Department of Statistics
University of Florida

PhD Qualifying Exam

August 2010

Instructions:

1. You have exactly four hours to answer questions in this examination.
2. There are 8 problems of which you must answer 6. You must do at least one problem from each of the four categories of Linear Models, Generalized Linear Models, Probability, and Inference. If there is doubt as to what area a particular problem covers, then ask.
3. Only your first 6 problems will be graded.
4. You will be given an identifying number for the exam. Write your identifying number on every page in the form SN-x, where x is your number.
5. Do not write your name anywhere on your exam.
6. Write only on one side each sheet of paper. For each problem you do, start the problem on a new page. At the end of the exam, for each problem, staple together all pages for that problem in order.
7. Clearly label each part of each question with the question number and the part, if any, e.g., **1(a)**.
8. You must show your work to receive credit.
9. While the eight questions are equally weighted, within a given question, the parts may have different weights.
10. Do not write near the upper left corner of the page where the pages will be stapled together.

1. Let $\beta_1, \beta_2, \beta_3$ be the interior angles of a triangle, so that $\beta_1 + \beta_2 + \beta_3 = 180$ degrees. Suppose we have available estimates Y_1, Y_2, Y_3 of $\beta_1, \beta_2, \beta_3$, respectively. We assume that $Y_i \sim N(\beta_i, \sigma^2)$, $i = 1, 2, 3$ (σ is unknown) and that the Y_i's are independent. What is the F-test for testing the null hypothesis that the triangle is equilateral?

2. This problem consists of three parts.

(a) Consider the multiple linear regression model

$$Y = X_1\beta_1 + X_2\beta_2 + \epsilon,$$

where Y is $n \times 1$, X_1 is $n \times p_1$, β_1 is $p_1 \times 1$, X_2 is $n \times p_2$, β_2 is $p_2 \times 1$, and ϵ is $n \times 1$. Suppose that in fact $\beta_2 = 0$, in other words, the model used by the experimenter is an overfitted model and the true model is

$$Y = X_1\beta_1 + \epsilon.$$

Let $\hat{\sigma}^2_{\text{overfit}}$ denote the usual estimate of variance based on the overfitted model, i.e., $\hat{\sigma}^2_{\text{overfit}} = Y'(I - P)Y/(n - p_1 - p_2)$, where P is the projection onto the space spanned by the columns of X_1 and the columns of X_2. Show that $\hat{\sigma}^2_{\text{overfit}}$ is an unbiased estimate of σ^2 even if the smaller model is true.

(b) For any standard linear model $Y \sim N(X\beta, \sigma^2 I)$, derive the expression for the usual 95% confidence interval for σ^2.

(c) Let $\hat{\sigma}^2_{\text{red}}$ be the estimate of σ^2 based on the reduced (and correct) model

$$Y = X_1\beta_1 + \epsilon.$$

Show that the expected length of the confidence interval for σ^2 based on the reduced model is smaller than under the overfitted model.

3. Suppose that Y_{ij}, $i = 1, \ldots, m$, $j = 1, \ldots, n_i$, follow a random intercept model of the form

$$\begin{aligned} &Y_{ij} | U_1, \ldots, U_m \sim \text{independent Poisson}(\lambda_{ij}) \\ &\log \lambda_{ij} = \boldsymbol{x}_{ij}^T \boldsymbol{\beta} + U_i \\ &U_1, \ldots, U_m \sim \text{i.i.d.} \end{aligned}$$

Let $Z_i = e^{U_i}$ and let γ be its coefficient of variation, i.e.,

$$\gamma = \frac{\sqrt{\text{Var}(Z_i)}}{E(Z_i)}.$$

(a) Show that $\text{Var}(Y_{ij}) = \mu_{ij}(1 + \gamma^2 \mu_{ij})$ and $\text{Cov}(Y_{ij}, Y_{ik}) = \gamma^2 \mu_{ij} \mu_{ik}$ for $j \neq k$, where $\mu_{ij} = E(Y_{ij})$.

