STA 215: Statistical Inference

HW 2 Due Wed Feb 15 2012

1. For the Gaussian linear model $Y \sim N_n(Z\beta, \sigma^2 I_n)$, $\beta \in \mathbb{R}^p$, $\sigma^2 > 0$, with log-likeliihood function

$$\ell_y(\beta, \sigma^2) = \text{const} - \frac{n}{2}\log\sigma^2 - \frac{(y - Z\beta)^T(y - Z\beta)}{2\sigma^2}$$

we calculated the log-profile-likelihood in $\beta$ to be

$$\ell_y^\dagger(\beta) = \text{const} - \frac{n}{2}\log\left\{1 + \frac{(\beta - \hat\beta_{\text{LS}})^T(Z^TZ)(\beta - \hat\beta_{\text{LS}})}{(n-p)s^2_{y|z}}\right\}$$

and said "some additional calculations" show that the log-profile likelihood in $\eta = A^T\beta$ equals

$$\ell_y^*(\eta) = \text{const} - \frac{n}{2}\log\left\{1 + \frac{(\eta - \hat\eta_{\text{LS}})^T\{A^T(Z^TZ)^{-1}A\}^{-1}(\eta - \hat\eta_{\text{LS}})}{(n-p)s^2_{y|z}}\right\},$$

where $\hat\eta_{\text{LS}} = A^T\hat\beta_{\text{LS}}$. This homework problem is about the additional calculations needed to establish the last expression.

(a) Consider the function

$$g(\beta) = -(\beta - \hat\beta_{\text{LS}})^T(Z^TZ)(\beta - \hat\beta_{\text{LS}})$$

over $\beta \in \mathbb{R}^p$ and define a new function $h(\eta) = \max_{\beta: A^T\beta = \eta} g(\beta)$ over $\eta \in \mathbb{R}^r$. Show that

$$h(\eta) = -(\eta - \hat\eta_{\text{LS}})^T\{A^T(Z^TZ)^{-1}A\}^{-1}(\eta - \hat\eta_{\text{LS}}).$$

[Hint: Fix an $\eta \in \mathbb{R}^r$. To find $h(\eta)$ we need to maximize $g(\beta)$ subject to the constraint $A^T\beta - \eta = 0$. So set up the Lagrange function $\tilde g(\beta, \lambda) = g(\beta) + \lambda^T(A^T\beta - \eta)$ in $\beta \in \mathbb{R}^p, \lambda \in \mathbb{R}^r$. Find $\hat\beta$ and $\hat\lambda$ that solve the system:

$$0 = \frac{\partial}{\partial\beta}\tilde g(\beta, \lambda) = \frac{\partial}{\partial\beta}g(\beta) + \lambda A = -2(Z^TZ)(\beta - \hat\beta_{\text{LS}}) + A\lambda$$
$$0 = \frac{\partial}{\partial\lambda}\tilde g(\beta, \lambda) = A^T\beta - \eta$$

and evaluate $h(\eta) = g(\hat\beta)$. To solve the system, note that the second equation merely imposes $A^T\beta = \eta$. Now premultiply both sides of the first equation by $A^T(Z^TZ)^{-1}$ to find the solution $\hat\lambda$. Substitute the expression of $\hat\lambda$ in the first equation and then solve in $\beta$.]

(b) Now suppose for some monotonically increasing function $f(x)$ on $x \geq 0$ we have

$$\ell^\dagger(\beta) = -f((\beta - \hat\beta_{\text{LS}})^T(Z^TZ)(\beta - \hat\beta_{\text{LS}})).$$

Show that $\ell^*(\eta) = \max_{\beta: A^T\beta = \eta} \ell^\dagger(\beta)$ equals

$$\ell^*(\eta) = -f\left((\eta - \hat\eta_{\text{LS}})^T\{A^T(Z^TZ)^{-1}A\}(\eta - \hat\eta_{\text{LS}})\right).$$

2. For Gaussian linear model $Y \sim N_n(Z\beta, \sigma^2 I_n)$, $(\beta, \sigma^2) \in \mathbb{R}^p \times (0, \infty)$, consider testing $H_0 : a^T\beta \le \eta_0$ for a given number $\eta_0$ and a given non-zero vector $a \in \mathbb{R}^p$. Any ML test can be characterized as "reject $H_0$ if $\eta_0 \le a^T\hat{\beta}_{\text{LS}} - cs_{y|z}/\sqrt{n_a}$" for some fixed cutoff $c > 0$. Show that such an ML test has size exactly equal to $1 - \Phi_{n-p}(c)$ where $\Phi_{n-p}$ is the CDF of $t(n-p)$.

3. In an experiment, $n_1 = 12$ infant rats were assigned to a high protein diet while $n_2 = 7$ rats were assigned to a regular diet. For each rat, body weight gain between 28th and 84th days after birth were recorded. Let $U_1, \cdots, U_{n_1}$ denote these measurements for the high protein group and $V_1, \cdots, V_{n_2}$ denote the same for the regular diet group. Consider the model $U_i \overset{\text{IID}}{\sim} N(\mu_1, \sigma^2)$, $V_j \overset{\text{IID}}{\sim} N(\mu_2, \sigma^2)$, $U_i$'s, $V_j$'s are independent, with model parameters $\mu_1, \mu_2 \in (-\infty, \infty)$ and $\sigma^2 > 0$.

