Ph. D. Qualifying Exam
Thursday, August 24, 2006

Please submit solutions to at most **seven** problems. You have four hours. No one is expected to answer all the problems correctly. Partial credit will be given. All problems are worth an equal amount of credit.

Put your solution to each problem on a separate sheet of paper.

Applied Statistics

Problem 1. Consider a 2^{5-2} fractional factorial design with generators **I=1234** and **I=135**. After analyzing the results from this design the researcher decides to perform a second 2^{5-2} design exactly the same as the first but with signs changed in column **3** of the design matrix.

(a) What is the defining relation of the first design? What is the resolution of the first design? List the aliases of main effects in this design.

(b) What is the defining relation of the second design? What is the resolution of the second design? List the aliases of two-term interactions of factors **1, 2,** and **3** in this design.

(c) What is the defining relation of the combined design (designs 1 and 2 together)? What is the resolution of the combined design? (Hints: combine the defining relations of the two designs together and remove all the defining words that have both"+" and "−" signs.)

(d) Can you modify the first and second designs such that the combined design has a higher resolution?

Problem 2. Consider the logistic regression model

$$\pi(x) \quad = \quad \frac{\exp(\alpha+\beta x)}{1+\exp(\alpha+\beta x)}, \qquad -\infty < x < \infty. \tag{1}$$

(a) For given constants α and $\beta > 0$, show that $\pi(x)$ is a distribution function. What is the density function of this distribution function?

(b) Show that the logistic density function

$$f(x) = \frac{\exp(x)}{[1+\exp(x)]^2}, \qquad -\infty < x < \infty$$

is symmetrical about zero. Based on this fact, derive the expectation of a random variable X that has $\pi(x)$ ($\beta > 0$) as its distribution function.

Problem 3. Consider the following multiple linear regression model based on centered data

$$Y_i - \bar{Y} = \beta_1(X_{1i} - \bar{X}_1) + \cdots + \beta_p(X_{pi} - \bar{X}_p) + e_i, \quad i = 1, \ldots, n. \quad (2)$$

where $\{e_1, \ldots, e_n\}$ are iid normally distributed with mean zero and variance σ^2.

(a) Derive the least squares estimates of $(\beta_1, \ldots, \beta_p)$ in model (2) using matrix notation. What is the least squares estimate of the variance σ^2?

(b) When the p predictors are orthogonal, i.e.,

$$\sum_{i=1}^{n}(X_{ji} - \bar{X}_j)(X_{ki} - \bar{X}_k) = 0, \quad \text{for} \quad k \neq j.$$

show that the least squares estimates of $(\beta_1, \ldots, \beta_p)$ based on model (2) are the same as those based on the simple linear regression models:

$$Y_i - \bar{Y} = \beta_j(X_{ji} - \bar{X}_j) + e_i, \quad i = 1, \ldots, n; \ j = 1, \ldots, p.$$

(c) Under the condition in part (b), show that the least squares estimates $\{\hat{\beta}_1, \ldots, \hat{\beta}_p, \hat{\sigma}^2\}$ are independent of each other.

Probability

Problem 4. Let U be a Uniform(0, 1) random variable. Define

$$Y_n = n1_{(0,1/n)}(U) - n1_{[1/n,2/n)}(U).$$

Use Vitali's theorem to show that $\{Y_n\}_n$ is not uniformly integrable, but that $\int Y_n dP \to 0$.

Problem 5. Let $X_1, \ldots, X_n$ be i.i.d. random variables with common cumulative distribution function $F(x) = 1 - e^{-x}$. Define $Z_n = \max_{1 \leq i \leq n} X_i$. Show that

$$\text{limsup} \frac{Z_n}{\log n} \geq 1, \quad \text{a.s.}$$

State the results you are using at every step.

Hint: You may want to use the known inequality $1 - x \leq e^{-x}$.

Problem 6. Let $X(t)$ denote Brownian motion with drift μ, and let T_a denote the time until this process hits the level a. Derive a partial differential equation which $f(a,t) \equiv P(T_a \leq t)$ satisfies.

Problem 7. Let $p_\theta(x)$ be a density on the real line that is Hellinger differentiable at θ_0, i.e., for $\theta \to \theta_0$,

$$\sqrt{p_\theta(x)} = \sqrt{p_{\theta_0}(x)} + \frac{1}{2}(\theta - \theta_0)\tau_0(x)\sqrt{p_{\theta_0}(x)} + r_{\theta,\theta_0}(x),$$

with $\int r^2_{\theta,\theta_0}(x)dx = o(|\theta - \theta_0|^2)$ and $\int \tau_0^2(x)p_{\theta_0}(x)dx < +\infty$. Show the following:

(a) The generalized score function has mean zero:

$$\int \tau_0(x)p_{\theta_0}(x)dx = 0\,.$$

(b) The squared Hellinger distance between P_θ and P_{θ_0} satisfies

$$H^2(P_\theta, P_{\theta_0}) = \frac{1}{4}(\theta - \theta_0)^2 \int \tau_0^2(x)p_{\theta_0}(x)dx + o(|\theta - \theta_0|^2) \quad \text{for } \theta \to \theta_0.$$

Theoretical Statistics

Problem 8. Consider the model $Y = X\beta + e$, where $e \sim N(0, \sigma^2 I)$, and the reduced model $Y = X_0\gamma + e$, $C(X_0) \subset C(X)$. Let M and M_0 be the perpendicular projection operators onto $C(X)$ and $C(X_0)$, respectively.

(a) Show that $M - M_0$ is the perpendicular projection operator onto $C(M - M_0)$.