(b) It is common to assume that the U_is are normally distributed (in which case the Z_is are log-normal random variables), but assume instead that $Z_i \sim \text{Gamma}(\alpha, 1/\alpha)$, for some $\alpha > 0$. What are μ_{ij} and γ^2 in this case? Write down the likelihood for $(\boldsymbol{\beta}, \alpha)$ and show that it can be expressed in closed form (i.e., with no unevaluated integrals) using the gamma function.

Note: in the present notation, $\text{Gamma}(\alpha, \zeta)$ *indicates the distribution with density*

$$f(z; \alpha, \zeta) = \frac{1}{\zeta^\alpha \Gamma(\alpha)} z^{\alpha - 1} e^{-z/\zeta}, \quad z > 0,$$

for $\alpha, \zeta > 0$*, with mean* $\alpha\zeta$ *and variance* $\alpha\zeta^2$*.*

4. Suppose that $Y_1, \ldots, Y_n$ are independent with $\mu_i = E(Y_i)$ satisfying

$$\log \mu_i = x_i \beta \qquad (x_i \text{ univariate})$$

and with

$$\text{Var}(Y_i) = \phi \mu_i.$$

(a) Give the quasi-likelihood estimating equation for β and find the asymptotic variance of $\tilde{\beta}$, the "maximum quasi-likelihood estimator" (MQLE) of β.

(b) Assuming that $Y_1, \ldots, Y_n$ are normally distributed (with means and variances as given above), derive the asymptotic variance of $\hat{\beta}$, the MLE of β.

(c) The "asymptotic relative efficiency" (ARE) of $\hat{\beta}$ with respect to $\tilde{\beta}$ is the ratio of the asymptotic variance of $\tilde{\beta}$ to that of $\hat{\beta}$. Use your results to find a formula for the ARE of the MLE $\hat{\beta}$ with respect to the MQLE $\tilde{\beta}$ assuming that $Y_1, \ldots, Y_n$ are normally distributed. Is the ARE greater or less than 1? How does it vary with ϕ?

5. Let $X_1, \ldots, X_p$ be p (≥ 3) independent $N(\theta_i, \sigma^2)$ random variables, where both $\boldsymbol{\theta} = (\theta_1, \ldots, \theta_p)^T \in R^p$, and σ^2 (> 0) are unknown. We write $\boldsymbol{X} = (X_1, \ldots, X_p)^T$, and $\boldsymbol{x} = (x_1, \ldots, x_p)^T$. Let $h(\cdot)$ be a real-valued function satisfying $E_{\boldsymbol{\theta},\sigma^2}|\partial^2 \log h(\boldsymbol{X})/\partial X_i^2| < \infty$ and $E_{\boldsymbol{\theta},\sigma^2}[\partial \log h(\boldsymbol{X})/\partial X_i]^2 < \infty$ for all $i = 1, \ldots, p$. Suppose that the loss incurred in estimating $\boldsymbol{\theta}$ by $\boldsymbol{a}$ is $L_{\sigma^2}(\boldsymbol{\theta}, \boldsymbol{a}) = \|\boldsymbol{\theta} - \boldsymbol{a}\|^2/\sigma^2$. Also, let U be a random variable distributed independently of the X_i's such that $U \sim \sigma^2\chi_m^2/(m + 2)$.

(a) Show that $\boldsymbol{T} = (X_1 + U(\partial \log h(\boldsymbol{X})/\partial X_1), \ldots, X_p + U(\partial \log h(\boldsymbol{X})/\partial X_p))^T$ improves on $\boldsymbol{X}$ for estimating $\boldsymbol{\theta}$ if $2\sum_{i=1}^p \frac{\partial^2 \log h(\boldsymbol{x})}{\partial x_i^2} + \sum_{i=1}^p (\frac{\partial \log h(\boldsymbol{x})}{\partial x_i})^2 < 0$ for almost all $\boldsymbol{x} \in R^p$.

(b) Take $h(\boldsymbol{x}) = \|\boldsymbol{x}\|^{-(p-2)}$. Show that with this choice of h, the estimator $\boldsymbol{T}$ given in (a) improves on $\boldsymbol{X}$ for estimating $\boldsymbol{\theta}$.

