Qualifying Exam: Topics in Statistics: January 2011

1. This question has two independent parts. Here $\boldsymbol{X}^T$ denotes the transpose of the matrix $\boldsymbol{X}$.

(a) Prove that the deviance for normal linear models is $D = \frac{1}{\sigma^2}(\boldsymbol{y}^T\boldsymbol{y} - \boldsymbol{b}^T\boldsymbol{X}^T\boldsymbol{y})$, where $\boldsymbol{y}^T$ denotes the transpose of the vector $\boldsymbol{y}$ and $\boldsymbol{b}$ denotes the OLS estimator of $\boldsymbol{\beta}$.

(b) For normal linear models when there is only one parameter, that is, when $E(Y_i) = \mu$ for all $i = 1, \ldots, n$, what is the expression for the OLS estimator b? Show that the deviance for this case reduces to $(n-1)S^2/\sigma^2$, where S^2 is the sample variance. Hence write down the exact distribution of the deviance for this case.

2. This question has two independent parts. Here $\boldsymbol{X}^T$ denotes the transpose of the matrix $\boldsymbol{X}$.

(a) For logistic regression show that the elements of $\boldsymbol{X}^T\boldsymbol{y}$ are sufficient statistics for $\boldsymbol{\beta}$.

(b) One form in which the negative binomial probability distribution is described is given by

$$f(y; \pi, \phi) = \binom{y-1}{r-1}\pi^r(1-\pi)^{y-r},$$

where r is assumed to be known. *For this particular form*, show that it belongs to the exponential family and find $E(Y)$ and $\mathrm{Var}(Y)$. Also derive the canonical link for this distribution.

3. In dose-response studies investigators are interested in the effect of various dose levels of a substance on the study subjects. The aim is to describe the probability of a desired outcome as a function of the single covariate, dose, denoted by x. Typically the data are the number of subjects n_i administered the dose level x_i, and the number y_i that experience the outcome of interest. Here i is from 1 to m for some $m > 0$.

(a) Describe completely the most appropriate model you will use to fit the data.

(b) Write down the log-likelihood function and the score equations for the parameters explicitly.

(c) Obtain the information matrix. You final answer must be an appropriate matrix whose elements must be specified explicitly.

(d) Describe clearly how you would obtain the MLEs of the parameters. For this purpose, you should clearly specify the iterative equation that will be solved.

4. This question has two independent parts.

(a) Prove or disprove the statement that for normal linear models the MLE of σ^2 is biased.

(b) For normal linear models, derive the joint distribution of the vectors $\hat{\boldsymbol{y}}$ and $\boldsymbol{e}$ and hence show that they are independent.

5. Provide a most appropriate experimental design for each case below. For each question, where appropriate, specify (a) your experimental design, (b) the reason for choosing the model; (c) factors and their types types (fixed, random, or mixed); (d) the statistical model.

(a) An industrial engineer is investigating the effect of four assembly methods (A,B,C,D) on the assembly time for a color television component. Four operators with different technical experience are selected for the study. Furthermore, the engineer knows that each assembly method produces such fatigue that the time required for the last assembly may be greater than the time required for the first, regardless of the method. That is, a trend develops in the required assembly time. What is your experimental design and why do you choose the model?

(b) The yield of a chemical process is being studied. The two factors of interest are temperature and pressure. The factor "temperature" has three different levels of low, medium and high, and the factor "pressure" has also three levels at 250, 260 and 270. That is, three levels of each factor are specifically selected. Due to restriction of the chemical laboratory, only nine runs can be made in one day. The experimenter runs a complete replicate of the design on each day. The experiments are repeated in the consecutive day.

(c) An experiment was conducted to compare the effects of four different drugs (A,B,C and D) in delaying atrophy of denervated muscles. A certain leg muscle in each of 48 rats was deprived of its nerve supply by surgical severing of the appropriate nerves. The rats were then put randomly into four groups, and each group was assigned treatment with one of the drugs. After 12 days, four rats from each group were killed and the weight W (in gram) of the denervated muscle was obtained. Theoretically, atrophy should be measured the loss in weight of the muscle, but the initial weight of the muscle could not have been obtained without killing the rat. Consequently, the initial total body weight X (in gram) of the rat was measured. It was assumed that this figure is closely related to the initial weight of the muscle.

6. Seven single-family homes were sampled in each of four location (SE,SC,NC,NW) in the metropolitan Newark area. House price average (in 1000 units) are $\bar{Y}_{SE} = 35.7$, $\bar{Y}_{SC} = 43.5$, $\bar{Y}_{NC} = 66.9$, and $\bar{Y}_{NW} = 89.2$. The ANOVA table is as follows:

	d.f.	Mean Square
Among Locations	3	4100.8
Within Locations	24	115.2
Total	27	4716.0

(a) Compute the estimate of the effect of living in the south central location.

(b) It is claimed the mean price of home in between the north west (NW) and the south east (SE) is the same as that between the north central (NC) and the south central (SC). Write the null and alternative hypotheses and carry out a test of hypothesis for testing H_0 at $\alpha = 0.10$.