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 \(h(t|z_i) = \lim_{\delta \to 0} P(Y_i < t + \delta | Y_i \geq t)/\delta\) 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 \(S(t|z_i) = \exp\{-\int_t^0 h(y|z_i)dy\}\).

**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 \(\beta\). 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 baseline 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 \(\Delta_i\). [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 \(\epsilon_i\)'s are independently distributed according to a known distribution free of \(h_0\) and \(\beta\). Describe the distribution of \(\epsilon_i\)'s. [Hint: \(X \sim \text{Exponential}(\lambda)\) means \(X = X_0/\lambda\) where \(X_0 \sim \text{Exponential}(1)\).]

2. A machine goes through 4 hazard levels \(\theta\), 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.

| \(\theta\) | \(f(0|\theta)\) | \(f(1|\theta)\) | \(f(2|\theta)\) | \(f(3|\theta)\) |
|----------------|----------------|----------------|----------------|----------------|
| 0               | \(\frac{4}{10}\) | \(\frac{3}{10}\) | \(\frac{2}{10}\) | \(\frac{1}{10}\) |
| 1               | 0               | \(\frac{2}{6}\) | \(\frac{2}{6}\) | \(\frac{1}{6}\) |
| 2               | 0               | 0              | \(\frac{2}{3}\) | \(\frac{1}{3}\) |
| 3               | 0               | 0              | 0              | 1              |
For observation $X = 2$, what is the ML p-value for testing $H_0 : \theta \in \{0, 1\}$ vs. $H_1 : \theta \in \{2, 3\}$?

3. Let $X = (X_1, \ldots, X_n)$ be modeled as $X_i \overset{\text{iid}}{\sim} \text{Normal}(\mu, \sigma^2)$, with both parameters unknown: $\mu \in \mathbb{R}$, $\sigma^2 > 0$.

   (a) Show that the profile log-likelihood function $\ell^*_{X}(\mu) := \max_{\sigma^2 > 0} \ell_{X}(\mu, \sigma^2)$ equals
   
   $$
   \ell^*_{X}(\mu) = \text{const} - \frac{n}{2} \log \left\{ \frac{n-1}{n} s^2_{X} + (\bar{X} - \mu)^2 \right\},
   $$
   
   where $\bar{X} = \frac{1}{n} \sum_{i=1}^{n} X_i$ and $s^2_{X} = \frac{1}{n-1} \sum_{i=1}^{n} (X_i - \bar{X})^2$.

   (b) Give the exact rule\(^{1}\) of the size-\(\alpha\) ML test for testing $H_0 : \mu = \mu_0$ vs. $H_1 : \mu \neq \mu_0$ based on $\bar{X}$, $s^2_{X}$.

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

   (a) Show that the model on $X = (X_1, \ldots, X_n)$ is an exponential family model, i.e., show that the joint pmf of $X$ can be written as $f(x|\alpha, \beta) = h(x) \exp\{\eta(\alpha, \beta)T(x) - B(\alpha, \beta)\}$ for some functions $h, \eta, T$ and $B$. Give formulas for these functions. You should be able to write it in a way so that $\dim(T(X)) = 2$, i.e., $T(X) = (T_1(X), T_2(X))^T$ for two scalar statistics $T_1(X)$ and $T_2(X)$.

   (b) Show that the profile likelihood $L^*_{X}(\beta) := \max_{\alpha > 0} L_X(\alpha, \beta)$ of $\beta$ equals
   
   $$
   L^*_{X}(\beta) = \text{const} \times \beta^{T_2(X)} A(\beta)^{-T_1(X)}, \quad \beta > 0
   $$
   
   where $A(\beta) = (\beta^a - 1)/(\beta - 1)$ for $\beta \neq 1$ and $A(\beta) = n$ for $\beta = 1$.

   (c) For this model, there is no closed form expression for the MLE, but it can be solved numerically. Set up the likelihood equations (i.e., first order conditions) in $\alpha$ and $\beta$ and simplify.

