{N}(0, \lambda_i)$. Derive the distribution of $Y_i^* = \mathrm{X}^{*\prime}\,\mathbf{e}_i$ when X is standardized to form $\mathrm{X}^{*\prime} = (\mathrm{X} - \boldsymbol{\mu})'\,\mathbf{\Sigma}^{-\frac{1}{2}}$.

-6- Use the posted file `Galtons_Height_Data.csv` for this question. The dataset is sourced from `https://www.randomservices.org/random/data/Galton.html`.

-a- Subset the data by using only those parent pairs who have 2 or more kids. Amongst them use data that pertains only to the first daughter and son listed. Your data should then have the following variables:

X_1 = Father's Height.

X_2 = Mother's Height.

X_3 = Daughter's Height.

X_4 = Son's Height.

-b- Perform PCA on the standardized and un-standardized datasets and analyze the results thoroughly. Compare and contrast wherever relevant.

-7- Generate a dataset of 100 samples from a multivariate normal dataset with $\boldsymbol{\mu} = [0, 25]$, $\rho_{12} = -0.4$, $\sigma_{11} = 3$, and $\sigma_{22} = 6$. Construct a scatterplot of the two variables (mean centered and uncentered), overlay their principal components and a 95% confidence ellipse for the mean vectors in each case. Take any other arbitrary positive definite 2×2 correlation matrix and find its eigenvectors. What do yo find and why?

-8- In R use the command `data("urine", package="boot")` to download urine analysis data for 77 patients. You can find documentation on this data set at

`https://stat.ethz.ch/R-manual/R-devel/library/boot/html/urine.html`

Note that it consists of 79 rows and 7 columns. The first column is categorical (indicates if calcium oxalate crystals were found in the urine or not), while the remaining six are numerical. Note that rows 1 and 55 each have a missing entry in them (recorded as an NA). Delete these rows entirely.

-a- Perform PCA on each of the standardized versions of the two data sets and report all results. Interpret PCs, if able. Do a subset of the PCs appear to sufficiently capture the variation in each dataset? Do the two datasets appear to be substantially different? In other words, does the presence of calcium oxalate seem to significantly influence the correlation structure of the cloud of data points?

-b- Construct bi-plots for each dataset and interpret the results.