   (a) Recall that Gaussian linear model representation of this data: $Y \sim N_n(Z\beta, \sigma^2 I_n)$, $n = n_1 + n_2$, $p = 2$, $\beta = (\mu_1, \mu_2)^T$.

$$Y = \begin{pmatrix} U_1 \\ \vdots \\ U_{n_1} \\ V_1 \\ \vdots \\ V_{n_2} \end{pmatrix}, \quad Z = \begin{pmatrix} 1 & 0 \\ \vdots & \vdots \\ 1 & 0 \\ 0 & 1 \\ \vdots & \vdots \\ 0 & 1 \end{pmatrix} \begin{matrix} \updownarrow n_1 \\ \\ \updownarrow n_2 \end{matrix}$$

   For an observation $U = u$, $V = v$, give neat expressions for $Z^T Z$, $\hat{\beta}_{\text{LS}}$ and $s^2_{y|z}$ in terms of $n_1, n_2$, $\bar{u}$, $\bar{v}$, $s^2_u$ and $s^2_v$ (the sample sizes, means and variances of the two groups).

   (b) Use the Gaussian linear model theory to derive a neat expression for a $100(1-\alpha)\%$ ML confidence interval for $\eta = \mu_1 - \mu_2$. [Show work]

   (c) Calculate the (ML) p-value for testing $H_0 : \mu_1 = \mu_2$ with the following data:

| Diet | Weight gain (grams) |
|---|---|
| High | 134 146 104 119 124 161 107 83 113 129 97 123 |
| Low | 70 118 101 85 107 132 94 |

   (d) For the same data calculate the (ML) p-value for testing $H_0 : \mu_1 \le \mu_2$. [Notice the choice of the inequality. The "status quo" is that the high protein diet is no better than the regular diet in terms of facilitating weight gain.]

4. 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 $\beta$ which captures whether the annuals counts are trending upward ($\beta > 1$), downward ($\beta < 1$) or staying flat ($\beta = 1$).

   (a) In HW 1 we derived the profile log-likelihood of $\beta$ to be

$$\ell^*_x(\beta) = \text{const} + (u_2 - u_1)\log\beta - u_1 \log A(\beta)$$

   where $u_1 = \sum_{t=1}^n x_t$, $u_2 = \sum_{t=1}^n t x_t$ and $A(\beta) = (\beta^n - 1)/(\beta - 1)$ for $\beta \ne 1$ and $A(\beta) = n$ for $\beta = 1$. For $n = 100$, $u_1 = 932$ and $u_2 = 51884$, this function is maximized at $\hat{\beta} = 1.006264$ with $\ddot{\ell}^*_x(\hat{\beta}) = -752132.7$ [I did this on `R` by using `optimize()` and `hessian()`]. Calculate an approximate 95% ML confidence interval for $\beta$.

(b) Calculate the (approximate ML) p-value for testing $H_0 : \beta = 1$. Do the same for testing $H_0 : \beta \leq 1$. [Again, notice the inequality, "status quo" is no increase in TC activity]

(c) The original log-likelihood function in both parameters is given by

$$\ell_x(\alpha, \beta) = \text{const} - \alpha A(\beta) + u_1 \log \alpha + (u_2 - u_1) \log(\beta).$$

This function is maximized at $(\hat{\alpha}, \hat{\beta}) = (6.738017, 1.006264)$ with

$$\ddot{\ell}_x(\hat{\alpha}, \hat{\beta}) = \begin{pmatrix} -20.52824 & -7518.775 \\ -7518.77491 & -3508375.969 \end{pmatrix}.$$

Repeat the two (approximate ML) p-value calculations based on this information and compare your answers with those in part (c).

5. 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 | $\frac{3}{6}$ | $\frac{2}{6}$ | $\frac{1}{6}$ |
| 2 | 0 | 0 | $\frac{2}{3}$ | $\frac{1}{3}$ |
| 3 | 0 | 0 | 0 | 1 |

(a) For the ML interval $A_{1/2}(x) = \{\theta \in \Theta : L_x(\theta) \geq \frac{1}{2} L_x(\hat{\theta}_{\text{MLE}}(x))\}$, calculate the coverage $\gamma(\theta, A_{1/2})$ at each $\theta \in \{0, 1, 2, 3\}$

(b) What is the confidence coefficient of $A_{1/2}$?

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

(a) For any arbitrary $k \in [0, 1]$, express the confidence coefficient of the ML set $A_k = \{\theta : L_x(\theta) \geq k \times L_x(\hat{\theta}_{\text{MLE}}(x))\}$ as a simple function of $k$.

(b) Calculate the 95% ML confidence interval for $\theta$ when observed data are $x$ =(10.4, 19.6, 12.8, 14.8, 1.3, 0.7, 5.8, 6.9, 8.9, 9.4).

(c) For the same data, find $\alpha \in (0, 1)$ such that the $100(1 - \alpha)$% ML confidence interval for $\theta$ has $\theta = 30$ on its boundary.

(d) For the same data, does the $\alpha$ from (c) give the (ML) p-value for testing $H_0 : \theta = 30$? If you answer "yes" explain why this is so. If you answer "no" give the correct p-value.