   (d) For observed data, an R-optimization routine produces $(\hat{\alpha}_{\text{MLE}}, \hat{\beta}_{\text{MLE}}) = (6.738017, 1.006264)$ with
   
   $$
   \vec{\ell}_{X}(\hat{\alpha}_{\text{MLE}}, \hat{\beta}_{\text{MLE}}) = \left( \begin{array}{c}
   -20.52824 \\
   -7518.775
   \end{array} \right)
   $$
   
   Calculate the (approximate) ML p-value for testing $H_0 : \beta = 1$ vs. $H_1 : \beta \neq 1$.

5. Time intervals (in minutes) $X_1, \ldots, X_n$ between successive eruptions of a geyser are modeled as $X_i \overset{\text{iid}}{\sim} \text{Exponential}(\lambda)$, $\lambda > 0$.

   (a) Give expressions for $\hat{\lambda}_{\text{MLE}}(X)$ and $I_{\text{OBS}}(X)$.

   (b) For observed data with $n = 272$ and $\bar{X} = 70.9$, calculate the p-value for testing $H_0$ : “median\(^{2}\) interval time $\geq 60$ minutes” vs. $H_1$ : “median interval time $< 60$ minutes”.

6. Eruption durations (in minutes) $X_1, \ldots, X_n$ of a geyser are modeled as $X_i \overset{\text{iid}}{\sim} \text{Uniform}(0, \theta)$, $\theta \in (0, \infty)$.

\(^{1}\)recall that for $X_i \overset{\text{iid}}{\sim} \text{Normal}(\mu, \sigma^2)$, $(\bar{X} - \mu)/s_{X} \sim t(n-1)$, the Student-t distribution with $n - 1$ degrees of freedom.

\(^{2}\)The median of a distribution on the real line is the number $m$ such that the distribution puts half of its total probability mass to the left of $m$ and the other to the right.
(a) Show that any ML test for $H_0 : \theta \geq \theta_0$ vs. $H_1 : \theta < \theta_0$ is given by “reject $H_0$ if $X_{\text{max}}^{-1} \geq c$” where $X_{\text{max}} = \max(X_1, \ldots, X_n)$.

(b) For observed data with $n = 272$ and $X_{\text{max}} = 5.1$, calculate the ML p-value for testing $H_0 : \theta \geq 6$ vs $H_1 : \theta < 6$.

7. Let $X_i$ be the number of arrivals at a counseling service counter on the $i$th of a sequence of $n$ days. A possible model for these data is to assume that customers arrive according to a homogeneous Poisson process and hence, the $X_i$ are a sample from a Poisson distribution with unknown parameter $\theta > 0$, the expected number of arrivals per day. Suppose a real number $\theta_0$ has been determined such that if $\theta \leq \theta_0$ then it is not worth keeping the counter open.

(a) Exhibit an optimal (UMP) test statistic for $H_0 : \theta \leq \theta_0$ vs. $H_1 : \theta > \theta_0$. Your test statistic should be as simplified as possible.

(b) For what levels can you exhibit a UMP test?

(c) You want to ensure that if the arrival rate is $\leq 10$ the probability of your deciding to stay open is $\leq 0.01$, but if the arrival rate is $\geq 15$, the probability of your deciding to close is also $\leq 0.01$. How many days must you observe to ensure that the UMP test achieves this? (You could use large sample approximations.)

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

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

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

$$
\ell^*_X(\rho) := \max_{\mu_1, \mu_2, \sigma_1^2, \sigma_2^2} \ell_X(\mu_1, \mu_2, \sigma_1^2, \sigma_2^2, \rho)
$$

and show that any ML test for $H_0 : |\rho| \leq 0.5$ vs $H_1 : |\rho| > 0.5$ is of the form “reject $H_0$ if $|\hat{\rho}_{\text{MLE}}| \geq c$.”

(c) It is known that $\hat{\rho}_{\text{MLE}} \sim AN(\rho, (1-\rho^2)/n)$. What is the p-value for the above $H_0$ based on the following observed data? Give details of how you calculated the p-value (and in particular how you handled the maximization over $\Theta_0$. I don’t need a clean mathematical proof, but basic calculations and a few graphs will be helpful.)

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