(c) Find an unbiased estimator of the risk improvement given in (b).

6. Let $X_1, X_2, \ldots, X_n$ be iid $N(\mu, \sigma^2)$, where μ (real) and σ^2 (> 0) are both unknown.

(a) Find the Fisher information matrix $I(\mu, \sigma)$.

(b) Define $\bar{X}_n = n^{-1}\sum_{i=1}^n X_i$ and $S_n^2 = n^{-1}\sum_{i=1}^n (X_i - \bar{X}_n)^2$. Show that $n^{1/2}(\bar{X}_n - \mu, S_n - \sigma)^T$ is asymptotically normal with zero means, variances σ^2 and $\sigma^2/2$ and covariance zero.

(c) Suppose now the normality assumption is dropped, but the X_i's are assumed to have a finite fourth moment. Let $\mu_3 = E(X_1 - \mu)^3$ and $\mu_4 = E(X_1 - \mu)^4$. Show that $n^{1/2}(\bar{X}_n - \mu, S_n - \sigma)^T$ is asymptotically normal with zero means, variances σ^2 and $(\mu_4 - \sigma^4)/(4\sigma^2)$ and covariance μ_3.

7. Let $\{X_n, n \geq 1\}$ be a sequence of independent random variables and let $\{b_n, n \geq 1\}$ be a sequence of positive constants with $\lim_{n\to\infty} b_n = \infty$. Suppose that

$$\lim_{n\to\infty} \sum_{j=1}^{n} P(|X_j| > b_n) = 0,$$

$$\lim_{n\to\infty} \frac{1}{b_n} \sum_{j=1}^{n} E(X_j I_{[|X_j| \leq b_n]}) = 0,$$

and

$$\lim_{n\to\infty} \frac{1}{b_n^2} \sum_{j=1}^{n} \mathrm{Var}(X_j I_{[|X_j| \leq b_n]}) = 0.$$

Set $S_n = \sum_{j=1}^{n} X_j, n \geq 1$.

(a) Prove that

$$\frac{S_n}{b_n} \xrightarrow{P} 0.$$

(b) Demonstrate by way of a suitable example that the above hypotheses do not necessary ensure that

$$\frac{S_n}{b_n} \to 0 \text{ a.c.}$$

8. Let $\{X_n, n \geq 1\}$ be a sequence of independent random variables with

$$EX_n = 0,\ 0 < EX_n^2 = \sigma_n^2 < \infty,\ n \geq 1$$

and set $S_n = \sum_{j=1}^{n} X_j$ and $s_n^2 = \sum_{j=1}^{n} \sigma_j^2, n \geq 1$. Define for all $n \geq 1$ and all $\varepsilon \in (0, \infty)$,

$$\Gamma_n(\varepsilon) = \frac{1}{s_n^2} \sum_{j=1}^{n} E(X_j^2 I_{[|X_j| > \varepsilon s_n]}) \ +\ \frac{1}{s_n^3} \left| \sum_{j=1}^{n} E(X_j^3 I_{[|X_j| \leq \varepsilon s_n]}) \right| \ +\ \frac{1}{s_n^4} \sum_{j=1}^{n} E(X_j^4 I_{[|X_j| \leq \varepsilon s_n]}).$$

(a) Prove that if

$$s_n^2 \to \infty,\ \frac{\sigma_n^2}{s_n^2} \to 0,\ \text{and}\ \frac{S_n}{s_n} \xrightarrow{d} N(0,1),$$

then

$$\lim_{n\to\infty} \Gamma_n(\varepsilon) = 0 \text{ for all } \varepsilon \in (0, \infty).$$

(b) Prove that if

$$\lim_{n\to\infty} \Gamma_n(\varepsilon_0) = 0 \text{ for some } \varepsilon_0 \in (0, \infty),$$

then

$$s_n^2 \to \infty,\ \frac{\sigma_n^2}{s_n^2} \to 0,\ \text{and}\ \frac{S_n}{s_n} \xrightarrow{d} N(0,1).$$