Homework Assignment 2

STAT 544

Due on February 26, 2021

1. Suppose we observe an iid sample from the BETA distribution with $\boldsymbol{\theta}_0 = (\alpha_0, \beta_0)' = (5, 2)'$, but we assume the data came from a KUMARASWAMY(a, b) distribution. Use KL divergence to estimate $\boldsymbol{\theta}^* = (a^*, b^*)'$, the true value of the parameter under the misspecified model. Show the BETA and KUMARASWAMY densities in the same plot. Comment.

2. Suppose we observe $\boldsymbol{Y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}$, where the errors are iid LAPLACE$(0, b)$ random variables (0 is the mean, $b \in \mathbb{R}^+$ is the scale). Find $\boldsymbol{\mathcal{I}}(\boldsymbol{\beta})$.

3. Recall that a nonparametric $(1-\alpha)100\%$ confidence band for $F(x)$ has lower and upper limits

$$
\begin{aligned}
L(x) &= \max\{\hat{F}_n(x) - \epsilon_n, 0\} \qquad (1)\\
U(x) &= \min\{\hat{F}_n(x) + \epsilon_n, 1\},
\end{aligned}
$$

where

$$\epsilon_n = \{(2n)^{-1} \log 2\alpha^{-1}\}^{1/2}.$$

In this problem you will compute a bootstrap confidence band for $F(x)$ for the nerve data and compare the results with (1).

   (a) R has a function called `ecdf` that computes $\hat{F}_n$. Familiarize yourself with this function.

   (b) To compute a bootstrap sample, we will pretend that $\hat{F}_n$ is the true cdf. We will simulate many datasets (1,000, say) from $\hat{F}_n$. For the kth simulated dataset, we will compute $\hat{F}_n^{(k)}$. We will then use sample quantiles for the $\hat{F}_n^{(k)}$ to construct our bootstrap confidence band.

   (c) We can sample from the empirical cdf by using R's `sample` function to draw a sample of size n, with replacement, from the original sample.

   (d) After you have computed $\hat{F}_n^{(1)}, \ldots, \hat{F}_n^{(n_b)}$, where n_b is the size of the bootstrap sample, use the `quantile` function to compute appropriate sample quantiles at each of the data points. Use $\alpha = 0.05$, and apply a Bonferroni correction to arrive at a confidence band (as opposed to pointwise confidence intervals).

(e) Plot $\hat{F}_n$ along with your bootstrap confidence band and the confidence band given by (1). Interpret the results. How sensitive is the bootstrap confidence band to the choice of n_b?

4. Before an election, a polling agency randomly samples $n = 100$ people to estimate π, the population proportion who prefer candidate A over candidate B. You estimate π by the sample proportion $\hat{\pi}$. I estimate π by $\frac{1}{2}\hat{\pi} + \frac{1}{2}(0.5)$. Which estimator is biased? For what range of π values does my estimator have smaller mean squared error? What does this problem illustrate? Explain what it means to say my estimator has a Bayesian flavor.

   (Mean squared error (MSE) is a commonly used measure of accuracy for estimators. It is given by $\text{MSE}(\hat{\theta}) = \mathbb{E}(\hat{\theta} - \theta)^2$ for estimator $\hat{\theta}$ and truth equal to θ. Note that MSE can be written as $\text{MSE}(\hat{\theta}) = \{\text{bias}(\hat{\theta})\}^2 + \mathbb{V}\hat{\theta}$.)

5. For x between 0 and 100, suppose the Gaussian linear model holds with

   $$\mathbb{E}Y = 45 + 0.1x + 0.0005x^2 + 0.0000005x^3 + 0.0000000005x^4 + 0.0000000000005x^5$$

   and $\sigma = 10$. Pseudorandomly generate 25 observations from the model, with x having a $\text{Uniform}(0, 100)$ distribution. Fit the simple model $\mathbb{E}Y = \beta_0 + \beta_1 x$ and true model $\mathbb{E}Y = \beta_0 + \beta_1 x + \cdots + \beta_5 x^5$. Create plots that show the data, the true relationship, and the model fits. For each model, measure the quality of the fit using the mean of $|\hat{\mu}_i - \mu_i|$. Summarize your findings and explain what this problem illustrates about model parsimony.

6. Suppose Y_i has a Poisson distribution with $g(\mu_i) = \beta_0 + \beta_1 x_i$, where $x_i = 1$ for $i = 1, \ldots, n_A$ from group A and $x_i = 0$ for $i = n_A + 1, \ldots, n_A + n_B$ from group B, with all observations being independent. Show that for any link function the GLM likelihood equations imply that the fitted means $\hat{\mu}_A$ and $\hat{\mu}_B$ equal the sample means.

7. The `MASS` package for R contains the `Boston` data file, which contains several predictors of the median value of owner-occupied homes for 506 neighborhoods in the suburbs of Boston, MA. Describe a model-building process for these data.