STA 215: Statistical Inference
HW 1 Due Wed Feb 1 2012

1. Survival analysis, a very important component of statistical applications to medicine, deals with *survival time* or *time to event* data related to various medical procedures. Typically one collects data (Y_i, z_i) on patients $i = 1, \cdots, n$, where Y_i's denote the survival times and z_i's are measurements (covariates) that are likely to influence Y_i's. A popular way to formulate this influence is the Cox proportional hazard model which we describe below, after introducing some notations.

   Notations: Let $f(t|z_i)$ denote the pdf of the survival time Y_i of a patient with covariate z_i. Define the survival and hazard functions $S(t|z_i) = \int_t^\infty f(y|z_i)dy$ and $h(t) = f(t|z_i)/S(t|z_i)$. Note that $S(t|z_i)$ gives the probability of surviving longer than time t and that $h(t|z_i) = \lim_{\delta \downarrow 0} P(Y_i < t + \delta | Y_i \geq t)/\delta$ gives the instantaneous failure rate at time t given survival up to that time. Note that $h(t|z_i) = -\frac{d}{dt} \log S(t|z_i)$ and so $S(t|z_i) = \exp\{-\int_0^t h(y|z_i)dy\}$.

   Cox proportional hazard model: In Cox proportional hazard model one takes

   $$h(t|z_i) = h_0(t) \exp\{g(z_i, \beta)\}$$

   where $h_0(t)$ is a *baseline hazard* and $g(z_i, \beta)$ is a known function except for the vector of coefficients β. A common choice is $g(z_i, \beta) = z_i'\beta$.

   In the following we will do some algebra to gain some insight about this model and represent it in a more standard regression form. Set $\Delta_i = \exp\{g(z_i, \beta)\}$. The base line hazard $h_0(t)$ gives rise to a baseline survival function $S_0(t) = \exp\{-\int_0^t h_0(y)dy\}$ and the associated pdf $f_0(t) = -\frac{d}{dt} S_0(t)$. Let T denote a random survival time with this pdf.

   (a) Show that $S(t|z_i) = S_0(t)^{\Delta_i}$.

   (b) Suppose $h_0(t)$ is strictly positive for all $t > 0$ and define $\tilde{Y}_i = -\log S_0(Y_i)$. Show that $\tilde{Y}_i$ has an exponential distribution with rate Δ_i. [Hint: $h_0(t) > 0$ means $S_0(t)$ is strictly decreasing.]

   (c) Find a monotone transformation $Q(t)$ such that $Y_i^* = Q(Y_i)$ can be written as $Y_i^* = g(z_i, \beta) + \epsilon_i$ where ϵ_i's are independently distributed according to a known distribution free of h_0 and β. Describe the distribution of ϵ_i's. [Hint: $X \sim Ex(\lambda)$ means $X = X_0/\lambda$ where $X_0 \sim Ex(1)$.]

2. Annual TC counts $X_1, \cdots, X_n$ from n consecutive years are modeled as $X_t \overset{\text{IND}}{\sim} Poi(\mu_t)$, $\mu_t = \alpha\beta^{t-1}$, $t = 1, \cdots, n$ with model parameters $\alpha \in (0, \infty)$ and $\beta \in (0, \infty)$. We are interested in β which captures whether the annuals counts are trending upward ($\beta > 1$), downward ($\beta < 1$) or staying flat ($\beta = 1$).

   (a) For observed data $x = (x_1, \cdots, x_n)$, derive neat formulas for $L_x(\alpha, \beta)$ and $\ell_x(\alpha, \beta)$ and show that $u_1 = \sum_{t=1}^n x_t$ and $u_2 = \sum_{t=1}^n t x_t$ form a pair of sufficient statistics.

   (b) Show that the profile likelihood $L_x^*(\beta)$ of β equals

   $$L_x^*(\beta) = \text{const} \times \beta^{u_2 - u_1} A(\beta)^{-u_1}$$

   where $A(\beta) = (\beta^n - 1)/(\beta - 1)$ for $\beta \neq 0$ and $A(\beta) = n$ for $\beta = 0$.