(b) Show that the reduced model is true if and only if

$$\frac{Y^T(M - M_0)Y/r(M - M_0)}{Y^T(I - M)Y/r(I - M)} \text{ is distributed as } F(r(M - M_0), r(I - M), 0).$$

Problem 9. Suppose $X_1, X_2, \ldots, X_n$ are iid $N(\theta, \theta)$, that is, the mean and variance are both equal to $\theta > 0$.

(a) What is the method of moments (MOM) estimator of θ?

(b) Is the MOM estimator asymptotically efficient? Find the limiting ratio of variances between the MOM estimator and the MLE as $n \to \infty$.

Problem 10. Define $X \sim \mathrm{DP}(\lambda)$ to mean that X is an integer-valued random variable which has the "Double Poisson" distribution with pmf (mass function) given by

$$P(X = x) = \begin{cases} \dfrac{\lambda^{|x|} e^{-\lambda}}{2(|x|!)} & \text{for } x \neq 0 \\ e^{-\lambda} & \text{for } x = 0. \end{cases}$$

Answer the following. Justify your answers.

(a) Suppose $X_1, X_2, \ldots, X_n$ are iid $\mathrm{DP}(\lambda)$. Find a complete sufficient statistic for λ.

(b) Suppose $X_1, X_2, \ldots, X_n$ are independent with $X_i \sim \mathrm{DP}(i\theta)$ (they are **not** identically distributed). Find the best unbiased estimator of θ.

Biostatistics

Problem 11. The proportional hazards model specifies that:

$$\lambda(t|x) = \lambda_0(t) \exp\{\boldsymbol{X}'\boldsymbol{\beta}\}$$

(a) Consider the following observations:

time	x	fail
1	3	1
3	2	0
5	4	1
9	6	0

Write out the partial likelihood associated with the data.

(b) The above contained no ties. Suppose the data were the following:

time	x	fail
3	3	1
3	2	1
5	4	0
9	6	1

Sketch at least two ways we could take the tie into account in forming the partial likelihood contribution at time 3.

Problem 12. Consider the following two by two table representing the results of a cohort study:

		Exposure		
		+	−	Total
Disease	+	a	b	a+b
	−	c	d	c+d
Total		a+c	b+d	n

(a) Define each of the following, both in terms of the table provided and in terms of probabilities:

1. Incidence
2. Prevalence
3. Odds Ratio
4. Relative Risk

(b) Use the delta method to derive an approximate standard deviation for the odds ratio.

Problem 13. Derive sample size formulae for the following scenarios. In each, you may assume an $\alpha = .05$ (one-sided) and a $\beta = 0.10$. You may also assume σ is known and for the two-sample case, is the same for both populations.

(a) You wish to test the hypothesis that the mean in a population is some specified value, i.e. $\mu = \mu_0$ versus the alternative that the mean is some specified but alternative value $\mu = \mu_A > \mu_0$. The test is to be based on a single sample of size n.

(b) You wish to test the hypothesis that the means in two populations are the same, i.e. $\mu_1 - \mu_2 = 0$ versus that they are different by some specified amount, i.e. $\mu_1 - \mu_2 = \Delta > 0$. The test will be based on two samples of size n.

(c) Suppose for some reason, it is much cheaper to select observations in one of the populations. How would you change the derivation of the above two sample problem to take into account different sample sizes (i.e., $n_1 \neq n_2$)?

Computational Statistics

Problem 14. Let $y \in \mathbb{R}$ and $x \in \mathbb{R}^n$ be the response and the predictor random variables in a linear regression problem, and one is interested in solving for $b^* \in \mathbb{R}^n$ such that: $b^* = \text{argmin}_{b \in \mathbb{R}^n} E[|y - x^T b|^2]$. However, the number of predictors n is very large and we will instead use another predictor vector $z \in \mathbb{R}^d$ ($d << n$) and solve the alternate problem: $\tilde{b}^* = \text{argmin}_{\tilde{b} \in \mathbb{R}^d} E[|y - z^T \tilde{b}|^2]$.

The criterion for choosing z is:

$$z = U^T x \ \text{ such that } \ E[\|x - Uz\|^2] \ \text{ is minimized },$$

where $U \in \mathbb{R}^{n \times d}$, $U^T U = I_d$ ($d \times d$ identity matrix), and $\| \cdot \|$ is the two-norm of a vector.

(a) Derive an optimal U and hence an optimal z for which the criterion is achieved.

(b) Suggest an algorithm for performing this linear dimension reduction step before performing linear regression.

(c) Comment on the variation of the regression error $E[|y - z^T \tilde{b}^*|^2]$ versus d.

Problem 15. Let $x_1, x_2, \ldots, x_t$ denote the evolving state of a system in $\mathbb{R}$, and $y_1, y_2, \ldots, y_t$ denote our observations of this system, also in $\mathbb{R}$. We make the following assumptions: (i) $\{x_t\}$ is a Markov process and the one-step transitional density $f(x_t|x_{t-1})$ is easy to sample from. (ii) Given x_t, the observation y_t is independent of all the previous states and previous observations. The likelihood function $f(y_t|x_t)$ is given and easy to evaluate for a given pair (x_t, y_t).

(a) State the nonlinear filtering equations for this system. In other words, write the relationship between the posterior density at time $t+1$, i.e. $f(x_{t+1}|y_1, y_2, \ldots, y_{t+1})$ and the posterior density at time t, i.e. $f(x_t|y_1, y_2, \ldots, y_t)$.

(b) We are interested in estimating the posterior mean

$$\int x_{t+1} f(x_{t+1}|y_1, y_2, \ldots, y_{t+1}) dx_{t+1} .$$

Suggest a sequential Monte Carlo algorithm for this purpose, assuming that you have n perfect samples, $x_t^1, x_t^2, \ldots, x_t^n$, from the previous posterior $(f(x_t|y_1, y_2, \ldots, y_t))$.