(c) Consider the prior pdf $\pi(\alpha, \beta)$ under which $\alpha \sim Ga(a_0, b_0)$ and is independent of β. Show that the integrated likelihood $\tilde{L}_x(\beta)$ equals

$$\tilde{L}_x(\beta) = \text{const} \times \beta^{u_2 - u_1} \{b_0 + A(\beta)\}^{-(u_1 + a_0)}.$$

(d) Make a graphical comparison of these two likelihood functions for data with $n = 100$, $u_1 = 932$, $u_2 = 51884$, where we choose $a_0 = b_0 = 1$.

3. A machine goes through 4 hazard levels θ, coded 0 through 3 (from low hazard to high hazard) with use over time. The hazard level can be measured by frequency of hazardous incidents X, again coded 0 through 3 (low frequency to high frequency). Suppose X is modeled with pmfs $f(x|\theta)$, $\theta \in \Theta = \{0, 1, 2, 3\}$ as given by the rows of the following table.

θ	$f(0\|\theta)$	$f(1\|\theta)$	$f(2\|\theta)$	$f(3\|\theta)$
0	$\frac{4}{10}$	$\frac{3}{10}$	$\frac{2}{10}$	$\frac{1}{10}$
1	0	$\frac{3}{6}$	$\frac{2}{6}$	$\frac{1}{6}$
2	0	0	$\frac{2}{3}$	$\frac{1}{3}$
3	0	0	0	1

(a) Give a simple expression for the MLE $\hat{\theta}_{\text{MLE}}(x)$ in terms of $x \in \{0, 1, 2, 3\}$.

(b) Describe the ML set $A_{1/2}(x) = \{\theta \in \Theta : L_x(\theta) \geq \frac{1}{2} L_x(\hat{\theta}_{\text{MLE}}(x))\}$ (list all elements of the set) for each of $x = 0, 1, 2, 3$.

(c) For the discrete uniform prior pmf $\pi(\theta)$ with $\pi(0) = \pi(1) = \pi(2) = \pi(3) = 1/4$, describe the posterior pmf $\pi(\theta|x)$ for each $x = 0, 1, 2, 3$.

4. In an opinion poll study, the number of supporters X among n college students is modeled as $X \sim Bin(n, p)$, $p \in [0, 1]$ and suppose the actual observation is $X = x$.

(a) Find expressions for $\hat{p}_{\text{MLE}}(x)$, I_x and $B_c(x)$ for an arbitrary $c > 0$.

(b) Show that for any prior pdf $\pi(p) = Be(a, b)$ with $a, b > 0$ the posterior pdf is $\pi(p|x) = Be(a'(x), b'(x))$ for some $a'(x), b'(x)$ and give formulas for these quantities. Give expressions of $c(x), d(x)$ where $N(c(x), d^2(x))$ is the Laplace approximation of $\pi(p|x)$.

(c) For $n = 50, x = 20$, calculate the ML interval $B_{1.96}(x)$. Also calculate 95% posterior interval based on prior $Be(1, 1)$ and give its Laplace approximation.

5. Annual TC counts $X_1, \cdots, X_n$ from n consecutive years are modeled as $X_i \overset{\text{IID}}{\sim} Poi(\mu)$, $\mu \in (0, \infty)$. Let $x = (x_1, \cdots, x_n)$ be the observed data.

(a) Find expressions for $\hat{\mu}_{\text{MLE}}(x)$, I_x and $B_c(x)$ for an arbitrary $c > 0$.

(b) Show that for any prior pdf $\pi(\mu) = Ga(a, b)$ with $a, b > 0$ the posterior pdf os $\pi(\mu|x) = Ga(a'(x), b'(x))$ for some $a'(x), b[(x)$ and give formulas for these quantities. Give expressions of $c(x), d(x)$ where $N(c(x), d^2(x))$ is the Laplace approximation of $\pi(\mu|x)$.

(c) For $x = (12, 14, 15, 12, 16, 14, 27, 10, 14, 16)$, calculate the ML interval $B_{1.96}(x)$. Also calculate 95% posterior interval based on prior $Ga(1, 1)$ and give its Laplace approximation.

6. Time intervals (in minutes) $X_1, \cdots, X_n$ between n successive eruptions of a geyser are modeled as $X_i \overset{\text{IID}}{\sim} Ex(\lambda)$, $\lambda \in (0, \infty)$. Let $x = (x_1, \cdots, x_n)$ be the observed data.

(a) Find expressions for $\hat{\lambda}_{\text{MLE}}(x)$, I_x and $B_c(x)$ for an arbitrary $c > 0$.

(b) Show that for any prior pdf $\pi(\mu) = Ga(a, b)$ with $a > 0, b > 0$, the posterior pdf os $\pi(\mu|x) = Ga(a'(x), b'(x))$ for some $a'(x), b[(x)$ and give formulas for these quantities. Give expressions of $c(x), d(x)$ where $N(c(x), d^2(x))$ is the Laplace approximation of $\pi(\lambda|x)$.

(c) For $x = (79, 54, 74, 62, 85, 55, 88, 85, 51, 85)$, calculate the ML interval $B_{1.96}(x)$. Also calculate 95% posterior interval based on prior $Ga(1, 1)$ and give its Laplace approximation.

7. Smile durations (in seconds) $X_1, \cdots, X_n$ of an eight week old baby are modeled as $X_i \overset{\text{IID}}{\sim} \mathit{Unif}(0, \theta)$, $\theta \in (0, \infty)$.

(a) Find expressions for $\hat{\theta}_{\text{MLE}}(x)$ and $A_k(x)$ for an arbitrary $k \in [0, 1]$.

(b) Show that for any prior pdf $\pi(\theta) = Pa(a, b)$, $a > 0, b > 0$ the posterior pdf is $\pi(\theta|x) = Pa(a'(x), b'(x))$ for some $a'(x), b'(x)$ and find expressions for these quantities.

(c) For $x = (10.4, 19.6, 12.8, 14.8, 1.3, 0.7, 5.8, 6.9, 8.9, 9.4)$, calculate the ML interval $A_{0.05}(x)$. Also calculate the 95% posterior interval based on prior $Pa(2, 15)$.

8. Average LSAT and GPA scores $(X_1, Y_1), \cdots, (X_n, Y_n)$ from n law colleges are modeled as $(X_i, Y_i) \overset{\text{IID}}{\sim} BvN(\mu_1, \mu_2, \sigma_1^2, \sigma_2^2, \rho)$, $\mu_1, \mu_2 \in (-\infty, \infty)$, $\sigma_1^2, \sigma_2^2 \in (0, \infty)$, $\rho \in (-1, 1)$, where the bivariate normal distribution $BvN(\mu_1, \mu_2, \sigma_1^2, \sigma_2^2, \rho)$ has pdf

$$f(x, y) = \frac{1}{2\pi\sigma_1\sigma_2\sqrt{1-\rho^2}} \exp\left[-\frac{1}{2(1-\rho^2)}\left\{\frac{(x-\mu_1)^2}{\sigma_1^2} + \frac{(y-\mu_2)^2}{\sigma_2^2} - 2\rho\frac{(x-\mu_1)(y-\mu_2)}{\sigma_1\sigma_2}\right\}\right]$$

over $x, y \in (-\infty, \infty)$.

(a) Find expressions for the MLE of $(\mu_1, \mu_2, \sigma_1^2, \sigma_2^2, \rho)$.

(b) Write down an expression for the profile log-likelihood $\ell_x^*(\rho) = \log L_x^*(\rho)$ of ρ.

(c) Give (an approximate) expression for $B_c(x) = \{\rho : \ell_x^*(\rho) \geq \max_{\rho'} \ell_x^*(\rho') - c^2/2\}$ for an arbitrary $c > 0$. Evaluate this interval for $c = 1.96$ and data

College	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15
LSAT	576	635	558	578	666	580	555	661	651	605	653	575	545	572	594
GPA	339	330	281	303	344	307	300	343	336	313	312	274	276	288	296